Document Title:	Classifying Adult Probationers by Forecasting Future Offending
Author:	Geoffrey C. Barnes, Ph.D., Jordan M. Hyatt, J.D., M.S.
Document No.:	238082
Date Received:	March 2012
Award Number:	2008-IJ-CX-0024

Opinions or points of view expressed are those of the author(s) and do not necessarily reflect the official position or policies of the U.S. Department of Justice.

Classifying Adult Probationers by Forecasting Future Offending

Final Technical Report

Grant No. 2008-IJ-CX-0024

Geoffrey C. Barnes, Ph.D
Principal Investigator

Assistant Research Professor
Department of Criminology
University of Pennsylvania
gbarnes@sas.upenn.edu

Jordan M. Hyatt, J.D., M.S

Senior Research Coordinator
Jerry Lee Center of Criminology
University of Pennsylvania
jhyatt@sas.upenn.edu

3718 Locust Walk, 483 McNeil Building
Philadelphia, PA 19104
phone: (215) 746-3537
fax: (215) 898-6891

.

This project was supported by grant No. 2008-IJ-CX-0024 awarded by the National Institute of Justice, U.S. Department of Justice. Points of view are those of the author and do not necessarily represent the official position or policies of the U.S. Department of Justice.

Statement of Support, Disclaimers and Acknowledgements

This project was supported by grant No. 2008-IJ-CX-0024 awarded by the National Institute of Justice, U.S. Department of Justice. All of the opinions, findings, and conclusions or recommendations expressed in the publication are those of the authors and do not necessarily reflect the views of the Department of Justice.

This research was conducted in partnership with Pennsylvania's First Judicial District and Philadelphia's Adult Probation and Parole Department (APPD). While their support was obviously crucial to the accomplishments detailed here, they also provided a fantastic and fun environment in which to collaborate. The openness and degree of consultation that we have had with this agency has been breathtaking in scope, and we have deeply enjoyed working with everyone there. Specifically, the authors would like to thank our APPD research partners, without whom this project would not have been possible:

Robert Malvestuto, Chief Probation and Parole Officer
Charles Hoyt, Deputy Chief Probation and Parole Officer
Ellen Kurtz, Ph.D., Director of Research
Lindsay Ahlman, Research Associate
Amanda Tudor, Research Associate

In addition to our colleagues at APPD, this project would not have been possible without the support of Dr. Richard Berk, whose statistical models and advice served as the backbone of this project. While Richard was never a named participant on this grant, he was so active in its work and giving of his time that he really should be regarded as the driving force behind everything described here.

The authors would also like to thank Dr. Lawrence Sherman for his guidance and support throughout the project and for laying the groundwork for this research partnership. Were it not for his skill and persistence at building relationships, all of the necessary partnerships would never have come to be.

Abstract

Random forest modeling techniques represent an improvement over the methodologies of traditional risk prediction instruments. Random forests allow for the inclusion of a large number of predictors, the use of a variety of different data sources, the expansion of assessments beyond binary outcomes, and taking the costs of different types of forecasting errors into account when constructing a new model. This study explores the application of random forest statistical learning techniques to a criminal risk forecasting system, which is now used to classify adult probationers by the level of risk they pose to the community.

The project principally focused upon creating a risk prediction tool within a partnership between University-based researchers and Philadelphia's Adult Probation and Parole Department (APPD). This report details the model building process, including an explanation of the random forest procedures, and sets out the issues in data management and in policy considerations that are associated with creating a prediction tool. The importance of developing strong researcher-practitioner partnerships, especially with regard to tailoring the prediction tool to real-world concerns, is also considered.

The prediction models developed as a part of this project have been used since 2009 to assess all incoming probation and parole cases. Each risk prediction is then used to assign offenders to risk-stratified supervision divisions. This report discusses the development and accuracy of the three generations of models that have been employed to make these predictions, as well as the salient features of each iteration. For the most recent version of the model, the influences of major predictor variables are discussed, with a focus on those that were most powerful in the Philadelphia sample. Additionally, the predictions of the three models are also validated using a sample of cases from 2001, a cohort not used to build any of the prediction models. The long-term offending patterns of these 2001 offenders, with regard to their assessed risk level, are also considered. Finally, suggestions and step-by-step instructions are offered for practitioners seeking to build a similar prediction instrument for use in a wide variety of criminal justice settings.

As a matter of policy, criminal sanctions that are encompassed under the umbrella of community corrections have become an increasingly prevalent punishment option. Given resource constraints, as well as concerns about public safety, statistical risk assessment tools – such as the one developed here – have begun to take increasingly prominent roles in determining levels of supervision. This report, and the partnership supporting it, highlights the promises and potential of the methodology.

Contents

Executive Summary

This report presents results from a demonstration project focused on the construction of a risk prediction model for a local probation and parole department in a large urban city. The resultant Risk Forecasting Tool is now used to assess all new probation cases at their outset, and allows the department to tailor its supervision protocols in a manner that reflects the danger that the individual probationers pose to the community. The results shown here demonstrate that the project's risk forecasting models, developed through a joint partnership between Philadelphia's Adult Probation and Parole Department (APPD) and University of Pennsylvania researchers, were able to increase the Department's ability to predict recidivism, and led to the restructuring of supervision protocols for the agency.

This project resulted in the construction of three different prediction models based on a statistical process known as "random forests". Each of these models was used to make on-demand, live forecasts for every single incoming probationer. These operational models, along with the software written to execute them, are designed to extract all the necessary data concerning each probationer's past, and then use these values to make the thousands of comparisons required to produce a forecast at the start of every new probation case. Since the first live forecasting tool was installed in 2009, nearly 115,000 new case starts have been run through these models, at a rate of almost 47,000 forecasts per year.

The Problem Addressed

Offenders under community supervision, comprised of both probation and parole, represent a significant portion of the total correctional population. Approximately 1 in every 45 adults in the United States is under some form of community correctional supervision (Pew Center on the States 2009). In Philadelphia, as in many other county-level agencies across the country, the traditional means of supervising a large caseload has often required that each offender, regardless of their conviction offense or prior criminal history, be subject to standardized conditions and reporting requirements. In an attempt to reframe their supervision policy, APPD partnered with the Jerry Lee Center at the University of Pennsylvania in order to bring supervision strategies in line with the Department's focus on protecting the community.

The goal of the partnership between Philadelphia's APPD and the University of Pennsylvania was to produce a method of forecasting which offenders posed the largest risk of committing new offenses – and especially new serious crimes – while under the department's supervision. Through these forecasts, the partnership developed the ability to provide an appropriate level of supervision for offenders who presented differing levels of risk.

Purpose & Hypothesis

As a demonstration project, this work was undertaken with three concrete and particular objectives:

1. To develop a random forest model that predicts the likelihood of recidivism within the population of a large urban probation and parole agency;
2. To develop the software needed to make these forecasts instantly available to the agency whenever they were required; and,
3. To test and validate the predictions made by that model, as well as its subsequent generations, using data that were not employed in their original construction.

Model Construction

As with most other statistical models, random forest forecasting has certain core requirements and parameters that must be determined before construction can begin. Detailed below, these choices represent the decisions that were made for this particular project and highlight the need to balance statistical concerns with pragmatic needs facing the agency and researchers.

Unit of Prediction. The unit of prediction defines the point in time at which the forecast should begin. Anything that has happened prior to this point in time can be used, if the data are available, to form predictor values, while everything that takes place afterwards is a potential outcome that can be forecasted. For the APPD models, an individual probation case start is used as the unit of prediction, and the beginning of its supervision period marks the starting point for each forecast. This means, in practice, that many of the department's offenders – namely those who are under concurrent supervision for multiple different convictions – often have more than one forecast active at any given time.

Time Horizon. The time horizon determines both how long each forecast can be considered valid, and how the data needed to construct the model will be defined. In all of these forecasting models, the time horizon is set at two years from the start of each new probation case. Each of these models, therefore, had to be constructed from a sample of cases that were at least two years old, so that the outcome of each case start could be fully known and measured.

Outcomes. The most fundamental decision when constructing models like these is determining what kinds of behavior the model should forecast, and how risk should be defined. Based on an analysis of the data available, and the APPD's overall mission, the following categorical outcomes were defined:

> **High Risk**: the offender was predicted to commit at least one serious offense – defined as murder, attempted murder, aggravated assault, robbery, or a sexual crime – during the first two years after their case start date;
>
> **Moderate Risk**: the offender was predicted to commit only non-serious offenses during the first two years after their case start date, and;

Low Risk: that the offender was not predicted to commit any new offenses, of any kind, during the first two years after their case start date.

One of the most significant considerations when defining these forecastable outcomes is need to balance competing requirements. In this case, the APPD needed to both supervise offenders within their appropriate risk classification, and to maintain manageable caseloads. The department determined that it could handle approximately 15% of its overall caseload in the "High Risk" category, with another 25-30% being classified as "Moderate Risk." The remaining 55-60% were placed in the "Low Risk" group, where they would be given the lowest amount of supervision. The random forest procedure, fortunately, can accommodate this difficult balancing act without sacrificing significant accuracy.

Predictors Used. One benefit of random forest modeling is that there is no theoretical limit on the number of predictors that can be included in the model. Throughout the duration of the project, hundreds of different predictor variables were drawn from electronically-available administrative records, and tested for possible use in these models. Of these, only 53 were ever employed for live forecasting. Broadly speaking, the universe of predictors used here included: age at case start, residential zip code, neighborhood demographics, the number of instant charges (categorized into offense types such as drugs, property, violent, serious, sexual, and firearms), the number of prior charges (again, categorized into different offense types), the offender's history of incarceration and probation sentences, juvenile criminal history, and others.

Models Constructed

A number of different models were constructed during the duration of the project. Most of them were to test the usefulness of newly available predictor variables, and to evaluate different approaches to balancing APPD's needs with the ability to maximize forecasting accuracy. Ultimately, three versions of the prediction tool (referred to as Models A, B, and C) were installed into the APPD's Risk Forecasting Tool and were used for the live prediction of offender behavior. Each of these models was constructed using data from the same sources, but varied in the size of the samples used to build them, the mix of predictor variables, the relative costs assigned to different kinds of forecasting errors, and the size of the model produced.

Model A was constructed in January 2009, and was used for incoming case forecasting at APPD from March 2009 through April 2010. To construct this model, a sub-sample of 50,000 cases was drawn from all APPD starts between January 1, 2002 and December 31, 2005. The 34 predictor values included prior criminal charge history, prior sentences received, time spent in county jails, and the demographics of each offender's neighborhood. Overall, the full specification for Model A included 4.57 million different decision points.

Model B, constructed in December 2009, and was used for live case forecasting at APPD from April 2010 through November 2011. The sample used to construct the model was selected from all of the agency's new case starts between January 1, 2002 and December 31, 2006. Unlike Model A, the set of predictor variables included juvenile offending data, increasing the total number of variables to 48. As before, 50,000 sample cases were randomly selected from the overall sample of 94,653 probation starts to build Model B. This model was roughly a quarter of the size of its predecessor, containing 1.20 million decision points.

Model C, which is currently in use for live prediction, was built in August 2011 and was installed for live forecasting in November 2011. In this case, the construction sample consisted of all 119,988 APPD cases which began between January 1, 2002 and December 31, 2007. Since more than a third of offenders in this sample had more than one case start in the data, these 119,988 case starts are spread across just 71,976 different offenders. Model C, with 8.74 million decision points, is the largest of the predictions models.

Analytical Findings

The random forest procedures used here provided measures of the relative power of individual predictor variables, and allowed the department to specify the cost ratios for different types of errors. Additionally, by using a subsample of 2001 case starts that were not used in the construction process, the three generations of models can be compared to each other and validated independently.

<u>The Relative Influence of Predictors</u>. The analysis showed that some predictors were more important to the overall accuracy of the model than others. For example, in Model C, the number of prior stays in the county prison system was of key importance, while the offenders' residential zip codes, the time elapsed since their most recent serious offense, current age, and age at the time of their first adult offense all combined to form a strong second tier of important predictors. Interestingly, the three least-important predictor variables for the model, when considered as a whole, were the onset age for juvenile offending, the number of serious-crime charges stemming from the case that resulted in the offender being placed on supervision, and the count of prior charges for sexual offenses. It is worth noting that, since there is little penalty for including additional predictors – even when they add little in the way of predictive power –a wide variety of different predictors can be used to construct these models.

<u>Cost Ratios</u>. The project's random forest prediction models were developed for use in a large and very busy agency and on actual cases. As with any forecasting effort, errors were inevitable, including both over and under-estimates of risk levels. Construction of these models therefore required the consideration of these different mistakes, and the relative harm that they may cause. To provide a concrete example, researchers and practitioners were forced to consider how much more costly it would be to mistakenly classify a future serious offender into a lower risk category (i.e., a High Risk false negative) than it is to supervise someone who would actually turn out to be a non-serious offender as though they really were High Risk (a false positive). Fortunately, these costs can be built into the model itself as a function of the random forest procedure.

Based on the data used to construct Model C, there were 11,700 false positives for High Risk, as compared to just 4,468 false negatives. The final cost ratio, therefore, turned out to be approximately 2.6; each false negative was deemed to be slightly more than 2½ times more costly than each false positive. As a result, these false negatives occur much less often. This ratio of differential costs was determined by the agency's leadership, in keeping with their missions to both protect public safety and allocate caseloads in a manageable way.

Given the cost ratios used to construct them, all three generations of these models were designed to generate more High Risk false positives than false negatives. In practice, this means that more offenders are placed into the forecasted High Risk category than will actually go on to commit a new serious crime. These cost ratios have important repercussions for caseload size and officer workload, and therefore are always a very delicate balancing act. In this instance, the agency slowly adjusted the cost ratios with each successive version of the model, and gradually allowed for an increasing number of cases – including more false positives – to be classified as High Risk.

Estimated Model Accuracy. In order to assess the overall accuracy of any forecasting model, its predictions must ideally be validated against cases that were not employed in its construction. The random forest procedures used here, however, can estimate each model's accuracy by using a unique sub-sample of cases that are held in reserve as different parts of the model are created. Although these cases are not drawn from a fully-independent validation sample, they represent the model's best estimates of how well it would perform with this kind of sample.

Overall, the most recent version of APPD's model (i.e., Model C) produced an accurate forecast for 79,299 of the 119,935 probation case starts in the construction sample. In overall terms, these estimates suggest that model can be correct nearly two-thirds (66.1%) of the time. However, a more reasonable method of measuring the model's accuracy is to examine forecasted and actual outcomes separately, focusing on each of the three different outcome categories. For example, of 18,812 new case starts that were forecasted to be High Risk, just 7,112 (37.8%) of them involved an offender who later committed a serious offense during the two year time horizon. It may be worth noting that, though this percentage may seem low, the actual prevalence of serious offending is rare and occurs in less than 10% of all cases.

Another way to consider Model C's accuracy is to compare cases where the actual risk category is known, and then measure how many were forecasted correctly by the model. All three of these estimates are in excess of 60%. For example, of those case starts that actually resulted in new serious offending, the model correctly identified 61.4% of them as High Risk. Those case starts where the offender turned out to be actually Moderate Risk featured a similar level of accuracy (61.9%), while those where the offender was not charged with any new crimes were forecasted correctly nearly 70% of the time.

External Validation of the Model. The estimated accuracy of Model C's predictions described above was computed using the same sample of cases that was used, during the construction of the model itself. A 'cleaner' way to assess the model is to rely upon cases that were not a part of its construction sample, such as a sample of earlier probation cases that began between January 1, 2001 and December 31, 2001.

Model C, when evaluated using the 2001 cohort data, has an overall accuracy of 57.8% within the model's two year time horizon. This represents a slight departure from the 66.1% accuracy estimated from the construction sample, though this difference may be an artifact of the slight differing sample characteristics. Overall, 35% of offenders who went on to become actually High Risk were accurately identified. The same holds for 38.7% of those who became actual Moderate Risks and 69.4% of actual Low Risks.

Since the 2001 cohort was not employed to build any of the models constructed for this project, it can also be used to compare each of them to one another. Despite variability in the predictors used and the size of the models themselves, all three models produced approximately the same degree of overall accuracy, forecasting the correct outcomes approximately 60% of the time. The cost ratios of the models, as noted earlier, changed over time. In general terms, these changes allowed a larger proportion of offenders to be forecasted as High Risk, increasing from 12% in Model A to 14.9% in Model C. These changes in cost ratio reflect shifts in APPD's policies and abilities. Although the current model may produce slightly more errors than prior versions, this increase is largely intentional and reflects APPD's growing level of comfort with larger High Risk caseloads.

Since the validation sample cases are significantly older, it is also possible to examine the offenders' actual behavior over a much longer period of time – well beyond the models' own two-year time horizon – than would be possible with a more recent cohort. Using the forecasts from Model C, more than a third (36%) of the forecasted High Risk cases resulted in new serious offending within five years of the case start date, compared to just 20% of forecasted Moderate Risk, and 10% of forecasted Low Risk case starts within five years for the assessment. After eight years, 45% of all forecasted High Risk offenders had committed a new serious offense, while only 27% of forecasted Moderates and 14% of forecasted Lows have done so. These long-term results suggest that the model's high risk false positives may not be simple and clear errors in forecasting. In many instances, these forecasts were not incorrect in forecasting new charges for serious crime, but instead erred merely by forecasting these offenses too soon.

When the focus is moved away from serious crimes, and new charges for any sort of offending are considered, it becomes clear that even the lowest risk probationers are quite likely to reoffend at some point in the (possibly distant) future. While cases in the forecasted Low Risk group were much less likely to lead to some form of new offending than those in the other two forecasted risk groups, more than half of them resulted in a new criminal charges by the time eight years had elapsed. In contrast, more than 80% of the other two groups were charged with new offenses within eight years of their original probation case start date.

Model Building Recommendations

Many of the problems encountered, and surmounted, during this project were not unique to Philadelphia. Indeed, an attempt to build a random forest prediction model in any jurisdiction would face similar challenges. The list of steps shown below may serve as a blueprint for those seeking to undertake such a task.

