Risk Assessment: An Art or Science?

Predicting Recidivism at the Time of Sentencing

Miguel Camacho-Horvitz, Mathematical and Computational Sciences
Jeremy Kim, Electrical Engineering
Stanford University

***Abstract*—Motivated by recent legislation that bases sentencing of criminals on their likelihood to recommit a crime, we developed models that would make a prediction about whether or not individual criminals would commit another crime post-release. Using various supervised learning techniques, we looked at features of relevant criminal, locational, and demographic information to make predictions based on the known outcomes of, in some ways, similar individuals. We also created a "balanced set" with equal numbers of positive and negative examples and trained and tested on that set. From there, we moved on to a hybrid unsupervised/supervised model. In this approach, we classified the data we had using k-Means Clustering before training our supervised learning algorithms on each data-cluster and looking at the average predictive error between clusters. Finally, we calculated Mutual Information statistics to infer which features were most informative of recidivism in an effort to decompose our hypothesized biases. In the end, we found that in the full data set the supervised learning models on their own could barely perform better than a null-hypothesis that no-one recommits, but when paired with the k-Means clustering, we saw slight but significant improvement in the prediction accuracy.**

I. Introduction

As a society, we seek to reduce crime. One important facet of this complex goal is the minimization of prisoner recidivism, or a relapse in criminal behavior following release. To this end there have traditionally been several strategies including rehabilitation services and prisoner vocational training. Recently, however, quite a different strategy has arisen in the national conversation. In 2010, Pennsylvania adopted into its state legislature the idea of using a "risk assessment" metric as a factor in determining the most appropriate incarceration sentences. By Title 42, Section 2154.7:

(a) General rule.–The commission shall adopt a sentence risk assessment instrument for the sentencing court to use to help determine the appropriate sentence within the limits established by law for defendants who plead guilty or nolo contendere to or who were found guilty of felonies and misdemeanors. The risk assessment instrument may be used as an aide in evaluating the relative risk that an offender will reoffend and be a threat to public safety. [1]

What if at the time of sentencing we could predict whether an offender in question, still awaiting his/her sentence, would commit another crime following his/her ultimate release from our penal system? Given such a metric, many have advocated that we could reduce the burden on our prison system by providing "low-risk" offenders early parole as well as enhancing public safety by keeping tabs on "high-risk" offenders. However, there has been significant push-back, with critics raising both moral and practical questions surrounding the usage of risk-assessment instruments. Indeed, former Attorney General Eric Holder commented that "by basing sentencing decisions on static factors and immutable characteristics – like the defendant's education level, socioeconomic background, or neighborhood – they may exacerbate unwarranted and unjust disparities that are already far too common in our criminal justice system and in our society."

Interesting as it may be, in this study we avoid any moral or ethical debating and focus on numbers. In particular, we hope to address perhaps the most obvious and practical question: *given the apparent consequential nature of a risk-assessment instrument, how well could such a model truly predict prisoner recidivism?*

II. Related Work

We originally came across this topic through a collaboration between well-known statistician Nate Silver's *FiveThirtyEight* and *The Marshall Project*, a non-profit organization studying criminal justice. Just months ago, in August of 2015, they co-published an in-depth piece titled "Should Prison Sentences Be Based On Crimes That Haven't Been Committed Yet?" [2] in which they brought up the idea of risk-assessment as a tool for criminal sentencing. Though acknowledging the controversies, they pointed to several counties across the US in which initial implementations of risk-assessment tools used in criminal sentencing has led to less severe sentences for "low-risk" offenders and a reduction or flat-line in recidivism.

Intrigued by this idea, we looked into how such counties, states, or outside organizations have constructed or determined the best risk-assessment instruments. We found that multiple other studies ([3], [4], [5], [6]) published statistics about rates of recidivism. In particular, we found that previous studies focused their analyses on identifying significant features. This is one reasonable application of such analysis on prisoner recidivism as there are so many human biases at play that there is a natural goal of demystifying which features are in fact most predictive as well as identifying the correlations between features. Unfortunately, these studies do *not* publish their predictive models nor discuss their methodologies.

Thus, the main difference between our work and these previous studies is that we will go beyond feature inference and resulting predictive error percentages and report specifically which models do best as well as interpreting our results.

III. Data

A. Data source

To perform any meaningful analysis on recidivism prediction, it was critical that we find a longitudinal study on prisoners who were tracked following their release from the prison system. While there were interesting questions to ask in the unsupervised setting, we had our minds set on a classification model that would make use of the binary outcome vector that stated, *did he/she reoffend?* feature. The United States Department of Justice's Bureau of Justice Statistics (BJS) was the only source that had large-scale studies of the type we were searching for. We could gain access to a study conducted from 1986-1989 through the National Archive for Criminal Justice Data. [7]

This data set, *RECIDIVISM OF FELONS ON PROBATION* is composed of 12,369 "cases" of felons released on probation in the year 1986. Each case has 149 associated features consisting of information ascertained from sentencing records, probation files, and criminal history files. Notably, the study includes whether the felon re-entered the prison system during the four years this study was conducted. It is worth noting that all of our analysis will be in terms of recidivism in this four-year period only.