1. Obtain access to reliable data that are consistently and electronically available.

2. Define the unit of prediction and desired time horizon.

3. Define the outcome risk categories.

4. Consider the practical implications for a risk-based supervision strategy and ensure adequate resource allocation relative to the distribution of risk scores.

5. Choose the predictor variables to be used in the model based on theoretical, practical and policy-based considerations.

6. Build the construction data in a single data file.

7. Estimate the relative costs of false positives and false negatives; allow the agency leadership to value the weight of these inaccuracies.

8. Build an initial model and evaluate the results.

9. Adjust the resultant model to reflect policy-based concerns regarding accuracy and proportional assignment to output categories, and construct additional test models where required.

10. Produce forecasts for those offenders already in the agency's caseload.

11. Create the user interface and back-end software needed to produce live forecasts.

12. Continuously monitor the results of the live forecasts.

Conclusion

The power and promise of these random forest forecasting methods is clear. In Philadelphia, their introduction has allowed the agency to stratify offenders by the risk they pose, to tailor supervision requirements, to focus resources in accordance with policy directives, and to balance caseload sizes in the face of budgetary constraints. Though these techniques may give rise to important questions of ethics and justice, they also represent an opportunity to advance the capabilities of the criminal justice system to protect communities.

Though it is already becoming apparent that these kinds of models and stratified supervision policies will become more widely used in the future, these forecasts will never be error-free. Policy-makers, in conjunction with researchers, can work to aid in the selection of appropriate predictors, to ensure the proper balancing of different kinds of errors, and to control the use of these predictions in order to assure the targeted and appropriate use of these powerful statistical tools.

Introduction

Statement of the Problem

The current correctional landscape is disproportionally comprised of offenders on probation or parole. Approximately 1 in every 45 adults in the United States is under some form of community correctional supervision (Pew Center on the States 2009), far exceeding the 1 in 100 adults representing the penal population (Pew Center on the States 2008). This distribution is repeated on the state level; for example, in Pennsylvania alone, 258,905 individuals were on probation or parole in 2007, a figure more than 5.6 times the inmate population in the state's correctional institutions (Emery, et al. 2008). It is clear that, given fiscal pressures, probation officials are going to – with escalating frequency – be asked to do more with less.

One of the most readily actionable ways to make probation more "effective," as well as evidence-based, is through the use of standardized, actuarial risk assessment procedures (Lowenkamp, Latessa and Holsinger 2006). Every day, probation officials are tasked with making difficult decisions with potentially serious consequences. In doing so, they must balance the allocation of scarce resources with an overarching mission to protect public safety. By deploying effective risk instruments, officials can:

- classify large populations in a consistent manner;
- supervise offenders based on their likely conduct;
- identify suitable intervention strategies;
- reduce rates of recidivism; and
- reduce overall costs of supervision.

Though the process can be complex, integrating universal risk assessments into case management systems, as was the goal in Philadelphia, allows probation officials to progress towards managing their populations in an effective manner. In order to do so, prediction tools must be accurate, efficient and usable; limitations in statistical capabilities and in instrument design have prevented many prior assessment tools from generating accurate, policy-relevant predictions of recidivism (Rhodes 2001).

Literature Review

Community-based supervision is considered a relatively inexpensive punishment to administer (Petersillia 1997). In the face of rising correctional costs, it will likely remain a heavily relied-upon sanctioning option. Although prisons and jails consume far more resources, the national population of offenders on community supervision remains more than twice the amount of those incarcerated (Glaze 2010). As caseloads expand and financial resources either shrink or remain constant, probation and parole agencies will be forced to reconsider the amount of supervision that they can reasonably deliver (Austin 2010). Risk forecasting offers one potential opportunity to meet this challenge.

Risk forecasting is more than simply a financial imperative. It can identify those offenders that, given the most accurate and up-to-date statistical evidence possible, present the largest danger to the community, and allow corrections officials to target this narrow group of offenders. Recent statistics indicate that over half of the jail inmate population was, at the time of their most recent arrest, under the supervision of probation, parole, or pretrial release authorities (National Research Council, Committee on Community Supervision and Desistance from Crime, Committee on Law and Justice, Division of Behavioral and Social Sciences and Edcuation 2007). Accurate risk predictions can allow scarce resources – including both supportive programming and increased levels of supervision – to be focused on these "power few" (Sherman 2007) in a preventative framework focused on reducing recidivism.

Despite a significant body of literature addressing the impact of probationary sentences, and the effectiveness of a number of programs and supervision protocols (Gill 2010), the assessment and forecasting of risk is typically an atheoretical endeavor. Rather than focusing upon any one explanation for how criminal behavior occurs, these efforts seek to pull the predictive power from any and all variables that add explanatory power to their models. This is not to say, however, that the practical implications of a risk-based supervision protocol are uninformed by theoretical criminology. Theory contains numerous hints about the impact of risk-stratified supervision, but the combined effect of these suggestions is simply inconclusive. For example, when identified low risk probationers are supervised less stringently, some theories, including deviant peer contagion (Dodge, Dishion and Lansford 2006), would expect reduced criminality, because probationers were not exposed to one another during more frequent visits to the probation department. Specific deterrence theory, on the other hand, would predict an opposite effect, as reduced reporting requirements could support a perception of reduced certainty, severity or celerity of sanctioning (Gibbs 1975) in response to any future criminal behavior (Barnes, et al. 2010).

Regardless of the forces underlying behavioral reaction, both practitioners and scientists have sought to harness the potential power of statistically-driven risk assessments for some time. The history of actuarially-developed forecasts demonstrates that they can out-perform subjective human judgments in most, if not all, situations (Gottfredson and Moriarty 2006). Despite the potential benefits, criminal justice practitioners have remained reluctant to rely on these kinds of forecasts, given their concerns about their impact on case management (Andrews, Bonta and Wormoth 2006). However, recent advances in statistical methodology have allowed for risk predictions to be made with ever-increasing accuracy and for the integration of data that, in previous generations, were simply not available.

Historically, forecasting processes in criminal justice, regardless of the outcomes being predicted, have been derivations of traditional linear regression (Berk 2008a, Berk 2008b) However, as Berk notes, "recent work. . . has addressed important concerns that result from model selection methods, symmetric loss functions, and overreliance on linear models (Berk 2008a, 236)." These newer techniques allow for the construction of adaptable risk prediction models that take account of differentially weighted kinds of errors. Unlike early attempts at prediction, these newer methods do not require binary outcome variables to represent the mere presence or absence of given event. Next-generation forecasting models like these have been successfully designed to predict homicide (Berk, Sherman, et al. 2009), violence in a correctional setting (Berk and de Leeuw 1999), as well as the role of race in capital punishment (Berk, Azusa and Hickman 2005). Similar models have also been used, within the same probation population as in this project, to identify those offenders who did not, at the time they began their sentence, pose a threat of serious recidivism (Barnes, et al. 2010). This body of work has served as the foundation and framework for the current research.

Statement of Hypothesis

As a demonstration project, the current effort did not focus upon any testable hypotheses per se. Instead, this project was undertaken with three overall goals:

- First, to develop a random forest model (described below) that adequately predicted the likelihood of both serious and less-serious recidivism within the population of a large urban probation and parole agency;
- Second, to develop the software needed to make these forecasts instantly available to the agency whenever they were required; and,
- Third, to test and validate the predictions made by that model, as well as its subsequent generations, using data that were not employed in their original construction.

The sections that follow set out the characteristics of the forecasting models developed through a joint partnership between Philadelphia's Adult Probation and Parole Department (APPD) and University of Pennsylvania researchers, and describe how this model was implemented to provide live, real-time forecasts. The precision of three generations of this model are then compared, with a focus on the accurate prediction of serious recidivism, as well as the logistics and processes necessary for replication of the random forest prediction process.

Background and History

In mid-2005, the leadership of Philadelphia's Adult Probation and Parole Department (APPD) approached the newly-created Department of Criminology at the University of Pennsylvania, and proposed the formation of a partnership. Philadelphia's APPD is a county-level probation and parole agency, and is encompassed within the First Judicial District of Pennsylvania. In Pennsylvania, each local department is charged with supervising any offenders who are sentenced to probation by the courts within its own judicial district. In addition, these departments also supervise a smaller number of county-level parolees. These parolees are offenders who were serving sentences of less than 2 years in their county's prison system, and who were granted parole by their sentencing judge. These local departments typically do not manage parolees from the state correctional system. Instead, those who are paroled from these longer terms of incarceration are supervised by a separate state agency, known as the Pennsylvania Board of Probation and Parole (PBPP).

At the time when the collaboration between Philadelphia's APPD and the University of Pennsylvania began, the agency was managing a majority of its offenders using a "one size fits all" supervision strategy. Each offender, regardless of their conviction offense or prior criminal history, was subject to roughly standardized conditions and reporting requirements. While APPD possessed some tools to assess each offender's risks and needs, these measures were considered too basic, cumbersome, and inaccurate to determine how much supervision should be provided in individual cases. In some instances, judicially-ordered conditions mandated that certain offenders receive specific amounts and forms of supervision. Usually, however, these judicial instructions were absent, and most offenders were simply placed in the general supervision pool.

Offenders in general supervision were largely burdened with the same requirements, regardless of their prior criminal history, or the amount of risk they presented for future offending. In most instances, these offenders were required to report in person once a month, when they would meet with their supervising officers for approximately 10-15 minutes. Thus, the bulk of the agency's offenders were supervised under a strategy that mandated only 2½ hours of interaction per year. The department's leaders expressed a strong desire to reform this policy, and expressed their desire to focus more supervision on those with the largest risk of future violence, while devoting far less resources to those who presented little or no risk of reoffending.

The initial goal of the partnership between the APPD and the University was therefore to produce a method of forecasting which probationers and parolees presented the largest risk of committing new offenses while under the department's supervision. Through these forecasts, the partnership hoped to achieve its overarching ambition of providing varying levels of supervision and different forms of treatment to offenders who presented different levels of risk. Over time, the University's department grew to include a number of scholars whose skills and interests were particularly suited to the Penn-APPD partnership. These additions included Professor Richard Berk, who joined the faculty in 2006, bringing with him a detailed knowledge of statistical learning procedures, and in particular a forecast modeling technique known as "random forests."

Eventually, the partnership constructed three different models that were put into daily use within the APPD, making on-demand, live forecasts of every single new probationer who reported for supervision. These models, along with the software written to execute them, are tasked with extracting all the necessary data concerning each probationer's past and current circumstances, and then use these values to make the several thousand comparisons required to come up with a unique prediction for each new incoming case. In just 10 or 15 seconds, the department has access to a valid and reliable forecast that it can use to govern the type of supervision that the offender should receive. Since the live forecasting tool was first installed in

2009, nearly 115,000 new case starts have been run through one of three live models, at a rate of almost 47,000 forecasts per year.

While the particular details of random forest models will be described later, the core requirements of creating one of them are generally similar to what one would need to do to create any predictive statistical model. A set of "predictors" – values that represent the conditions that exist at a specific point in time – are defined and collected. For forecasting criminal behavior, these predictors are generally drawn from each offender's past, and can include anything which seems, at the time, theoretically relevant to future offending. The predictor set will therefore usually include measures such as prior arrests, present age, age of onset, previous incarceration, and the neighborhood where the offender currently resides. These predictors are then used to forecast a single "outcome," which represents the offender's criminal behavior after this same specific point in time.

Using a set of observations that are old enough for their outcomes to be fully known, the predictors and outcomes are then combined into a "construction sample" (also known as a "training sample"). For random forest models in particular, it is best if the construction sample is as large as possible, so that even the most rarely-occurring outcomes are sufficiently represented within the data. Every available case can be used to build the model, provided that the values of all predictors are known and are not missing in the data. Unlike other modeling techniques, there is no pressing need to divide the construction sample in two and retain a portion of the observations as a "validation sample" that is later used to test (but not build) the model. The random forest procedure essentially does this on its own, randomly selecting different portions of the construction sample to validate each individual micro-model (known as a "tree;" described below) within the larger overall model.

With these key terms defined, we turn to the question of how a municipality could begin to construct a similar model for its own use.

Unit of Prediction and Time Horizon

It is important to remember that, even though these risk forecasts focus on the behavior of individual offenders, the unit that is really being subjected to forecasting is not the individual offender, but instead that offender's behavior over a specific period of time. Two fundamental decisions must therefore be made at the start of any criminal forecasting effort, namely when this time period begins, and how long it lasts.

The first step in constructing a new model is defining the "unit of prediction" that will be used. In basic terms, this means choosing the moment during the lifespan of an offender's criminal case when the forecast should begin. This could be the moment when bail is set, when prosecutors decide whether to proceed with the case, when sentence is being determined, when the convicted offender first moves into the state correctional system, when parole decisions are being made, when the offender reports for community supervision, or any other moment that makes sense for a particular situation. Anything that has happened prior to this point in time can be used, if the data are available, to form predictor values, while everything that takes place afterwards is a potential outcome that can be forecasted. In essence, this means that each forecast makes predictions about the unit of interest (e.g., each bail decision), and not about the offender generally.

Once the starting point of each forecast is determined, the second step is to decide the "time horizon" that should be applied to each forecast. This time horizon determines both how long each forecast can be considered valid, and how the data needed to construct the model will be defined. As always, however, there are tradeoffs when choosing a time horizon. It may be tempting to choose the longest time horizon possible, in hopes that doing so will allow the model to forecast as many negative outcomes as possible. For example, a parole board constructing a forecasting model in 2012 may feel that it needs to predict any serious offending by potential parolees for at least 10 full years after they return to their communities from prison. In order to construct such a model, however, the board would be forced to use a very limited amount of data from the rather distant past. Only parole cases which had already completed 10 full years of post-parole time could be used to create such a model, because those are the only cases in which the offenders' behavior over the full time horizon are conclusively known. In this example, that means that the model would be constructed using offenders who were granted parole in 2001 and earlier.

Clearly, a great deal of things are likely to have changed between 2001 and today. Today's parolees will be released into very different economic circumstances, the board's own procedures and guidelines are likely to have changed, supervisory techniques have probably been altered, and the machine-readable data that is available on cases from more than a decade in the past is almost certainly different from what the board gathers, records, and has available to it today. There is little reason to think that forecasts based on what transpired 10 years ago will be equally valid and accurate in today's circumstances. For this reason, and a variety of others, it may make more sense to focus on a shorter time horizon, and use more contemporary data to produce the model.

In the case of Philadelphia's probation forecasting models, the unit of prediction was defined as the start of a new term of probation or parole supervision, and the time horizon was set at two years from this initiating event. Each forecast made by one of Philadelphia's forecasting models was created based upon a single instance of probation or parole. For the sake of brevity, these units will be referred to as "probation cases" or "probation starts" here, even though a sizable minority of offenders were on county-level parole after release from short prison sentences. In essence, the models did not produce forecasts about offenders, but instead made forecasts on individual probation cases. If an offender arrived for forecasting after being sentenced on multiple cases, each case would receive a unique forecast. Once live forecasting began, a new prediction was obtained every time an offender began a new instance of probation, and each of these predictions looked forward for two years.

Although simple in concept, this plan often resulted in many overlapping forecasts. Each individual offender can be supervised by APPD on many different occasions over the course of their lifetime. Within a single period of continuous APPD supervision, moreover, an offender can often have multiple different arrests, court cases, and sentences that resulted in the offender's assignment to APPD for supervision. Even when a given offender was already under supervision due to earlier cases, the start of each new case – perhaps brought about by a new conviction and sentence, a new county parole release, or the start of a previously-sentenced period of supervision after a period of state incarceration – resulted in a new forecast.

Because of these complications, it is entirely possible for some offenders to have had multiple, overlapping forecasts active at any one time. An offender who began three new probation cases over a single 16-month period, for example, would have received three different forecasts, and each of these forecasts would cover its own unique 2-year time horizon. It is also possible, therefore, for these multiple forecasts to have conflicting results, and for individual offenders to have had multiple different levels of forecasted risk active at the same time.

Outcomes Forecasted by the Models

Once a model's unit of prediction and time horizon have been determined, the next step is deciding what types of outcomes the model should be set up to predict. One fundamental part of this task is determining how many different forecasted risk groups the model should create. Unlike many more traditional regression techniques, the random forest models used here are not limited to simple binary outcomes, with only two possible results. Instead, these models are capable of dividing the forecasts into three or more different forecasted risk groups. As long as the outcome categories are mutually exclusive, and are capable of being defined within the construction data, it is quite possible to use one single model to sub-divide cases into many different risk groups.

The Philadelphia forecasting effort provides an excellent example of this flexibility, in that it originally began by producing traditional binary-outcome models, and later moved to using three different forecasted risk categories. When the collaboration between the University of Pennsylvania and the APPD started, the initial efforts at forecasting – none of which were ever employed for the live forecasting of incoming probationers – focused only upon the prediction of what was referred to as "murderous conduct". Offenders were placed into one of just two risk categories. Those who were forecasted to be charged with either murder or attempted murder, within two years of their probation case start, were placed in the highest risk category, while all other case starts were placed into a less-risky category.

This simplistic two-category division had many flaws. Most notably, it placed a number of offenders who had committed very serious offenses short of "murderous conduct" – including aggravated assault, robbery, rape, and other sexual offenses – into the lowest risk category. When the current project began, the partnership's focus shifted to the providing live forecasting for all new probationers at the time of intake. Now that the modeling was no longer a simple academic exercise, and would affect the levels of supervision given to actual probationers, the APPD expressed a strong desire to include these other serious offenses in the highest risk category. After substantial analysis of the construction data (i.e., the data that would eventually be used to build the model), the partnership arrived at the following three outcome categories, based on the types of offenses that the offenders were known to have committed during the standardized two-year follow-up period:

- **High Risk**; meaning that the offender was predicted to commit at least one serious offense – defined as murder, attempted murder, aggravated assault, robbery, or a sexual crime – during the first two years after their case start date;
- **Moderate Risk**; meaning the offender was predicted to commit only non-serious offenses during the first two years after their case start date, or;
- **Low Risk**; meaning that the offender was not predicted to commit any new offenses, of any kind, during the first two years after their case start date;

While these three outcome categories seem fairly easy to conceptualize and define, they were not created arbitrarily. Each one of these definitions had to be tested to see how large the three actual risk groups were (i.e., how many of the construction sample's new case starts fell into each category, based on the offenders' actual behavior). These numbers then had to be compared to APPD's own estimates to how many offenders the agency could ultimately supervise in each of the forecasted risk groups. For obvious reasons, the model would not be very useful if more than half of the incoming cases were ultimately forecasted as "High Risk," since no agency could respond effectively when the majority of its caseload was given the highest priority.

Instead, APPD's leaders spent a number of weeks examining their caseloads and staffing levels. They also spent a great deal of time planning the supervision procedures that were to be used for each of the three risk levels, and estimating how large the caseloads could be for the officers who would work in each of the three risk-stratified divisions. After extensive analysis, the department determined that it could handle, at a maximum, approximately 15% of its caseload in the forecasted "High Risk" category, with another 25-30% being forecasted as "Moderate Risk". The remaining 55-60% of cases were to be placed in the forecasted "Low Risk" group, where they would be given the lowest level of supervision.

In practical terms, these real-world limitations on the distribution of predicted outcomes meant that those case starts which involved actual "High Risk" offenders (i.e., those who actually committed a “serious” offense) needed to be defined in such a way as to be somewhat smaller than the 15% of total caseload that APPD thought it could handle in the predicted High Risk category. Since only 15% of new case starts could reasonably be allocated to a forecast of High Risk, then even fewer of the observations in the construction sample – perhaps as few as 10% of the observations – could be defined as actual High Risk outcomes. To understand why this limitation exists, we must first consider the types of errors that forecasting models can make, and how different kinds of errors are more costly than others.

Errors and Costs

Whenever any kind of forecast is made about future events, there is always a chance that the prediction will be wrong. In the most basic terms, there are two different kinds of errors that can occur in this situation. A “false positive” error takes place when a certain event is forecasted to occur, but this event does not come about. An example of a false positive in Philadelphia’s risk forecasting system would be those case starts that are predicted to be High Risk, but the offenders reach the end of the two-year period without committing a new serious offense. The second type of error is known as a “false negative”. False negatives occur when a specific outcome is forecasted to ***not*** take place, but this event ends up arising anyway. In Philadelphia, one very important form of false negative are those cases that are forecasted to be something other than High Risk (i.e., they are predicted to be in either the Low or Moderate categories), but the offender in question actually ends up being charged with a new serious offense before his two-year time horizon comes to a close.

The full range of possible High Risk forecasting outcomes, based on the data from Philadelphia’s most recent model, are illustrated in Table 1. In statistical terms, this table is a simplified form of the “confusion matrix” that was formed from the construction sample when the most recent model was created. Since the matrix stems from construction data, every observation included in the table has both a forecasted result and a known actual outcome available for analysis. This matrix simply compares the outcomes that were forecasted to take place with those that actually occurred during the first two years after the offenders began their term of APPD supervision. For simplicity’s sake, only the High Risk forecasts and actual outcomes are described in Table 1. Those observations which fell into the Moderate and Low risk categories are collapsed into a single category of “Non-High” risk case starts.

Table 1: Simplified confusion matrix for the most recent Philadelphia forecasting model (i.e., Model C), based on construction sample

	Actual	Actual		
	High	Non-High	Totals	Percent
Forecast High Risk	A 7,112	B 11,700	18,812	15.7%
Forecast Non-High	C 4,468	D 96,655	101,123	84.3%
Totals	11,580	108,355	119,935	
Percent	9.7%	90.3%		

The forecasted outcome was correct in the two green cells of the table (labeled A and D). Those in cell A are "true positives" (i.e., they were predicted to be high risk, and actually did commit a new serious offense), while those in cell D are "true negatives" (i.e., they were forecasted to refrain from serious offending, and actually did so). When it comes to distinguishing those who go on to commit serious offenses from those who do not, this most recent model does remarkably well. Fully 86.5% of the construction sample observations were correctly forecasted with respect to being either High Risk vs. something other than High Risk. Nevertheless, errors are unavoidable in any forecasting effort. These forecasting errors are listed in the red cells (B and C), with B containing the false positives for the High Risk predictions, and C containing the false negatives.

While errors in forecasting are inevitable, these two forms of High Risk forecasting errors are far from equally desirable. In one case (cell B), we have offenders who have been labeled as High Risk, but who have not gone on to commit any new offenses. These offenders will end up being more closely supervised than their actual behavior would require, and thus represent (to a certain degree) an unwarranted expenditure of the agency's resources. On the other hand (cell C), we have offenders who have been incorrectly forecasted as being something other than serious offenders – and who will be less closely-supervised as a result – who then go on to be charged with a serious crime within two years of starting their new probation case. With public safety as one of their agency's core functions, it was clear to APPD's leadership that false negative High Risk forecasts were vastly more of a concern than false positive ones. To put the situation in economic terms, failing to predict an actual serious offender was deemed to be a much more costly mistake than incorrectly supervising an offender as High Risk.

One important advantage to random forest modeling is its ability to take the differential costs of these errors into account when the model is being constructed. The challenge here, however, is that the numerical ratio of these costs must be provided (or at least approximated) before the model can be constructed. It is not enough to simply state that false negatives are generally more costly than false negatives. Instead, an actual value must be provided. Before model construction can begin, someone must answer the following question: Precisely how much more costly is it to mistakenly classify a future serious offender into a lower risk category, as compared to the costs of supervising someone who is actually a non-serious offender as though they really were High Risk?

An example from the early days of model development in Philadelphia helps to illustrate how difficult it can be to come up with appropriate values. When asked to provide an initial estimate of the cost ratio (i.e., false negative to false positive) for High Risk forecasts, the APPD leadership initially provided a value of 10.0. Missing one actual serious offender, in other words, was determined to be equal in cost to mistakenly supervising 10 non-serious offenders under the High Risk supervision protocol. After constructing a model with this cost ratio, however, the results presented a very obvious problem. Because this early model used such an extreme cost ratio, it placed far more offenders placed into the forecasted High Risk category – the vast majority of whom were actually Moderate or Low Risk – than the department could ever hope to manage.

After a few iterations of model building, however, it became easier to determine a cost ratio that could ensure that false positives occurred more often than false negatives, while not presenting APPD with a forecasted High Risk caseload that was larger than it could reasonably handle. The results can be seen in Table 1. Based on the data used to construct the model, there were 11,700 false positives for High Risk (cell B), as compared to just 4,468 false negatives (cell C). The final cost ratio, therefore, turned out to be approximately 2.6; each false negative was deemed to be slightly more than 2½ times more costly than each false positive, and false negatives therefore occur less often.

One consequence of this kind of cost ratio, however, is a certain degree of inflation in the overall size of the forecasted High Risk group when it is compared to the actual High Risk group. If false positives will occur more often than false negatives, then the forecasted High Risk group must almost certainly be larger than the number of case starts that go on to produce an actual High Risk result. Again, the implications can be seen in Table 1. Within the construction sample, just 9.7% of the case starts actually led to a new serious offense within the two-year time horizon, but fully 15.7% of the sample ended up being forecasted as High Risk.

In the end, therefore, the desired cost structure must be balanced against every other decision that goes into creating this kind of model. Once it became clear, for example, that the department's ability to supervise High Risk offenders was capped at around 15%, both the time horizon and offenses that defined "High Risk" needed to be reconsidered. If a 5-year time horizon had been desired instead of a 2-year one, then the size of the actual High Risk group would have increased to a value in excess of this 15% limit. A random forest model could still have been constructed, but it would have required a cost structure that made false negatives more common than false positives, which was deemed unacceptable. Similarly, other less-serious offenses – such a burglary and simple assault – could have been included in the definition of "High Risk," but that would also have enlarged the actual High Risk group, and would have required the use of either a shorter time horizon, a less-differentiated cost structure, or both.

There is therefore no one right answer in determining the unit of prediction, the time horizon of the forecasts, the ways in which the outcomes are defined, or the cost structures that should be provided to govern the construction of a new model. All of these factors influence one another, and any adjustment in one could easily require a change to all of the others. Each municipality that sets out to construct these kinds of forecasting models must make its own set of decisions about how these elements should be defined, and each model that is constructed will be a unique reflection of the aspects that are considered important in each different context.

Three Different Live Models

Over the course of this project, Philadelphia has constructed several dozen different forecasting models using the random forest technique. Most of these models were built to test new ideas, balance the desired cost ratios, and examine how different sets of predictor variables worked with one another. Thus the majority of these efforts were never used on a "live" basis to predict the agency's incoming caseload. Only three different models reached the required standard, and were approved for use with the live forecasting system.

In many ways, these three models are more similar than they are different. All of them used the same unit of prediction, time horizon, and had largely equivalent cost structures. The largest differences centered upon the predictor variables that were used by the models to produce the forecasts. Even here, however, there was one very important similarity across all three sets of predictor variables. All of the predictors used within these models were required to stem from data sources that were readily accessible, in machine-readable form, within the APPD's data network.

Many other possible predictor variables – including any abuse and neglect history from the offenders' childhoods, foster care placements, tax and employment histories, and juvenile incarceration histories – were, at least in theory, available from various departments in Philadelphia city government. Many of these external predictors might have proven very useful in improving the accuracy of the model. Despite these potential advantages, however, these data sources were not immediately accessible for database queries originating from the department's live forecasting software. Because our project's focus was on the development and installation of a model that could be used at the instant when a new offender arrived at intake, these data sources – regardless of what benefits they could lend to forecasting accuracy – were simply not usable for our purposes. Thus forecasting accuracy was (potentially) sacrificed, at least to some extent, in the name of usability and speed.

One consequence of this focus on immediately-available data sources was a narrow geographic focus. All of the data used to construct these models were based upon the criminal history data for Philadelphia alone, since data from other states – and even from other Pennsylvania jurisdictions – were simply not available for instantaneous access within the APPD's data network. Although it was possible to obtain samples of such outside data sources for the purposes of constructing a model, getting immediate access to the them at the moment when a new offender arrived at APPD's intake unit, and live forecast was needed, would have required a great deal of effort. Moreover, testing revealed that the inclusion of these outside predictors did not seem to appreciably increase the accuracy of the models' forecasts.

This geographic limitation affected these models in terms of both the predictors they employed to make their forecasts, and the outcome offenses that they predicted. In other words, all three of the live models used the offenders' prior criminal histories in Philadelphia to forecast any future offenses that they were likely to commit within Philadelphia. Thus the forecasts produced in Philadelphia don't necessarily indicate each offender's overall or universal level of risk. Offenders who represent a serious danger outside of the city limits could very easily be forecasted as Low Risk within these boundaries, particularly if they usually live, work, and offend elsewhere.

Another important similarity across all three models was their use of charge counts – as opposed to conviction counts – to represent each offender's criminal history. In some ways, the decision to use charging data in this manner was a difficult one. Philadelphia's APPD is run by the court system and its leadership reports to the judges of the First Judicial District (FJD) of Pennsylvania instead of to the city's executive branch. Understandably, therefore, there is a certain desire to structure supervision around what the offenders were convicted of in court, instead of the offenses that were merely charged (but not proven) with committing. In statistical terms, however, the charging data was thought to contain more information about the offenders' backgrounds, notably due to the plea bargaining and trial processes, than data limited solely to convictions.

The first of the three models used in live forecasting, known as Model A, was constructed in January 2009, and was used for incoming case forecasting at APPD from March 2009 through April 2010. To construct this model, data were drawn from all APPD case starts between January 1, 2002 and December 31, 2005. For each of the probation or parole cases which began during this time, a large set of predictor values were drawn to represent each offender's criminal charge history prior to the case start date, the amount of time they had spent on probation and in the local county-level prison system in the past, their residential zip code on the day that the case began, and a variety of other factors. The final set of predictors used in the model included 34 different values. In addition, any criminal charges for offenses which took place during the subsequent two years were categorized and counted to form the outcome variable for each case. Although 70,728 probation cases were available in the data (spread across a total of 48,529 different offenders), the model was constructed using a sub-sample of 50,000 case starts. In total, Model A included 4.57 million different decision points, although only a small fraction of them would be used to produce the forecast for any particular case.

The second model (Model B) was constructed in December 2009, and was used for live case forecasting at APPD from April 2010 through November 2011. The sample used to construct the model was pulled from all of the agency's new case starts between January 1, 2002 (i.e., the same beginning date as Model A) and December 31, 2006 (i.e., one year later than Model A). The primary reason for building this new model was to make use of juvenile offending data, which had become available after negotiations with other agencies in Philadelphia government. Due to the addition of these juvenile offending variables, the total number of predictors increased to 48 different values. The construction data included a total of 94,653 probation starts (60,373 offenders), from which a sample of 50,000 were used to build Model B. In overall size, this model was roughly a quarter of the size of its predecessor, containing 1.20 million decision points.

The third and most recent model (Model C) was built in August 2011, and was installed for live forecasting in November 2011. The sample used to construct it consisted of all 119,988 APPD cases which began between January 1, 2002 and December 31, 2007 (i.e., two more years of data compared to Model A, and one additional year compared to Model B). Since more than a third of offenders in this sample had more than one case start in the data, these 119,988 case starts are spread across just 71,976 different offenders.

While the primary reason for replacing Model B with Model C was the introduction of a new database system for the juvenile offending data, this model also included a number of new features and techniques. Instead of using a sub-sample of probation case starts, as had been used in the past, the entire sample was used to construct the model. In addition, the total number of predictors was greatly reduced, so that just 12 different values – those which were shown to have the most influence on the desired outcomes – were used to produce the forecasts. Another important change was shift away from using census data to represent the conditions in each offender's residential zip code. Instead, the 29 most common zip code values for APPD offenders were entered under a single categorical variable, so that the overall effect of living in one of these locations could be estimated and used to influence the forecasts made by the model. The result was a much larger than either of the two previous models, with 8.74 million different decision points across all of Model C.

Predictors Used

In the course of developing the three live forecasting models, 53 different predictor variables have been used to predict future offending by Philadelphia's incoming probationers. Different types, numbers, and combinations of these predictors have been featured in all three of the models. Some predictors have been used only once, and then discarded, while others have played a role in every model created since the beginning of the project. A checklist of all these predictor variables is presented in Table 2. To save space, the names of these predictors have been abbreviated somewhat. To better understand what each of these predictors really represents, however, the following descriptions are likely to be helpful:

ProbationStartAge. The offender's age at the start of the new probation case.

Table 2: Predictor variables used to construct the three live forecasting models

Predictor Variable	Model A	Model B	Model C
ProbationStartAge	✓	✓	✓
CalculatedGender	✓	✓	
ZipBase5Top29			✓
ZipPopulation	✓	✓	
ZipHouseholdIncome	✓	✓	
ZipHouseValue	✓	✓	
ZipPersonsPerHousehold	✓	✓	
ZipCityLimitDistance	✓	✓	
ZipOutsideCityLimits	✓	✓	
FirstAdultAnyChargeAge	✓	✓	✓
FirstAdultViolenceChargeAge	✓	✓	✓
FirstJuvAnyChargeAge		✓	✓
FirstJuvViolenceChargeAge		✓	
InstantMurderChargeCount	✓		
InstantSeriousChargeCount		✓	✓
InstantViolenceChargeCount	✓	✓	
InstantSexualChargeCount	✓	✓	
InstantPropertyChargeCount	✓	✓	
InstantFirearmChargeCount	✓	✓	
InstantDrugChargeCount	✓	✓	
InstantProbationSentenceCount		✓	
InstantProbationDaysConcurrent		✓	
InstantIncarcerationSentenceCount		✓	
InstantIncarcerationDaysConcurrent		✓	
PriorAdultAnyChargeCount	✓	✓	✓
PriorAdultUcrPersChargeCount	✓		

CalculatedGender. The offender's gender, as calculated from all available data sources. This value is available from more than one of the databases used to produce predictors the model. Most of the time, these sources all agree on whether the offender is male or female. When disagreement occurs, or when some of these values are missing, this value is calculated by using the gender value from the criminal records data (if available), and the value from the probation case management system where the criminal records value is missing.

ZipBase5Top29. This variable forms a categorical list of 31 distinct values to indicate the 5-digit zip code where the offender was residing at the time that the instant probation case began. These values are made up of the 29 most prevalent valid zip code values among probation case starts, along with 2 other coded values to indicate whether the offender was residing in some other valid zip code. If the offender was living in one of the 29 most-frequent zip codes – all of which are located within the city limits of Philadelphia – this variable is coded with that offender's five-digit zip code value. When the offender did not reside in any of these 29

Table 2 (continued): Predictor variables used to construct the three live forecasting models

Predictor Variable	Model A	Model B	Model C
PriorAdultSeriousChargeCount		✓	
PriorAdultViolenceChargeCount	✓	✓	✓
PriorAdultSexualChargeCount	✓	✓	✓
PriorAdultSexRegChargeCount	✓	✓	
PriorAdultPropertyChargeCount	✓	✓	
PriorAdultWeaponChargeCount	✓	✓	
PriorAdultFirearmChargeCount	✓	✓	
PriorAdultDrugChargeCount	✓	✓	
PriorAdultDrugDistChargeCount	✓	✓	
PriorJuvAnyChargeCount		✓	
PriorJuvSeriousChargeCount		✓	
PriorJuvViolenceChargeCount		✓	
PriorJuvSexualChargeCount		✓	
PriorJuvPropertyChargeCount		✓	
PriorJuvWeaponChargeCount		✓	
PriorJuvFirearmChargeCount		✓	
PriorJuvDrugChargeCount		✓	
PriorJuvDrugDistChargeCount		✓	
PriorAdultSeriousChargeLatestYears	✓		
PriorSeriousChargeLatestYears		✓	✓
PriorProbationCount	✓	✓	
PriorFailureToAppearCount	✓	✓	
PriorAbsconderCount	✓	✓	
PriorJailStays	✓	✓	✓
PriorJailDays	✓	✓	✓
PriorConfinementSentenceCount	✓		
PriorIncarcerationSentenceCount	✓	✓	

specific zip codes, the value is coded as “99998” when the offender lived elsewhere inside the city limits, and “99999” when the offender lived elsewhere outside the city limits. Offenders with missing or invalid zip code values are excluded from the model construction data.

ZipPopulation. The total population, based on 2000 census data, in the zip code where the offender was residing at the start of the new probation case. This predictor, like all of the zip code-based demographic values, was used only for Models A and B.

ZipHouseholdIncome. The average household income in the offender’s home zip code.

ZipHouseValue. The average house value in the offender’s home zip code.

ZipPersonsPerHousehold. The average number of persons residing in each household in the offender's home zip code.

ZipCityLimitDistance. The number of statute miles between the offender's home zip code and the Philadelphia city limits. Coded as zero for all observations where the offender resided within the city.

ZipOutsideCityLimits. A binary variable which indicates whether the offender's home zip code is outside of the Philadelphia city limits.

FirstAdultAnyChargeAge. The offender's age at the time of the first offense which resulted in charges in adult criminal court.

FirstAdultViolenceChargeAge. The offender's age at the time of the first violent offense which resulted in charges in adult criminal court. When the offender has never been charged as an adult with a violent offense, this value is coded as 100 years.

FirstJuvAnyChargeAge. The offender's age at the time of the first offense which resulted in charges in juvenile court. When the offender no record of juvenile offending, this value is coded as 100 years. This variable is used only in the Models B and C, and reflects the addition of juvenile predictors to the model.

FirstJuvViolenceChargeAge. The offender's age at the time of the first violent offense which resulted in charges in juvenile court. When the offender no record of violent juvenile offending, this value is coded as 100 years. This variable is used only in the Model B.

InstantMurderChargeCount. The total number of charges for murder or attempted murder that appear in the court records for the instant case. The instant court case is the one that resulted in the offender being placed on APPD supervision for this instance of probation or parole. This variable is used only in the Model A. It was dropped from the later models because only a very small number of cases which involve charges this serious result in the offender being placed on APPD supervision.

InstantSeriousChargeCount. The total number of charges for serious offenses – defined as murder, attempted murder, aggravated assault, robbery, and sexual crimes – in the instant case. This variable is used only in the forecasting models from Model B onward. It replaces the number of instant charges for murder or attempted murder.

InstantViolenceChargeCount. The total number of charges for violent offenses in the instant case. Violent offenses include all serious offenses, as well lesser crimes such as simple assault.

InstantSexualChargeCount. The total number of charges for sexual offenses in the instant case.

InstantPropertyChargeCount. The total number of charges for property offenses in the instant case.

InstantFirearmChargeCount. The total number of charges for firearm offenses in the instant case.

InstantDrugChargeCount. The total number of charges for drug offenses in the instant case.

InstantProbationSentenceCount. The total number of sentences to probation that appear in the court records as a result of the instant case. This variable was used only in Model B. It was added, along with the other instant sentencing variables, to provide an indication of how dangerous the sentencing judge thought the offender to be.

InstantProbationDaysConcurrent. The maximum number of days sentenced to probation as a result of the instant case, assuming that all sentences are to be served concurrently. This variable is used only in Model B.

InstantIncarcerationSentenceCount. The total number of sentences to incarceration as a result of the instant case. This variable is used only in Model B.

InstantIncarcerationDaysConcurrent. The maximum number of days sentenced to incarceration as a result of the instant case. This variable is used only in the Model B.

PriorAdultAnyChargeCount. The total number of charges for offenses which were dealt with in adult criminal court, and which took place prior to the start of the new probation case.

PriorAdultUcrPersChargeCount. The total number of charges for Uniform Crime Report (UCR) Part I Personal offenses which were dealt with in adult criminal court, and which took place prior to the start of the new probation case. These offenses include murder, aggravated assault, robbery, and forcible rape. This variable was used only in Model A. It was dropped from later models because it did not include some non-forcible sexual offenses, such as statutory rape, that are included in the models' definition of serious crime.

PriorAdultSeriousChargeCount. The total number of charges for serious offenses which were dealt with in adult criminal court, and which took place prior to the start of the new probation case. This variable is used only in Model B, where it replaced the number of prior charges for UCR personal offenses.

PriorAdultViolenceChargeCount. The total number of charges for violent offenses which were dealt with in adult criminal court, and which took place prior to the start of the new probation case.

PriorAdultSexualChargeCount. The total number of charges for sexual offenses which were dealt with in adult criminal court, and which took place prior to the start of the new probation case.

PriorAdultSexRegChargeCount. The total number of charges for sex offender registration offenses (i.e., violations of the registration requirements in Megan's Law) which were dealt with in adult criminal court, and which took place prior to the start of the new probation case.

PriorAdultPropertyChargeCount. The total number of charges for property offenses which were dealt with in adult criminal court, and which took place prior to the start of the new probation case.

PriorAdultWeaponChargeCount. The total number of charges for weapon offenses which were dealt with in adult criminal court, and which took place prior to the start of the new probation case.

PriorAdultFirearmChargeCount. The total number of charges for firearm offenses which were dealt with in adult criminal court, and which took place prior to the start of the new probation case.

PriorAdultDrugChargeCount. The total number of charges for drug offenses which were dealt with in adult criminal court, and which took place prior to the start of the new probation case.

PriorAdultDrugDistChargeCount. The total number of charges for drug distribution offenses which were dealt with in adult criminal court, and which took place prior to the start of the new probation case.

PriorJuvAnyChargeCount. The total number of charges for offenses which were dealt with in juvenile court, and which took place prior to the start of the new probation case. This variable is used only in Model B, and reflects the addition of juvenile predictors to the model.

PriorJuvSeriousChargeCount. The total number of charges for serious offenses which were dealt with in juvenile court, and which took place prior to the start of the new probation case.

PriorJuvViolenceChargeCount. The total number of charges for violent offenses which were dealt with in juvenile court, and which took place prior to the start of the new probation case.

PriorJuvSexualChargeCount. The total number of charges for sexual offenses which were dealt with in juvenile court, and which took place prior to the start of the new probation case.

PriorJuvPropertyChargeCount. The total number of charges for property offenses which were dealt with in juvenile court, and which took place prior to the start of the new probation case.

PriorJuvWeaponChargeCount. The total number of charges for weapon offenses which were dealt with in juvenile court, and which took place prior to the start of the new probation case.

PriorJuvFirearmChargeCount. The total number of charges for firearm offenses which were dealt with in juvenile court, and which took place prior to the start of the new probation case.

PriorJuvDrugChargeCount. The total number of charges for drug offenses which were dealt with in juvenile court, and which took place prior to the start of the new probation case.

PriorJuvDrugDistChargeCount. The total number of charges for drug distribution offenses which were dealt with in juvenile court, and which took place prior to the start of the new probation case.

PriorAdultSeriousChargeLatestYears. The number of years since the offender's most recent serious offense which resulted in charges in adult criminal court. When the offender has never been charged as an adult with a serious offense, this value is coded at 100 years. This variable is used only in the Model A, and was amended in later models to include juvenile offending information.

PriorSeriousChargeLatestYears. The number of years since the offender's most recent serious offense, regardless of whether that offense was dealt with juvenile or adult criminal court. When the offender has never been charged with a serious offense, this value is coded as 100 years. This variable was not used until Model B, and reflects the addition of juvenile predictors to the model.

PriorProbationCount. The total number of cases which were placed under APPD supervision prior to the start of the new probation case.

PriorFailureToAppearCount. The total number of bench warrants taken out against the offender, prior to the start of the new probation case, due to a failure to appear in court.

PriorAbsconderCount. The total number of arrest warrants taken out against the offender, prior to the start of the new probation case, due to absconding from supervision.

PriorJailStays. The total number of entries into the Philadelphia County prison system which took place prior the start of the new probation case.

PriorJailDays. The total number of days spent incarcerated in the Philadelphia County prison system prior to the start of the new probation case.

PriorConfinementSentenceCount. The total number of sentences to confinement – which includes incarceration, house arrest, and electronic monitoring – that the offender received prior to the start of the new probation case. This variable is used only in the January 2009 forecasting model. It was dropped from later models because it strongly mirrors the incarceration sentence count variable, discussed below, and added little unique information.

PriorIncarcerationSentenceCount. The total number of sentences to incarceration that the offender received prior to the start of the new probation case.

In addition to the predictor variables described above, all of which were used in at least one of the three live forecasting models, an enormous number of other variables were tried or tested at various points during the current project. One of the many attractive features of random forest modeling is that, unlike traditional regression methods, there is no real limit on the number of predictors that can be included in a forecasting model.

Because of the way that the different predictors compete with one another as these models are constructed, any variables that offer only a weak amount of predictive power simply do not play much of a role in determining the forecasted outcomes. This fact allowed the APPD to include predictors that were important to various stakeholders in the city's criminal justice system, even if these additional predictors didn't seem to make the forecasting results any more accurate. In addition, it's entirely possible that certain predictors may not increase the predictive accuracy in a noticeable way across the entire model, but do prove useful under certain conditions.

For example, tests conducted upon Model B showed that the number of prior warrants issued for absconding from probation did not contribute a great deal of important information, and that the model would have lost less than 2% of its accuracy if this predictor had not been included. But perhaps for some kinds of offenders – those who are younger, have a history of drug-related offending, and have been on probation 3 or more times in the past, for example – a failure to comply with the reporting requirements of probation really is an indicator of a larger set of problems. Random forest modeling allows these kinds of predictors to exert their influence in these potentially-rare situations where they really matter. To understand why, one must understand exactly what one small part of such a model looks like, and how the different predictors combine within the many regression trees that make up an entire random forest.

Trees, Forests, and How the Model Functions

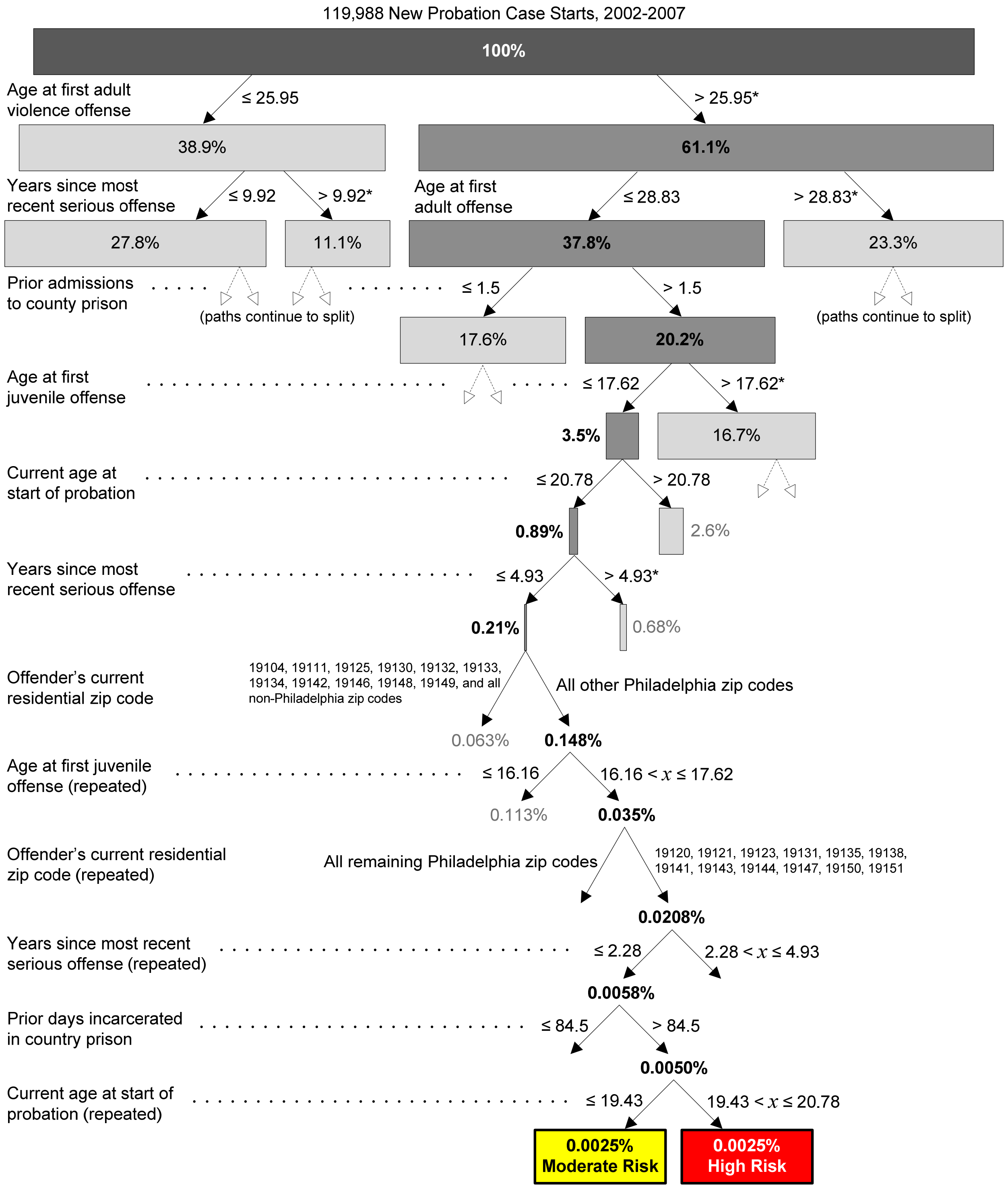
A random forest is not really a single unified model, but could instead be better described as an amalgamation of hundreds of individual regression trees. These trees are formed using a technique called classification and regression trees (CART), which is combined with the power of modern computers in a way that selects predictors at random, repeats the procedure to build several hundred trees, and then allows these randomly-selected predictor sets to average themselves into a single resultant outcome.

Figure 1 illustrates just one possible path that a particular probation case start might take on its way through just one of the many trees (Tree #404) that make up Model C. This path begins with 100% of all 119,988 cases in the model's construction sample. When this tree was created, 3 of the model's 12 possible predictor values were selected at random to compete for the job of splitting this sample[1] into two separate groups. There is no way to know all three of the predictors that participated in this competition, but the winner was the age of the offender at the time of their first adult violent offense, and the optimal split point was computed to be 25.95 years. All of the offenders whose adult violence onset took place prior to this age are moved to the left, while all of those whose onset was later (or who simply had never been charged with such an offense) are moved to the right.

At this stage, the full sample has been divided into two separate "nodes" of the tree. A smaller group of cases with early onset offenders, amounting to 38.9% of the sample, have been placed in the left node, while a 61.1% majority of the remaining cases have moved onto the right node. While this division might seem fairly simplistic, it sets the stage for everything that comes later. From this point onwards, these two nodes are treated as completely separate sub-samples of the construction data. Every new split that stems from the one on the left, for example, will involve only those cases where the offenders began their adult violent offending comparatively late in life (or even not at all). All the decision points that follow each node, in other words, apply only to a very specific sub-sample of the entire construction sample.

The tree shown in Figure 1 now moves onto the next set of divisions. Both the left and the right nodes get their own randomly-selected set of three predictors that compete with one another to split these two sub-samples even further. On the left side, the number of years since the offender committed their most recent serious offense is chosen to split the sub-sample even further. For the right node, an entirely separate competition between a different set of 3 predictors leads to this sub-sample being further divided by the onset age when the offenders committed their first offense (of any kind) as an adult. As shown in Figure 1, the left "daughter" node from this split includes the 37.8% of case starts where the offender began their adult violent offending after 25.9 years of age (based on the first split of the model), but who also started adult offending generally prior to 28.8 years (based on the next split in the chain).

[1] In reality, when this tree was created, the full sample of 119,998 new case starts would not have been used to build it. Instead, a number of the full sample's observations would have been set aside, at random, as a separate validation sample for this one particular tree. These unused cases are referred to as "out of bag" observations, and each individual tree in the forest has its own unique set of them. For ease of exposition, however, Figure 1 is presented as using the 100% of the available sample.


119,988 New Probation Case Starts, 2002-2007
100%
Age at first adult violence offense
≤ 25.95
> 25.95*
38.9%
61.1%
Years since most recent serious offense
≤ 9.92
> 9.92*
Age at first adult offense
≤ 28.83
> 28.83*
27.8%
11.1%
37.8%
23.3%
Prior admissions to county prison
(paths continue to split)
≤ 1.5
> 1.5
(paths continue to split)
17.6%
20.2%
Age at first juvenile offense
≤ 17.62
> 17.62*
3.5%
16.7%
Current age at start of probation
≤ 20.78
> 20.78
0.89%
2.6%
Years since most recent serious offense
≤ 4.93
> 4.93*
0.21%
0.68%
Offender's current residential zip code
19104, 19111, 19125, 19130, 19132, 19133, 19134, 19142, 19146, 19148, 19149, and all non-Philadelphia zip codes
All other Philadelphia zip codes
0.063%
0.148%
Age at first juvenile offense (repeated)
≤ 16.16
16.16 < x ≤ 17.62
0.113%
0.035%
Offender's current residential zip code (repeated)
All remaining Philadelphia zip codes
19120, 19121, 19123, 19131, 19135, 19138, 19141, 19143, 19144, 19147, 19150, 19151
0.0208%
Years since most recent serious offense (repeated)
≤ 2.28
2.28 < x ≤ 4.93
0.0058%
Prior days incarcerated in country prison
≤ 84.5
> 84.5
0.0050%
Current age at start of probation (repeated)
≤ 19.43
19.43 < x ≤ 20.78
0.0025% Moderate Risk
0.0025% High Risk

* including offenders with no prior record within the indicated offense category

Figure 1: An example of one path through one tree in Philadelphia's latest random forest model

The path for this 37.8% subgroup continues onwards, with further splits repeatedly dividing it into smaller and more specific subgroups. These splits proceed, first by using the offender's age at the time of their first juvenile offense (leaving just 3.5% of the sample where the offender's juvenile onset took place at or prior to 17.62 years of age), and then progressing further with splits based on the offenders' age at the time when their probation case began and the number of years since the occurrence of their most recent serious offense. By this point, the sub-sample has shrunk to include just 0.21% of the original 119,998 probation case starts that started the chain. Even within this tiny group of just 253 observations, however, there is still a diversity of different actual outcomes, with 89 actual High Risk outcomes, 98 Moderate Risk, and 66 Low Risk. The overall result from this path of the tree is therefore still unclear, and further splits are required.

The next division in the chain highlights one of the unique features built into Model C. The cases are split using the categorical representation of the offender's residential zip code at the time when their probation case began. In previous versions of Philadelphia's forecasting model, this zip code value was used solely to look up a variety of demographic variables based upon data from the 2000 census. By the time that Model C was developed, however, these data were already more than a decade old. While these 2000 census figures would still work well for use in the older construction sample, they would become increasingly removed from the conditions in these neighborhoods today, and therefore less and less useful for the live forecasting of new probation case starts.

To fix this problem, the APPD-Penn partnership decided to use the actual value of this zip code itself to represent everything about each of these geographic areas, rather than relying on demographic variables to present only a narrow part of this picture. Unlike the other predictors in the model, however, the zip codes are not ordinal in nature, and these divisions could not proceed merely by dividing the sample at a specific numerical split point. Instead, individual zip code values needed to be considered as separate categories, and the splits needed to allow different zip codes to move to different nodes based on the distribution of actual risk groups that they contained. Software limitations required that this division take place using just 29 of Philadelphia's 47 residential zip codes, so the 29 zip codes most common to APPD's offender population were used to make these splits, while all other Philadelphia zip codes were aggregated together into a single dummy value.

In Figure 1, the results of this technique move the observations from 11 of these Philadelphia zip codes (along with those from outside the city limits) to the left, while case starts from all other Philadelphia zip codes proceed to the right. The path then moves along to the next step in the chain, where the offender's juvenile onset age is used to split the sample again, despite the fact that it was already used to do so at an earlier point. Based on this earlier split, all of the offenders in this branch of the tree must already have an onset age that is less than (or equal to) 17.62 years of age. This next split divides this group even further. Those case starts with offenders whose first juvenile offense took place prior to age 16.16 are moved to the left, while those where the offender’s juvenile onset took place between 16.16 and 17.62 years of age are moved to the right.

Additional divisions continue to take place along with path, with the (at this stage rather small) sub-sample split once again by residential zip code, by the number of years since the most recent serious offense (which was also used at an earlier point in the chain), and the total number days that the offender had spent in the local prison system prior to their probation start date. The final split along this branch of this one tree occurs based on the offender’s age at the start of the instant probation case. Since this predictor was already used much earlier along this path, all of the offenders in this penultimate node are already less than (or equal to) 20.78 years of age. This last division narrows this range down even further, with all those younger than 19.43 years sent to the left, while those older than this split point are moved to the right.

The resultant two daughter nodes are referred to as “terminal nodes,” and mark the end points of their own unique paths through this particular CART tree. Any observations that fall into the left terminal node are given a forecasted result of Moderate Risk, while all of those which land in the right terminal node will be categorized as High Risk. By the time we reach these decisions, each terminal node contains just 3 of the 119,998 observations in the construction sample, amounting to just 0.0025% of the sample that we began with. Further splits of these micro-samples are either unnecessary (i.e., all of the observations in the terminal node have the same actual outcome) or impossible (i.e., none of the competing predictor variables are able to form nodes that are any more homogenous in terms of actual outcome).

It is important to note that this one single path through this one tree in the random forest is but a very tiny part of the overall model. In total, the entire path shown in Figure 1, from beginning to end, encompasses just 14 total nodes (12 which are divided into additional nodes, plus 2 terminal nodes). Within the one tree that this path is drawn from, however, there are 17,509 different nodes, and the full model of 500 separate trees consists of 8.74 million nodes. Moreover, even the tiny percentage of cases that fall into the High Risk node at the end of this particular path may not be ultimately forecasted as High Risk by the model as a whole. While the result from this one tree would constitute one “vote” for a High Risk result, the other 499 trees each get a vote as well. When the votes from all 500 trees are counted, whichever outcome has received the most votes will be presented as the forecasted outcome for the entire model. If any of these vote counts are tied, the APPD risk forecasting software defaults to the highest risk category.

The real power of random forest modeling ultimately lies in this extremely large number of separate nodes, along with the random selection of individual predictors to split them. This combination allows the influence of each predictor to be averaged over a wide variety of unique sub-samples throughout the model, and reduces the influence of any one particular tree to just one vote out of hundreds. Even if one particular branch or one entire tree proves to be somewhat inaccurate under certain conditions, therefore, its biases can easily be compensated for by the millions of other paths that cases take through the model as a whole.

The Influence of Predictors on Forecasted Outcomes

Although twelve different predictor variables were used to build Philadelphia's most recent forecasting model (Model C), not all of the variables made an equal contribution to the accuracy of the model's predictions. Figure 2 shows how much each of these variables contributed to the forecasting accuracy of the model as a whole (i.e., the accuracy across all three outcome groups combined). These values were determined by temporarily setting the values of each variable to a set of random numbers, and then observing how much the model's accuracy was reduced when that particular predictor was no longer providing any useful information.

It is clear that some predictors were very important to the overall accuracy of the model. The number of prior stays in the county prison system was of key importance, while the offenders' residential zip codes, the time elapsed since their most recent serious offense, current age, and age at the time of their first adult offense all combined to form a strong second tier of important predictors. The three least-important variables for the model as a whole were the onset age for juvenile offending, the number of serious-crime charges stemming from the case that resulted in the offender being placed on supervision, and the count of prior charges for sexual offenses.

It should be noted that the earliest versions of this model did not include these three variables at all, since they did not appear to contribute much in the way of forecasting accuracy. Nevertheless, the APPD leadership felt that the inclusion of certain predictors was politically necessary in order for them to defend the use of the model to various stakeholders associated with the city's criminal justice system. At a minimum, it seemed necessary to have a least one measure of offending from each offender's juvenile years, at least one indicator of the seriousness of the instant offense, and one predictor which focused exclusively on prior involvement in sexual crime. Because random forest modeling suffers little or no penalty for including additional predictors (however weak they may be) into the model, the desire forof these politically-desirable variables was easy to accommodate.

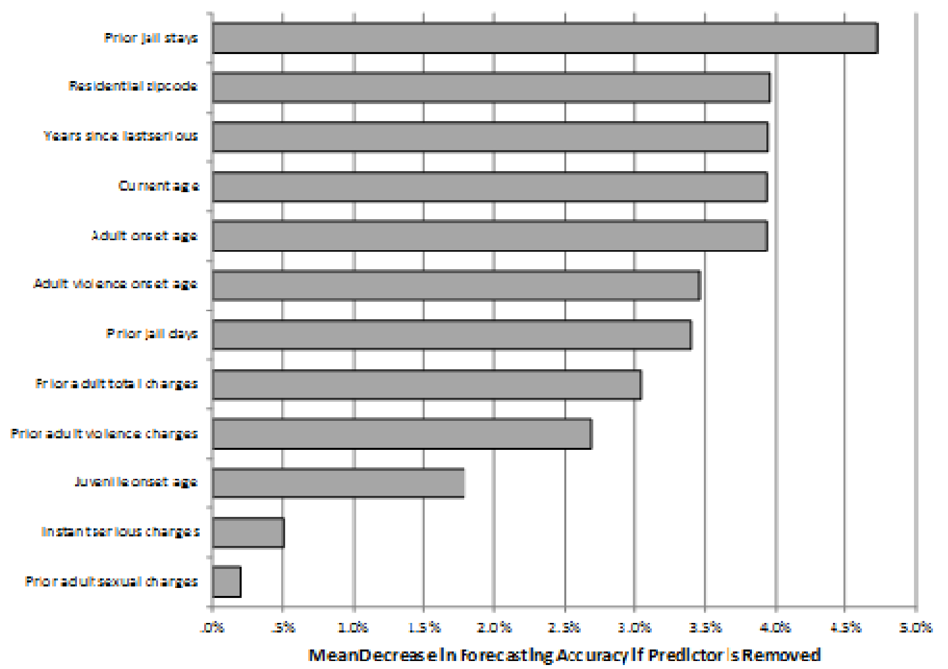


Figure 2: Importance plot for predictors in the latest forecasting model; all three forecasted outcome categories combined

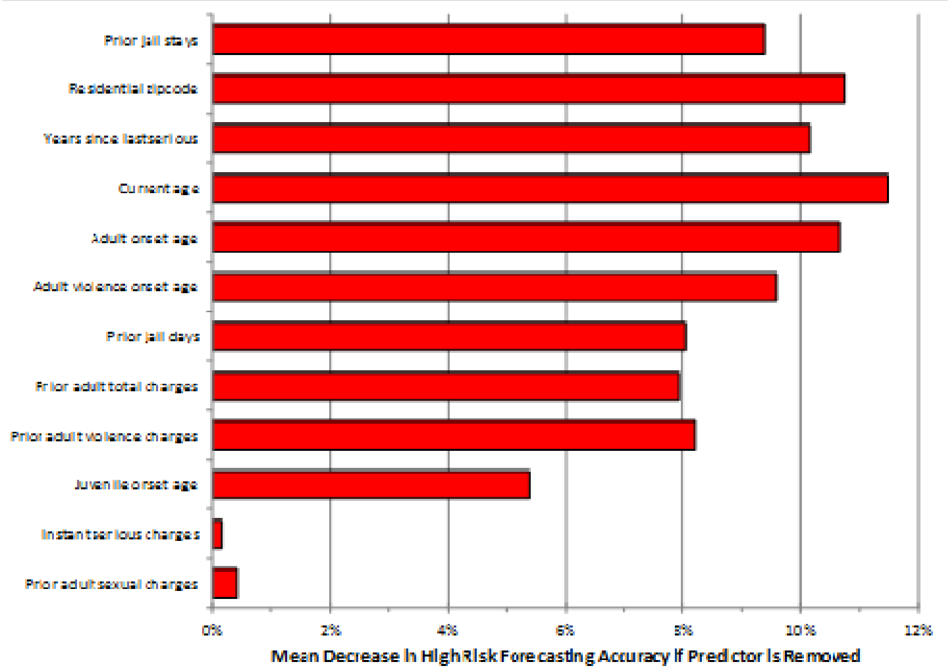


Figure 3: Importance plot for predictors in the latest forecasting model; High Risk outcomes only

While Figure 2 shows the relative importance of each predictor to the model as a whole, each of the three outcome categories – High, Moderate, and Low Risk – has its own set of variables that are uniquely important to forecasting that specific type of risk. Figure 3 shows the same variable importance plot as Figure 2, but limited only to the forecasts of serious offending within the first two years of supervision. Two things are clear from Figure 3. First, the relative order of these predictors in their importance for forecasting High Risk offending is rather different from their importance to the model as a whole. Current age and the ages of adult onset (for both offending generally and for violence) seem much more crucial to forecasting High Risk outcomes – at least in relative sense – than they were when the overall accuracy of the entire model is considered.

Secondly, the entire scale of these importance measures was much greater when it comes to predicting serious offending than it was for all three outcome categories combined. The absence of usable information for the most important predictor in the entire model (i.e., prior jail stays) would have reduced forecasting accuracy by just over 4.5%, while the loss of the most important High Risk predictor (current age) would cause the accuracy of these forecasts to drop by more than 11%. Even the importance of a relatively minor predictor, such as the age of juvenile onset (5.4%), was greater for High Risk forecasts alone than the most important predictor (prior jail stays; 4.7%) was when all three outcomes were combined.

Along with allowing different predictors to play different roles in forecasting different outcomes, another central characteristic of random forest modeling is its use of highly non-linear effects for each individual predictor. Figure 4, for example, shows the bivariate relationship between an offender’s current age and the likelihood that Model C would forecast them as High Risk. As one might expect, the youngest probationers seem to present the biggest danger. A bit more surprising, however, is how quickly the probability of a high risk forecast drops as the offenders get just a few years older. By the time that an incoming probationer turns 27, the likelihood of receiving a High Risk forecast is not appreciably different from that of a 40-year-old. After age 40, however, the amount of risk seems to drop once again, until it reaches a level that is effectively zero at age 50 and beyond.

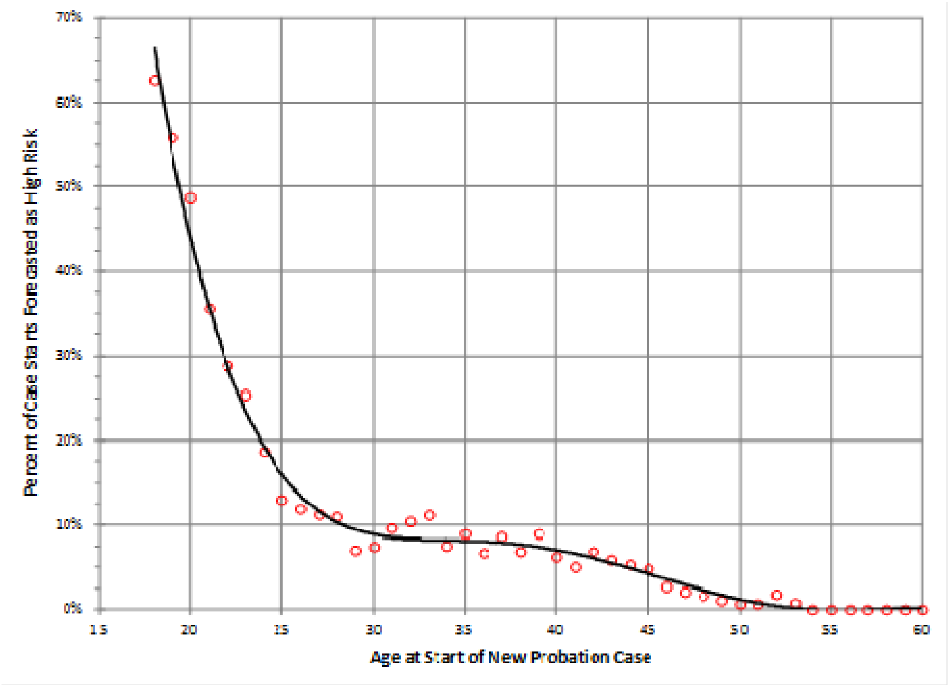


Figure 4: Probability of a High Risk forecast vs. current age in the latest forecasting model (i.e., Model C)

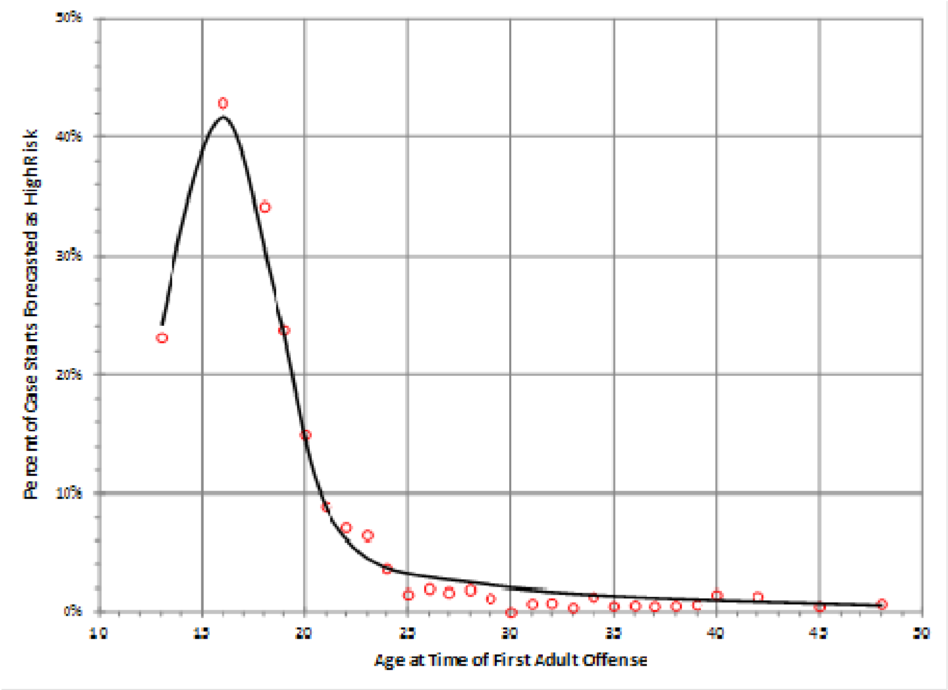


Figure 5: Probability of a High Risk forecast vs. age of adult onset

A

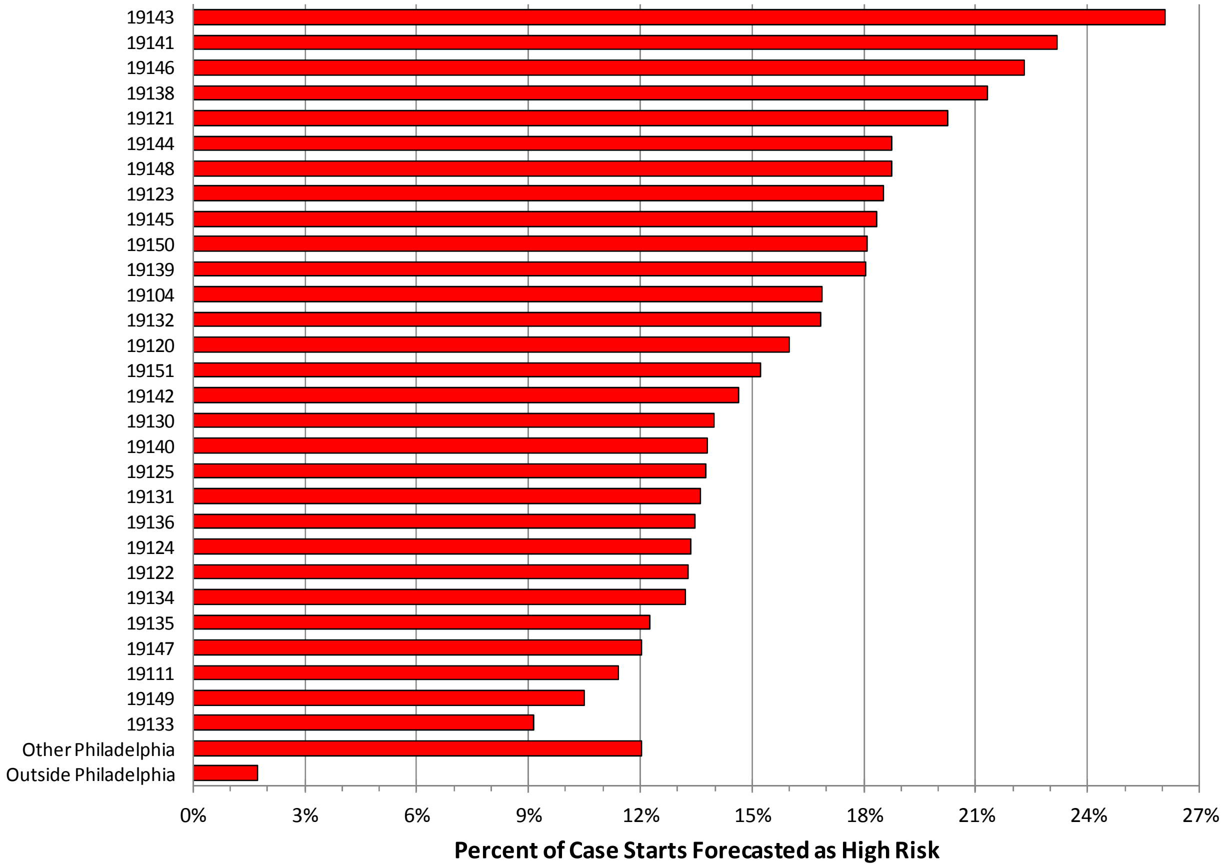


Figure 6: Probability of a High Risk forecast vs. residential zip code at the time of probation case start

Another important predictor for the High Risk forecasts was the onset age for adult offending. The relationship between this predictor and the prevalence of forecasted High Risk outcomes can be seen in Figure 5. Perhaps surprisingly, those who were charged with adults at the very youngest ages (12, 13, and 14) are considered less of a risk by the model[2] than those who onset somewhat later, but still prior to the usual age of adult responsibility (i.e., 15, 16, and 17). After this high point, however, the risk drops off quite rapidly. Those who manage to refrain from adult offending until their late 20s or later are comparatively unlikely to be forecasted as High Risk by this particular prediction model.

The offender's residential zip code is a very different sort of predictor than the others used in the model. While zip code is clearly an important predictor to both the model in general and to producing accurate High Risk forecasts, it is not a continuous variable and is treated by the model as 31 separate conditions within a single categorical variable. The differential effects within each of these zip codes on the likelihood of a forecasted High Risk outcome can be seen in Figure 6. More than 1 in 4 case starts with offenders living in 19143 (located in the southern

[2] To be fair, however, there are very few offenders in the sample who managed to be charged as an adult at such a young age. Charging offenders this young as adults normally takes place only when the offense is extraordinarily serious in nature, and it's therefore quite possible that this tiny group of offenders is simply different from those who are charged as adults in their later juvenile years.

part of West Philadelphia, quite close to the University of Pennsylvania campus) were forecasted as High Risk by the most recent model. Another common location for High Risk forecasts (19146) is situated just across the Schuylkill River in South Philadelphia. Two other areas that produced a relatively large number of High Risk outcomes (19141 and 19138) are also located next to one another, in the northern portion of the city.

The zip code that would appear to have been the least risky in Figure 6 (19133, located just north of Temple University's main campus) must be considered within the specific context of this particular predictor variable. The 29 explicit zip code values that were contained within this categorical variable were identified by finding the zip codes that had the most probationers living in them. The 19133 zip code produced nearly 28,000 APPD case starts between 2002 and 2007, ranking it sixth in terms of new case initiations. When this amount of probation activity is considered against the area's population, this one zip code produced 203 case starts for every 1,000 residents, making it by far the most probation-active residential zip code in the entire city. For comparison, the second-ranked zip code in this regard (19123) yielded just 159 case starts per 1,000 residents. So even though this zip code ranks below 28 others in the likelihood of a High Risk forecast, it also contains a very large number of offenders who are on APPD supervision.

Two final predictors that seemed especially important to the accuracy of the model's High Risk forecasts were the amount of time that had passed since the offender's most recent serious offense, and the number of times that the offender had been previously admitted into the county prison system. The relationship between these two predictors and the likelihood of a forecasted High Risk outcomes can be seen in Figures 7 and 8, and largely conform to what one might expect given current research on offending patterns. Those whose criminal histories reflect more recent involvement with serious offending are generally more likely to be categorized as High Risk by Philadelphia's newest forecasting model. For stays in prison, an increasing number of prior experiences with incarceration are associated with an increased likelihood of a High Risk forecast.

In both cases, however, these relationships appear to be non-linear in form, and the influence of both predictors seems to approach a natural limit as they increase in value. This diminishing influence of predictors as they increase to larger values also makes a certain amount of intuitive sense. The difference between an offender who has never been charged with a violent offense and one who has been charged with three such offenses might be quite important indeed. But after these values reach a certain level, the question of whether an offender has a substantial history of violent behavior has already been answered. Thus the difference between having 33 versus 36 prior charges for violent crimes may not be all that crucial for forecasting future serious offending.

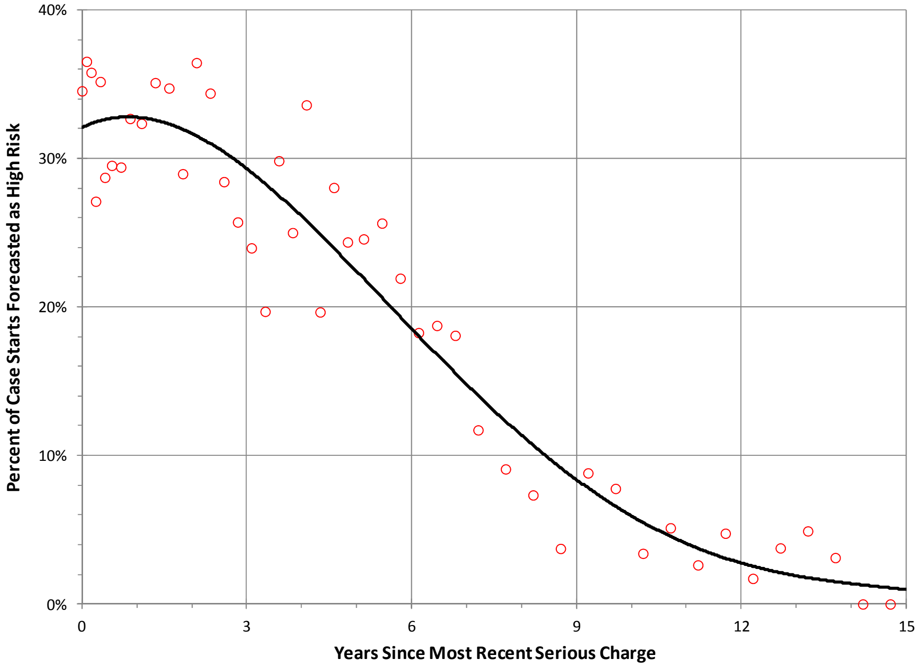


Figure 7: Probability of a High Risk forecast vs. the number of years since the offender's most recent serious offenses

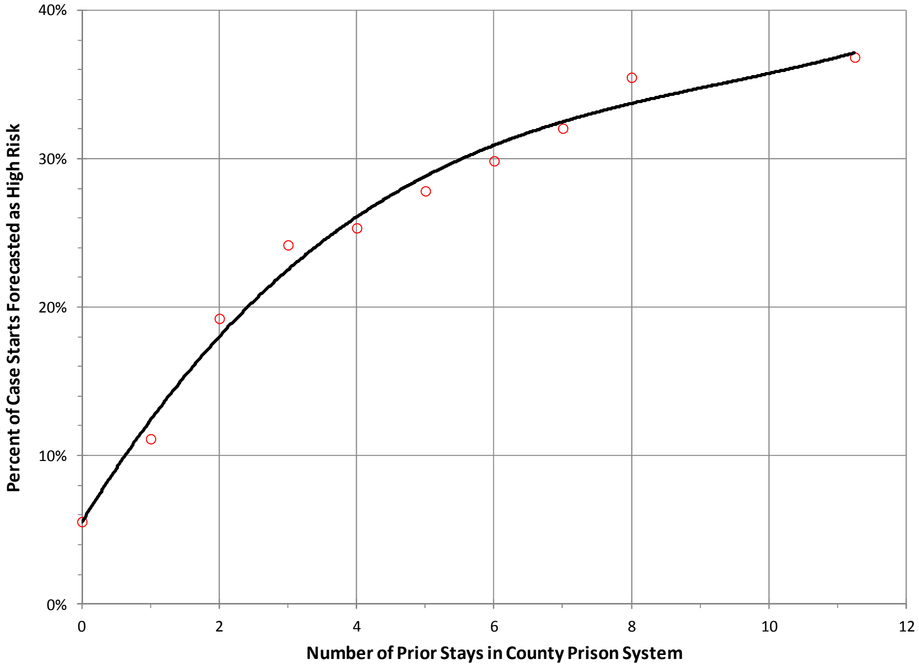


Figure 8: Probability of a High Risk forecast vs. the number of previous admittances to Philadelphia's county-level prison system

Random forest forecasting models allow a mix of predictors to operate in conjunction with one another, in ways that serve to form a very complex set of interrelated and curvilinear relationships. To the extent that these relationships mirror the natural reality of how prior experiences and present-day conditions blend together to influence future behavior, the accuracy of these forecasts can become quite strong. In the next section, we begin to examine just how accurate the three forecasting models used in Philadelphia have turned out to be.

Forecasting Accuracy

In order to fully understand the accuracy of a particular model (i.e., Model C), we must first examine the full confusion matrix. While a simplified version of this matrix, focusing only on High Risk outcomes, was shown earlier as Table 1, the complete version – which examines all three outcome categories – is shown in Table 3. This particular matrix is derived from the construction sample that was used to build Model C, and thus represents the forecasted results from a sub-sample (unique for each of the model's 500 trees) that was held in reserve as each individual tree was created. While values shown in the table aren't drawn from a fully-independent validation sample, they do represent the model's own best estimates of how well it could perform with such a sample.

As was the case earlier in Table 1, the accurate forecasts are shown in three green cells of the table (i.e., A, E, and I). In total, the most recent Philadelphia model produced an accurate forecast for 79,299 of the 119,935 probation case starts in the construction sample. Thus in overall terms, the model was correct 66.1% (nearly two-thirds) of the time.

While it is tempting to focus upon this one value to represent the model's overall accuracy, it is important to remember how the differential costs of various errors played a role in constructing this random forest model. A similar degree of overall accuracy, after all, could have been easily achieved by simply labeling every single probation case start in the entire sample as Low Risk. Doing so would have produced a model that was 59.4% accurate. All of the errors made by this simplistic forecast, however, would have been precisely the sort that Philadelphia's APPD most wanted to avoid; every time that the model was wrong, it would have under-estimated the offender's actual level of risk. Thus the real achievement of Model C is not that it is right two-thirds of the time, but that it produces this level of accuracy by balancing the relative costs of the different kinds of errors.

A more reasonable method of measuring the accuracy of the model's forecasts is to examine forecasted and actual outcomes separately, and then focus upon each of the three different outcome categories. To begin this process, one might first examine all of the case starts that were forecasted to result in a certain type of outcome, and then examine how many of them actually turned out in a manner that matched the predicted result. Of 18,812 new case starts that were forecasted to be High Risk, for example, just 7,112 (37.8%) of them included a probationer who went on to commit the predicted serious offense within two years of the case's start date.

Table 3: Confusion matrix and summary statistics for the most recent Philadelphia forecasting model (i.e., Model C), based on construction data

	Actual	Actual	Actual		
	High	Moderate	Low	Totals	Percent
Forecast High Risk	A 7,112	B 4,553	C 7,147	18,812	15.7%
Forecast Moderate Risk	D 2,248	E 23,000	F 14,867	40,115	33.4%
Forecast Low Risk	G 2,220	H 9,601	I 49,187	61,008	50.9%
Totals	11,580	37,154	71,201	119,935	
Percent	9.7%	31.0%	59.4%		

Total percent of forecasts that were accurate within 2 years:	66.1%
Of those forecast to be High Risk, percent which actually were:	37.8%
Of those forecast to be Moderate Risk, percent which actually were:	57.3%
Of those forecast to be Low Risk, percent which actually were:	80.6%
Of those actually High Risk, percent forecasted accurately:	61.4%
Of those actually Moderate Risk, percent forecasted accurately:	61.9%
Of those actually Low Risk, percent forecasted accurately:	69.1%
False Positive / False Negative ratio, High Risk:	2.62
False Positive / False Negative ratio, Moderate Risk:	1.21
False Positive / False Negative ratio, Low Risk:	0.54
Cautious error / Dangerous error ratio:	1.89
Very Cautious error / Very Dangerous error ratio:	3.22

While this percentage may seem low, it is important to note that actual serious offending in the APPD caseload is fairly rare, and occurs in less than 10% of all new case starts. Compared to this low base rate of serious offending, the 37.8% of forecasted High Risk offenders who go on to exhibit this behavior is, in reality, remarkably large. In fact, those who are forecasted by Model C as High Risk are more than 13 times more likely to commit a new serious offense within the two-year time horizon than those who are forecasted to as either Moderate or Low Risk.

Even more important than the low base rate of actual High Risk offending, however, are the cost ratios that were built into the model. As described earlier, the APPD leadership expressed a strong preference for erring on the side of caution when it came to making High Risk forecasts. The model is expressly designed to generate more High Risk false positives than false negatives, which in turn means that more people must be placed into the forecasted High Risk category than will actually engage in any amount of serious crime. Thus, the 37.8% rate seen in Table 3 is more a function of how the model was designed to function, and less a function of what level of accuracy it might be capable of producing.

With so many errors occurring (as desired) among those forecasted to be High Risk, it's hardly surprising to see substantially better performance within the case starts placed into the forecasted Moderate and Low Risk categories. Of the cases forecasted as Moderate Risk, 57.3% actually went on to result in one or more (non-serious) offenses within the next two years. Within the set of cases starts forecasted to be Low Risk, more than 80% reached the end of the two year period with no new offenses of any kind.

An alternate way to assess the accuracy of Model C's forecasting is to examine those case starts where the actual risk category is known, and measure how many were forecasted correctly. All three of these values are in excess of 60%. Of all the construction sample case starts which actually resulted in new serious offending, the model correctly identified 61.4% of them as High Risk. Those where the offender was actually Moderate Risk featured virtually identical accuracy, at 61.9%. New case starts where the offender was charged with no new offending at all during the next two years, meanwhile, were forecasted correctly nearly 70% of the time.

The various errors produced by the model are shown in both red (cells D, G, and H) and blue (cells B, C, and F). In a three-by-three confusion matrix such as this one, there are no single values that represent false positives and false negatives. Each of the three different outcome categories have their own sets of cells to represent these two types of errors, and every error cell plays different roles as both a false positive and a false negative, depending upon which outcome category is being discussed. For instance, the High Risk false positives (i.e., forecasted to be High Risk, but were not charged with a new serious offense) can be found by combining cells B and C, but cell B is also one of the false negative conditions for Moderate Risk, and cell C represent one of the false negatives for Low Risk.

As demonstrated earlier, in the discussion of Table 1, High Risk predictions used a cost ratio that produced 2.6 false positives for each false negative. The Moderate Risk cost ratio, on the other hand, is far more complex. A false positive, for example, could involve a forecasted Moderate Risk case who either committed a serious offense, or who engaged in no criminal behavior at all. With this mix of error types, the model's Moderate Risk cost ratio was fairly close to 1.0, meaning that it produced nearly as many false negatives as false positives. In the case of Low Risk outcomes, the cost ratio was established to favor false negatives. The APPD leadership desired a certain degree of confidence that those who were placed into the forecasted low risk group would largely turn out to be non-offenders when the two-year time horizon was reached. False positives for Low Risk – those who were forecasted to be non-offenders, but who actually ended up being charged with a new offense – were therefore assigned a larger cost than Low Risk false negatives.

Given that the meaning of "false positive" and "false negative" shifts depending upon which risk category is being discussed, it may instead be helpful to think of the errors in the confusion matrix as being either "cautious" or "dangerous" in nature. The cautious errors are those with a blue background (cells B, C, and F), and feature cases being forecasted to a higher level of risk than was necessary, based on the offenders' eventual behavior. The dangerous errors, meanwhile, are shown with a red background (cells D, G, and H). In these instances, the model made forecasts that were lower in risk level than the offenders' actual behavior required. Clearly, the APPD leadership preferred that the model make more cautious errors than dangerous ones. As can be seen in Table 3, the result was a model that was about 1.9 times more likely to err on the side of caution than it was to underestimate the actual risk posed by a given probation case. Moreover, in terms of the most extreme errors – forecasting High Risk for an actual Low Risk (cell C), or predicting no offending at all for a case that would actually result in serious crime (cell G) – the cost ratio becomes even more pronounced. For every one of the most dangerous errors (G), the model places 3.2 actual non-offenders into the forecasted High Risk category (C).

It seems clear that, based on the data it was constructed from, this most recent forecasting model (i.e., Model C) meets the requirements that governed its creation. It produces an impressive degree of accuracy, while ensuring overestimates of risk levels are much more common than underestimates. These results, however, are produced from the model's own construction data, and they don't allow us to compare Model C to the two earlier versions. In the next section, we will use an older set of 2001 probation case starts as a validation sample, allowing us to determine how the models have evolved over time.

Validation and Comparison Using 2001 Cohort

All three of the Philadelphia's live forecasting models have used construction samples which began on January 1, 2002. In order to validate the models using an independent data source (i.e., data that was never used to construct any of them), probation case starts that occurred between January 1, 2001 and December 31, 2001 are an obvious choice, albeit a somewhat problematic one. Since the cases in this validation sample are nearly a full decade old, they may not adequately represent what is happening with today's probationers. Ideally, more recent case starts could be used, but most of these were used in the construction of Model C. Of those that remain, the majority simply have not yet reached the two-year time horizon needed to measure the offenders actual risk level at the end of the forecasted period.

Despite these detriments, the 2001 validation sample also presents some unique advantages. Because so much time has passed since these probation cases began, we can measure actual offender behavior not only up to the end of the two-year time horizon, but well beyond it. It may be the case, for example, that the forecasting accuracy at the two-year point tells only part of the story, and that the forecasted risk groups become even more distinct from one another as time goes on. The offenses stemming from these 2001 probation cases have not only had more time in which to occur, but have also had ample time to be reported, investigated, and result in criminal prosecution. Perhaps most importantly, the cohort of 2001 case starts is

Table 4: Confusion matrix for the most recent Philadelphia forecasting model (i.e., Model C), based on 2001 validation sample

	Actual	Actual	Actual		
	High	Moderate	Low	Totals	Percent
Forecast High Risk	A 472	B 763	C 1,018	2,253	14.9%
Forecast Moderate Risk	D 435	E 1,627	F 1,899	3,961	26.3%
Forecast Low Risk	G 443	H 1,809	I 6,622	8,874	58.8%
Totals	1,350	4,199	9,539	15,088	
Percent	8.9%	27.8%	63.2%		

equally independent for all three of the live forecasting models, and therefore allows all of them to be compared to one another.

The degree of forecasting accuracy in this 2001 validation sample can be quite different from what was estimated using the construction sample. Table 4 presents yet another confusion matrix for Model C, drawn from the older validation sample. A number of things stand out from these values. First, the overall sample size is much smaller, since only 15,088 new probation cases began in 2001. Since that time, Philadelphia's APPD has become much more heavily utilized, and has seen an increase in new case starts per year. In 2007 (i.e., the last year of case starts that were used to construct Model C), the agency saw 25,335 new cases begin, an increase of 69% from 2001.

In addition to being somewhat smaller than cohorts from more recent years, the 2001 validation sample seems to have been slightly less criminally active. At the two-year point following the case start date, 8.9% of the 2001 cohort had committed a new serious offense (compared to 9.7% in the Model C construction sample), and 36.8% had committed a new offense of any kind (compared to 40.6% in the construction cohort). The result is that far more of the validation sample is categorized as actual Low Risk (63.2%) than was the case in the more recent construction sample (59.4%). The forecasts produced by Model C within this 2001 cohort generally reflect this lower amount of actual risk, with proportionally fewer High and Moderate Risk forecasts, and an increased amount of Low Risk forecasts.

A comparison of all three live forecasting models, using data from the 2001 validation cohort, is presented in Table 5. While Model B's overall accuracy in the validation sample was nearly equal to the value estimated from its construction sample, neither Model A nor Model C performed as well in validation as their construction estimates suggested. In general, all three models produced about the same degree of overall accuracy, forecasting correctly approximately 60% of the time. While the accuracy of Model C is slightly lower than its two predecessors, this difference seems to stem from a shift in cost ratios. Over time, the models have become increasingly cautious in nature, placing a higher cost on underestimating the offenders' actual risk levels when compared to overestimating them.

Table 5: Summary statistics for all three live forecasting models, based on 2001 validation sample

	Model A	Model B	Model C
Construction sample starting year (January 1):	2002	2002	2002
Construction sample ending year (December 31):	2005	2006	2007
Construction sample size:	50,000	50,000	119,935
Estimated overall accuracy from construction sample:	67.4%	59.7%	66.1%
<u>January 1 - December 31, 2001 Validation Sample</u>:			
Percent forecast as High Risk:	12.0%	13.5%	14.9%
Percent forecast as Moderate Risk:	20.8%	21.6%	26.3%
Percent forecast as Low Risk:	67.2%	64.9%	58.8%
Total percent of forecasts that were accurate within 2 years:	60.5%	60.0%	57.8%
Of those forecast to be High Risk, percent which actually were:	22.3%	21.5%	20.9%
Of those forecast to be Moderate Risk, percent which actually were:	43.4%	43.9%	41.1%
Of those forecast to be Low Risk, percent which actually were:	72.6%	73.3%	74.6%
Of those actually High Risk, percent forecasted accurately:	29.9%	32.4%	35.0%
Of those actually Moderate Risk, percent forecasted accurately:	32.4%	34.1%	38.7%
Of those actually Low Risk, percent forecasted accurately:	77.1%	75.3%	69.4%
False Positive / False Negative ratio, High Risk:	1.49	1.75	2.03
False Positive / False Negative ratio, Moderate Risk:	0.63	0.66	0.91
False Positive / False Negative ratio, Low Risk:	1.28	1.11	0.77
Cautious error / Dangerous error ratio:	0.90	1.02	1.37
Very Cautious error / Very Dangerous error ratio:	1.32	1.71	2.30

The ratio of cautious errors to dangerous ones steadily increased in all three iterations of the model, from 0.90 in Model A to 1.37 in today's Model C. This evolution in cost structure mirrors the desires of the APPD leadership. As the agency gained experience using these forecasting models, its management grew more comfortable with the size of its High Risk caseload, and was better able to understand how many offenders it could supervise under these conditions. When Model A was originally constructed, APPD had no historical reference for understanding how many individual offenders would fall into each of the three forecasted risk categories. The model's confusion matrix provided some guidance, but the forecasts were based on case starts, while caseloads were calculated based on individual offenders. In a live forecasting environment, some forecasts are effectively rendered moot by those that have come before them. For example, once an offender has been forecasted as High Risk, it no longer matters how many additional High Risk forecasts they may receive over the next two years, since the supervision requirements for this one offender can no longer increase.

Over time, the management of the APPD found that it could tolerate a larger percentage of case starts being forecasted as High Risk, because High Risk case forecasts did not always create new High Risk offenders. A sizable number of these case starts turned out to stem from offenders who were already being supervised in this manner. Moreover, slightly under half of the entire High Risk caseload tended to be incarcerated (either pre-trial or post-conviction) on any given day, which reduced the active workload of the High Risk probation officers even more. Thus the later models, in an overall sense, allowed for an increasing number of High Risk forecasts, which necessarily meant an increase in High Risk false positives and a proportional downward adjustment to the model's overall accuracy. The latest model, in other words, may produce slightly more errors than the earliest one, but this increase is largely intentional, and reflects APPD's growing level of comfort with both cautious errors and larger High Risk caseloads.

Long Term Offending Patterns in the 2001 Validation Cohort

Because the validation sample is constructed from cases that began during 2001, it is possible to examine the offenders' actual behavior over a much longer period of time than would be possible with a more recent cohort. Figure 9 presents a survival analysis of the forecasted High, Moderate, and Low Risk groups, and shows the "time to failure" until the offenders committed their first new serious crime after beginning their probation case. Since this graph focuses upon serious crime, it essentially represents the proportion of each forecasted risk group that became actually High Risk at any time between 0 and 8 years after the instant probation case began.

At end of the model's two-year time horizon, 21% of the forecasted High Risk group had fulfilled their prediction, and committed a new serious offense. This proportion can be compared to the 11% of forecasted Moderate Risk cases and 5% of forecasted Low Risk cases which defied the model's predictions and instead became actually High Risk. These same values can be easily calculated using the confusion matrix provided in Table 4.

After two years, we move beyond the span of time forecasted by the model, but the behavioral trends of the three forecasted risk group not only continue, but become more pronounced. Within five years of starting their new probation case, more than a third (36%) of the forecasted High Risk cases resulted in new serious offending, compared to just 20% of forecasted Moderate Risk and 10% of forecasted Low Risk case starts. After eight long years – when the majority of the instant probation cases have long since expired – the differences are even more noteworthy. At this stage, nearly 45% of all forecasted High Risk offenders have committed a new serious offense, while only 27% of forecasted Moderates and 14% of forecasted Lows have done so. To some extent, it would seem that many of the High Risk false positives errors were not so much incorrect about whether the offender would commit a serious crime in the future, but instead erred merely by forecasting this event too soon.

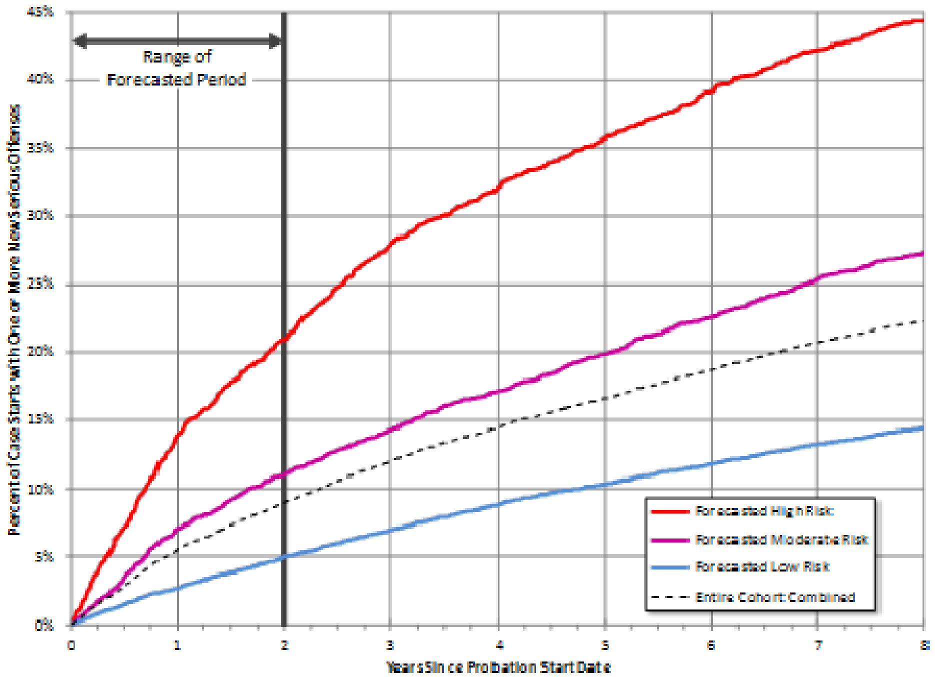


Figure 9: Survival function across all three forecasted risk groups; time until the commission of the first serious offense after the probation case start date

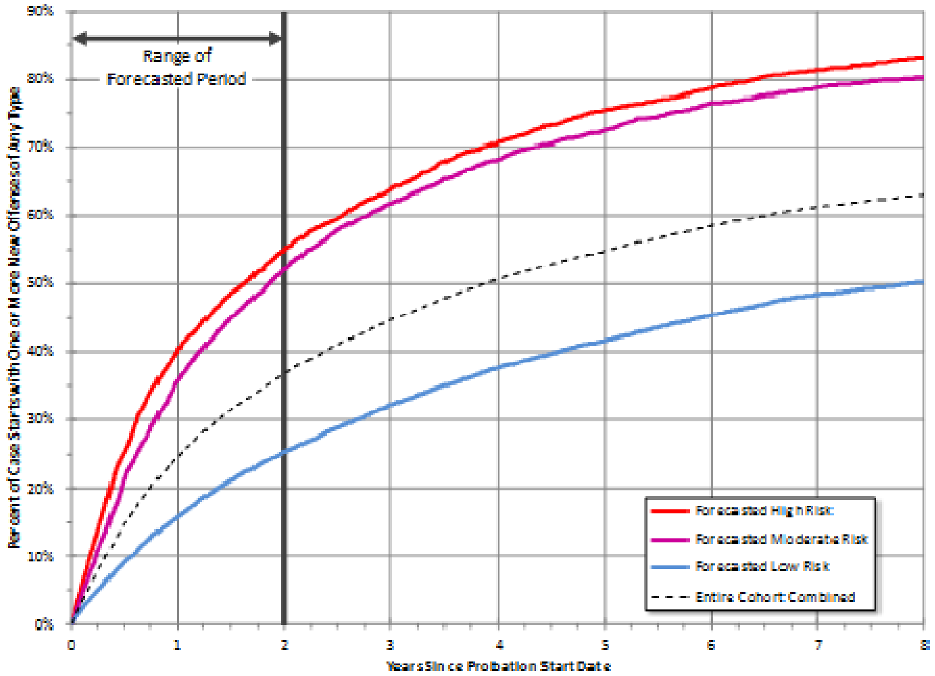


Figure 10: Survival function across all three forecasted risk groups; time until the commission of the first offense of any kind after the probation case start date

While the forecasts produced by Model C for the validation sample appear to have substantial long-term accuracy for predicting serious offending, their relationship with future offending generally (i.e., new offenses of any type) is more complicated. The survival curves shown in Figure 10 portray the time to failure until the first offense of any kind, broken down by the forecasted High, Moderate, and Low Risk outcome categories. While cases in the forecasted Low Risk group were much less likely to lead to new offending than those in the other two forecasted risk groups, offending was still quite common. More than half of the forecasted Low Risk probation cases experienced a new offense of some kind by the time that eight years had elapsed since their start date.

The forecasted Moderate and High Risk groups, meanwhile, were nearly tied with each other through the entire 8-year period shown in Figure 10. The members of both groups seemed equally likely to return to offending in some form, at any given time after their instant probation cases began. It is apparent that the model's forecasts produced quite different results for future serious offending (Figure 9) than they did for offending in general. What seems particularly noteworthy was the extraordinarily high rate of reoffending for those in the forecasted High and Moderate Risk conditions. In both groups, in fact, new offending appeared to be nearly certain to occur at some point. More than four in every five Moderate or High Risk forecasts resulted in a new offense by the time that eight years had elapsed since the probation start date.

In general, Philadelphia's probation population appears to consist of very active offenders, and a strong majority of the APPD's cases will result in at least one new offense at some point. Nearly 63% of all cases (High, Moderate, and Low Risk, combined) which started in 2001 exhibited some degree of new offending by the time that 8 years had passed. Nevertheless, the most recent forecasting model does an admirable job at differentiating the three risk groups, and those forecasted to be Low Risk, while still more likely than not to return to crime at some point, are also much less likely to do so than those in the other two forecasted risk categories.

While Figures 9 and 10 provide information on the prevalence of new offending, they do not indicate how many of these new offenses occur. While Figure 10 demonstrates, for example, that Moderate and High Risk cases were almost equally likely to result in at least one new criminal offense, it remains possible that the High Risk cases produced a larger number of new offenses than those that were forecasted to be Moderate Risk. The data displayed in Figure 11 address this concern, and show the mean number of criminal charges for new offenses (of any kind), in each of the three forecasted risk groups, throughout the full 8-year follow-up period available to the validation sample.

Despite the fact that forecasted High and Moderate Risk cases were almost equally likely to result in some degree of renewed criminal behavior, the total number of charges were noticeably higher in the High Risk group. More than 80% of the case starts in both groups exhibited a new offense within 8 years, but the total amount of offending, based on the number

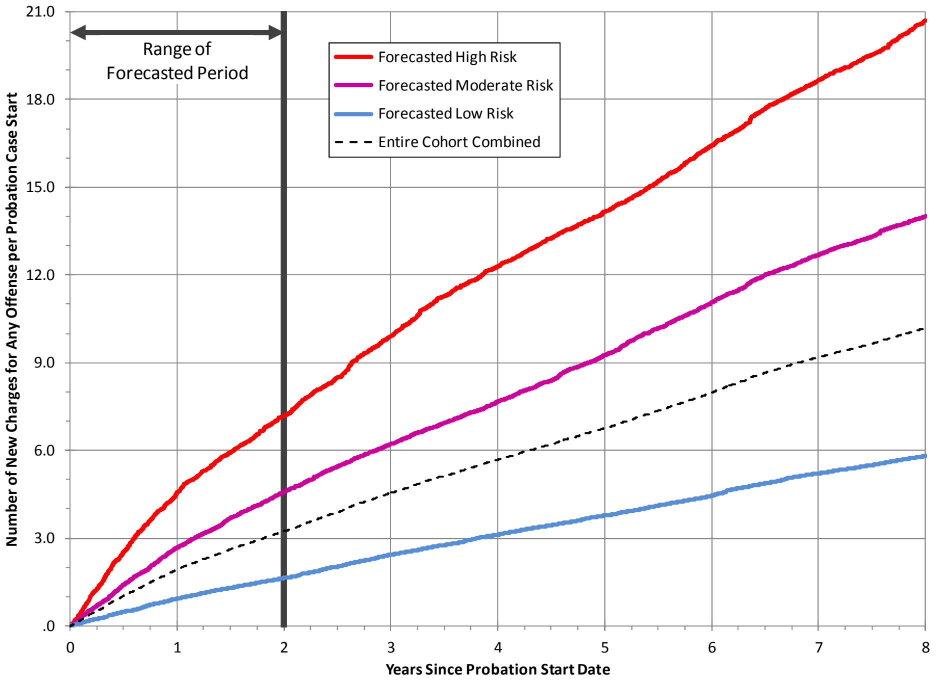


Figure 11: Number of criminal charges for new offenses of any type vs. time since probation case start date

of new charges filed per probation case start, was 55% higher for cases that were forecasted as High Risk when compared to those forecasted as Moderate Risk. Thus even though the model was constructed to forecast the likelihood of new offending – and was not designed to predict these differences in the total amount of criminal behavior – it still achieves some noteworthy success in this regard.

Steps Needed to Implement Forecasting

Should a jurisdiction wish to pursue a forecasting program similar to Philadelphia's, it would need to surmount many of the same challenges faced in this project. The following list summarizes the steps that will likely be necessary to produce a model and make it available for the live, on-demand forecasting of offenders.

1. *Obtain access to data*. Most agencies will already have a number of large and complex data systems available for use. In some cases, the necessary access may be limited to Information Technology employees, and their assistance will be needed. The key, however, is to ensure that all of the data sources that are used to create any construction data are also sources that are immediately accessible through the agency's data network.

Note that these data sources do not necessarily need to be up-to-the-minute accurate in order to be useful. For more than a year, for example, the initial juvenile predictor variables in Philadelphia were drawn from a legacy database which was updated only once a month. Since all of the clients under APPD's supervision were already (by definition) certified as adult offenders by the courts, there was no compelling reason for their juvenile records to change over time, and monthly updates were deemed acceptable.

In addition, it may become necessary to create new data sources where they are not already available. In Philadelphia, the jail admission and discharge data was sent out daily via a data stream, but was not available in database form within the APPD's network. To solve this problem, Penn researchers created a new database that was updated automatically from the daily feed sent out by the prison system.

2. *Define the unit of prediction and desired time horizon*. The moment when live predictions will be generated (i.e., once the system is fully operational) will be more obvious for some applications than for others. In Philadelphia, the logical point to produce these on-demand forecasts had to come after an offender was sentenced, but before they were assigned to a specific probation officer for supervision. These events could often be just an hour or two apart, which made it clear that the forecasts had to be generated when the offenders initially reported to the agency's intake unit. This requirement, in turn, immediately narrowed the unit of prediction down to the level of individual probation cases.

In other situations, immediate forecasting may not even be required. A parole agency, for example, may have a month or two to prepare before an inmate is released into their supervision. In this case, the unit of prediction will likely become each release of an offender onto parole, but the exact moment when the forecast will requested will be less clear.

Choosing the appropriate time horizon is likely to present a challenge during these early stages of the model construction process. Care must be taken to ensure that the horizon is set to a value that is long enough to be meaningful, but not so long as to require extremely out-of-date information in order to build the model. Also, once a time horizon is chosen, it should be viewed as a starting point rather than an inflexible requirement. Later analysis may show that the chosen horizon value is incongruent with the goals of the forecasting effort, and it may need to be either lengthened or shortened to produce the desired model.

3. *Define the outcome risk categories*. Although Philadelphia defined three different risk categories, other models could easily be limited to just two possible outcomes, or could be expanded to three or more risk categories. In either case, the categories must be mutually exclusive, and it must be impossible for a given case to fall into more than one of the outcome groups. In addition, the highest-risk category should be defined in such a way that it occurs in only a small minority of cases.

4. *Make preliminary plans for responding to forecasts*. In order to really finalize the unit of prediction, time horizon, and outcome risk categories, an agency must begin to think about how it will respond when the live forecasts become available. If the agency's plan is to focus an intense amount of resources on the highest risk category, for example, it will be necessary to define things so that this outcome occurs fairly rarely. Once the agency has estimated how many

offenders it can accommodate in each of its desired risk category, some preliminary data analysis may be needed. If the number of offenders who actually fall into the highest risk groups is larger than the agency can handle, it may be necessary to make adjustment to how these categories are defined. Ideally, the number of actual offenders in this category will end up being substantially smaller than the agency's desired caseload. This will allow room for false positives to be forecasted into this risk group, depending on the cost structure that the agency decides to use.

5. *Choose the predictors that will be used in the model*. While other forecasting methods place some strong limitations on how many predictors can be used, and how these predictors can relate to one another, random forest models are rather tolerant in this regard. At least for the initial versions of a model, a large number of predictors can be used, including predictors that are strongly correlated with one another. Those which prove to be relatively unimportant to forecasting the outcome, or whose predictive power can be better represented by other variables, will simply be used less often within the trees, and will become have less importance to the resultant forecasts.

In producing the very first forecasting model, the exact blend of predictors may be less important than where they come from. As discussed above, it is crucial to choose predictors that stem from instantly-accessible sources that will be available for making live forecasts later one.

6. *Build the construction data in a single data file*. While the need for a construction data set is clear, and the task seems relatively straightforward, this step may be one of the most complicated aspects of building a new forecasting model. The data required to construct this file likely reside in a number of different systems and different storage formats. Integrating them may require detailed knowledge of many different database management systems. In addition, the source data are very unlikely to be stored using the desired unit of prediction that was defined earlier. In Philadelphia, for example, a large relational database was the source of all criminal history data, and each row of each table represented a different unit within the database. None of these tables, however, were built around probation case starts, and all of the criminal history data therefore needed to be redefined and aggregated into the desired unit of prediction.

7. *Estimate the relative costs of false positives vs. false negatives*. These costs are ideally defined not by statisticians, but by the operational leadership of the agency that will rely upon the model. Precise accuracy is not required at this stage, and the cost estimates used here will almost certainly be redefined in later iterations of the forecasting model. Some quantitative estimate of these costs, however, must be provided in order to create an initial version of the model that will be useful enough to inform any necessary adjustments that come later.

8. *Build an initial model and evaluate the results*. The models used in Philadelphia were all constructed using a statistical package known as "R". R is an open source programming language, and the software needed to work with it can be downloaded at no cost (www.r-project.org). Within the R environment, additional packages can be downloaded (also for free)

to handle individual tasks. The “randomForest” package was used here (http://stat-www.berkeley.edu/users/breiman/RandomForests), and is one of several thousand different packages written for R.

The learning curve for using both R and the randomForest package can be enormously steep. The documentation is often unclear, and the completion of routine tasks may occasionally require additional a large amount of research and time. The easiest way to approach this step is to identify someone who already has experience working in R, and who understands its many challenges. In Philadelphia, the expertise of Professor Richard Berk at the University of Pennsylvania was essential in this regard.

The computing requirements of the “randomForest” package are also quite high, especially when a large number of cases or predictors are used in the construction sample. R will use an exceptional amount of memory to generate such a model, and a higher end desktop computer will be needed to do the work with any kind of speed. Depending on the method used to achieve live forecasting, this same computer may be used to produce on-demand forecasts later in the project.

9. *Make adjustments as needed and construct additional test models*. Once the first model is built, everyone involved with the forecasting project will need to become very comfortable with its resulting confusion matrix. The senior management at the participating agency must clearly understand how accurate the model is, both overall and for each of its outcome risk categories. The model will make errors, but it is the balance of the different kind of errors that must be fully understood and discussed. Balancing these different types of errors with the model’s overall accuracy rate is not a job for a statistician. The agency’s leadership will have to live with the consequences of any errors that occur once the forecasting effort goes live, and so they must be the ones to decide what level of accuracy they can live with, and the balance of errors that they prefer.

Once the results of the initial model are fully understood, it is almost certain that further adjustments will be needed. At this stage, nearly every prior decision can be revisited. If needed, the unit of prediction, time horizon, outcome categories, predictors, and cost ratios can be adjusted as needed. In addition, any weaker predictors from the initial model may be dropped entirely, since their presence increases the model’s memory requirements, causes unnecessary bloating in the size of the final model, and will decrease the speed of producing the on-demand forecasts.

In all likelihood, several iterations of the modeling process will be needed before a final model will be identified. In Philadelphia, we quickly lost count of how many different models were produced over lifespan of the project. By conservative estimates, it seems safe to say that at least 30 different random forests were built, examined, and discarded in the production of the three live forecasting models.

10. *Produce forecasts for the standing caseload*. The first step in using the new forecasting model is likely to focus on those who are already in the agency's existing caseload. Producing these forecasts may require, ironically enough, violating some of the very rules that were used to construct the model in the first place. For example, the Philadelphia model was predicated on obtaining forecasts when each new probation case began. Offenders in the standing caseload, however, often had several different cases active at a time, and many of these cases began years earlier. It made little sense, however, to use the model to make a forecast of what risk level an offender would have posed long ago, when their present cases began. Instead, the forecasts for the standing caseload were all produced as if the offenders were starting a brand new probation case at the present time.

Once these forecasts become available, the effort required to sort the different offenders into their risk-appropriate supervision levels may be quite substantial. Philadelphia chose to create four different divisions. Each of the three forecasted risk levels were given a separate division, while the fourth contained a number of specialized supervision programs that offenders could be judicially mandated to receive, independent of their forecasted risk level. In order to staff the three risk-stratified divisions, however, the officers who were being assigned to them had to first be stripped of their existing caseloads before being given a new set of offenders from the appropriate risk category. This process involved a delicate balancing of caseloads (some of which were only temporary in nature) to keep the agency's workload on an even keel while nearly every offender was moved to a new supervising officer. The entire process ultimately took four or five full months to complete.

11. *Create the user interface and back-end software needed to produce live forecasts*. Once the final model has been agreed to, there remains a fair bit of work in making it accessible to the everyday users who will generate the live forecasts. Much of the effort in this area will depend upon who these users are, and how often such forecasts are needed. If the forecasts will be needed on individual offenders who arrive with little or no notice, then a custom user interface will likely be required. If, on the other hand, the forecasts can be produced for larger batches of offenders on weekly or monthly basis, then perhaps R can be used directly by one or two well-trained employees.

In Philadelphia, new cases arrived in a trickle throughout the workday, and a custom Risk Forecasting Tool was required. The system employed a Windows-based front end application that an intake worker could use to communicate with a series of database servers that provided the necessary predictor values. These values were then passed to a separate database server that held all of the millions of different nodes that made up the trees in the forest. Next, the forecasting server used these predictor values to "drop down" each case through the unique set of paths that these values defined, and counted the "votes" that resulted from the hundreds of resultant terminal nodes. Based on these votes, a final overall forecasted result was then communicated back to the originating user and displayed on their screen. In addition, the software that was written for APPD recorded the forecasted risk result in the agency's own case management system, and identified the appropriate unit and probation officer to handle the offender's case, base on the forecasted risk level.

In technical terms, the Philadelphia approach to designing its live risk forecasting tool is far from the only option available. The state of Maryland, for example, has taken an entirely different approach, installing R directly on a web server and using an applet to communicate with R and initiate new forecasts. In the end, the type of infrastructure used to produce live forecasts will depend heavily on the set of skills available for use, and the types of employees who will be tasked with producing the forecasts.

12. *Monitor the results of the live forecasts*. Once the model is installed and live forecasting begins, line employees will almost immediately begin to disagree with the model's results. It may often be the case that external data – information that was not available for immediate use, and hence was not included in the model's predictors – will suggest that an offender presents a higher level of risk than the model has forecasted. The question, "How can this guy be low risk?" will be asked quite frequently, often followed by, "He was convicted of [serious offense] in another county 10 years ago!" Indeed, it was precisely these sorts of complaints that led Philadelphia to seek live access to juvenile offending data (which was not used in Model A), and to request a set of Pennsylvania statewide offending data for testing purposes.

These kinds of reactions are likely to continue for some time after live forecasting begins, but in Philadelphia they eventually became fairly rare events. The agency's leadership set the tone that the model's recommendations were to be adhered to, even in the face of the most vehement disagreements, and eventually the APPD staff learned to, if not trust, then to at least to accept the model's forecasts as the primary determinant of an offender's supervision level.

Throughout this entire time, however, the agency's senior management staff were constantly examining the forecasts produced by the model, the caseloads sized in the three risk-stratified divisions, and whether the procedures used by these divisions was achieving the desired impact on offender behavior. These efforts included a series of randomized controlled trials (RCTs), which randomly assigned groups of offenders, all of whom had the same forecasted risk levels, into different supervision conditions. While some of these offenders were assigned to be supervised in a manner that matched their forecasted risk levels, the rest remained in conditions that mirrored the agency's traditional case management techniques. Of the RCTs which have concluded, the results to date confirm that the new supervision plans did not increase the amount of offending among Low Risk offenders (Barnes, et al. 2010). In addition, a High Risk RCT is currently underway to determine the best method for managing offenders who are forecasted to commit serious crime.

After monitoring the results of the forecasting model for a number of months, an agency may find it necessary (or at least desirable) to alter the manner in which they are used. These changes could involve a series of rules that determine when (and if) the model's forecasts should be overridden. In Philadelphia, for example, when a new offender is assigned to Low Risk supervision, their officer performs a full multi-state criminal records search. If the offender has any history of sexual offending, they are transferred into Moderate Risk supervision instead of remaining in Low Risk. Such a plan may not work in every case, and each agency must make its own decisions about overriding their model's recommendations.

Conclusion

As technology progresses, forecasting models such as the one in Philadelphia will become increasingly more sophisticated, accurate, and easy to produce. There is no question that such models – based upon random forests or some other statistical techniques – will play an important role in the future of America's criminal justice system. The use of these forecasts to make real-world decisions will, just as assuredly, invoke an understandable and justifiable degree of discomfort. Our legal traditions allow us to treat offenders differently based on their past conduct, but have never before been confronted with the availability of accurate forecasts about what these same offenders are predicted to do in the future.

This is not to say that traditional sentencing and supervision guidelines have ignored the possibility future offending. One of the reasons that we sentence some offenders to longer prison terms is to prevent the crimes that they would otherwise commit in the future if they were not incarcerated (Piquero and Blumsteim 2007). Up until this point, however, these predictions of anticipated offending have often been very simplistic, based on only a coarsely-grained examination of the offender's previous criminality. Because they were based on such limited information, it seems likely that many, if not most of these predictions were deeply flawed and inaccurate, but they at least stemmed from predictors that were deemed acceptable for use in this regard.

In the end, it may be the choice of predictors that determines the acceptability of these new forecasting methods. Would it ever be permissible, for example, to include an offender's racial background as a predictor variable in one of these models? If not, what about the use other predictors, such as residential location or familial circumstances, which could indirectly communicate the offender's racial identity into the forecasting model? Could it be permissible to use these more controversial predictors in lower-stakes forecasting models, such as those used to control admission into a treatment program or govern supervision decisions, while prohibiting their use for higher-stakes outcomes such as sentencing? There are no easy answers to these questions, but they will almost certainly need to be addressed as these forecasting techniques become more and more integrated within criminal justice decision-making.

The power of these forecasting methods is clear. Their use in Philadelphia has allowed the city's adult probation agency to stratify offenders by the risk they pose, and adjust the amount of supervision delivered accordingly. While the system was never designed to help control costs, the fact that it arrived just as the economy went into a deep recession has proven to be enormously fortuitous. Rather than expending resources on offenders who were unlikely to reoffend, regardless of how they were supervised, APPD has been able to focus its limited resources on the much smaller number of offenders who require more active supervision. Despite a hiring freeze and the natural attrition of case-carrying officers, the agency has been able to handle a 28% increase in its overall caseload with a staff that is 15% smaller than it was before the introduction of forecasting. According to APPD's Chief Probation and Parole Officer, this feat simply would not have been possible without the use of risk forecasting (Elliott-Engel 2011).

If for no other reason than to help control costs, the pressure to use sophisticated forecasting techniques is sure to expand across the entire criminal justice system. The limits on how these models can be constructed and used are, as yet, undefined. The exact nature of any future forecasting models is therefore somewhat unclear; their existence, however, is not. These kinds of models will become more widely used in the future, and their forecasts will never be error-free. How we choose our predictors, balance these errors, and control the use of these predictions is likely to determine how accepted these models become.

Works Cited

Andrews, D. A., James Bonta, and Steve Wormoth. "The Recent Past and Near Future of Risk and/or Need Assessment." *Crime and Delinquency*, 2006.

Austin, James. “Reducing America's Correctional Populations: A Strategic Plan.” *Justice Research and Policy* 12, no. 1 (2010): 32-35.

Barnes, Geoffrey C., Lindsay Ahlman, Charlotte Gill, Lawrence W. Sherman, Ellen Kurtz, and Robert Malvestuto. "Low-Intensity Community Supervision for Low-Risk Offenders: A Randomized, Controlled Trial." *Journal of Experimental Criminology*, 2010: 159-189.

Berk, Richard A. "Forecasting methods in Crime and Justice." *Annual Review of Law and Social Science* (Annual Reviews Press) 4 (2008a): 219–238, 236.

—. *Statistical Learning from a Regression Perspective.* New York: Springer, 2008b.

Berk, Richard A., and J. de Leeuw. "An Evaluation of California's Inmate Classification System Using a Generalized Regression Discontinutity Design." *Journal of the American Statistical Association* 94 (1999): 1045-1052.

Berk, Richard A., Lawrence W. Sherman, Geoffrey C. Barnes, Ellen Kurtz, and Lindsay Ahlman. "Forecasting Murder Within a Population of Probationers and Parolees: A High Stakes Application of Statistical Learning." *Journal of the Royal Statistical Society, Seriese A: Statistics in Society* 172, no. 1 (2009): 191-211.

Berk, Richard A., Li Azusa, and Laura J. Hickman. "Statistical Difficulties in Determining the Role of Race in Capital Cases: A Re-analysis of Data from the State of Maryland." *Journal of Quantitative Criminology* 21, no. 4 (2005).

Dodge, K. A., T. J. Dishion, and J. E. Lansford. *eviant peer influences in programs for youth.* New York, NY: Guilford Press, 2006.

Elliott-Engel, Amaris. "Shortage of Probation Officers May Imperil Reforms, FJD Leaders Say." *The Legal Intelligencer*, April 15, 2011.

Emery, Mark, Bonnie Gasswint, Michael Hartman, and Dan Lategan. *Annual Statistical Report, Pennsylvania Department of Corrections, 2007.* Harrisburg, Pennsylvania: Pennsylvania Department of Corrections, 2008.

Gibbs, J. B. *Crime, punishment and deterrence.* New York, NY: Elsevier, 1975.

Gill, Charlotte E. *The effects of sanction intensity on criminal conduct: A randomized low--intensity probation experiment.* Doctoral Dissertation, Philadelphia, PA: University of Pennsylvania, 2010.

Glaze, Lauren E. *Correctional Populations in the United States, 2009.* Washington, DC: Bureau of Justice Statistics, 2010.

Gottfredson, Stephen, and Laura Moriarty. "Statistical Risk Assessment: Old Problems and New Applications." *Crime & Delinquency* 52 (2006): 180.

Lowenkamp, C., E. Latessa, and A. Holsinger. "The Risk Principal in Action: What Have We Learned." *Crime and Delinquency* 52, no. 1 (2006): 77–93.

National Research Council, Committee on Community Supervision and Desistance from Crime, Committee on Law and Justice, Division of Behavioral and Social Sciences and Edcuation. *Parole, Desistance from Crime, and Community Integration.* Washington, D.C.: The National Academy Press, 2007.

Petersillia, Joan. "Probation in the United States." *Crime and Justice* 22 (1997): 149-200.

Pew Center on the States. "One in One Hundred: Behind Bars in America." 2008.

Pew Center on the States. *One in Thirty One: The Long Reach of American Corrections.* Pew Center on the States, Washington, DC: Pew Center on the States, 2009.

Piquero, Alex, and Alfred Blumsteim. "Does Incapacatation Reduce Crime?" *Journal of Quantative Criminology*, no. 23 (2007): 267-285, 267.

Rhodes, William. "Predicting criminal recidivism: A research note." *Journal of Experimental Criminology*, 2001: 57-71.

Sherman, L. W. "The Power Few Hypothesis: Experimental Criminology and the Reduction of Harm." *Journal of Experimental Criminology* 3, no. 4 (2007).