B. Processing

Let us now define a "risk assessment instrument", as the Pennsylvania legislature did, to mean simply an empirically-based model which uses known information about an individual that are relevant in predicting recidivism to do just that. In order for us to construct such a model, we first had to process our data such that our features were discrete and ordered (as it did not make sense to run regression analysis on categorical data). Secondly, in order to draw conclusions as to predicting recidivism at the time of sentencing, we needed to ensure that all of the features we included in our models were known *at the time of sentencing* and not afterward.

To this end, we identified the number of previous convictions, type of crime committed, whether there was a pre-sentence investigation, number of address changes, drug abuse history, education, and a host of demographic information as suitable features. One of the challenges we faced early on was that many features we would have liked to include, such as employment and income, were pre-processed, assigned weights, or coded in a way that made them unavailable to us. Additionally, several features had to be discretized and the categorical features needed to be transformed into binary indicator variables; type of crime committed, for example, was divided into nineteen new binary features that corresponded to indicator functions of whether or not the crime was a certain crime.

Perhaps the greatest challenge we had with our data set was in dealing with missing data. When the outcome, if the subject reoffended or not, was not known we threw those examples away. Unfortunately, though not unexpectedly, many of the input data points were listed simply as 'unknown'. Moreover, the distribution of unknowns was largely uniform so it was unfeasible to throw away certain features or examples. In the end, we decided it best to deal with this using inter-feature correlations.

Given that a feature vector X_j had an example $x_j^{(i)}$ that was missing, we determined the feature vector X_k most correlated with X_j. We then found the value of $x_k^{(i)}$ for that training example. Then, for every example $x_k^{(n)}$ in X_k where $x_k^{(n)} = x_k^{(i)}$, we found the corresponding value $x_j^{(n)}$ in X_j. We then found the average (mean for ordered data, mode for categorical data) of all the associated values $x_j^{(n)}$ to assign to $x_j^{(i)}$.

Finally, this left us with an 11,712 example by 38 feature data set that was ready for analysis.

IV. Models

A. Supervised Learning

For this study, we looked at 4 supervised learning algorithms: Naive Bayes, Logistic Regression, Support Vector Machines, and Random Forest Classifiers.

Naive Bayes - The Naive Bayes algorithm looks at the conditional probabilities of an input given a certain output as well as the marginal probabilities of both that input and output. It then calculates the conditional probability of an output given that input with:

$$p(y|x) = \frac{\prod_i p(x^i|y)p(y)}{\prod_i p(x^i)}$$

and predicts the output with the highest condition probability.

Logistic Regression - The logistic regression algorithm assumes that conditional probabilities of outputs given input features follow a logistic function as follows:

$$p(y = 1|x) = \frac{1}{1 + e^{-(\theta_0 + \theta_1 x)}}$$

. By looking at a joint likelihood function:

$$L(\theta) = \prod_i p(y^i|x^i; \theta)$$

and setting the partial derivative with respect to θ equal to 0, the algorithm solves for the maximal parameters θ. For each new test example, the features x are put into the logistic function, and if the value is greater than 0.5, output 1 is predicted.

Support Vector Machine - A Support Vector Machine (SVM) uses the Lagrangian dual form of the optimal margin classifier to construct a separating hyperplane between positive and negative examples. The general form of the dual optimization problem for a linear SVM is:

$$max_\alpha W(\alpha) = \sum_i \alpha^i - \frac{1}{2} \sum_{i,j} y^i y^j \alpha_i \alpha_j < x^i, x^j >$$

such that:

$$\alpha^i \geq 0, \forall i$$

$$\sum_i \alpha_i y^i = 0$$

For nonlinear classification, we use Kernels to define feature mappings to represent non-linear data as feature vectors in terms of only inner products and create a hyperplane in the

transformed space. We tried 3 different Kernel's in our tests: polynomial, sigmoid, and Gaussian, defined (respectively) as:

$$K(x,z) = (x^T z + c)^d$$
$$K(x,z) = tanh(ax^T z + r)$$
$$K(x,z) = -\frac{||x-z||^2}{2\sigma^2}$$

Random Forest Classifier - A random forest classifier is an ensemble learning algorithm that captures the average or the mode of many regressions. The data is split into B different branches of n samples each, called X_b, y_b. Each branch then uses logistic regression to get a training rule f_b and calculates the mean of these rules as:

$$\hat{f}_b = \frac{1}{B}\sum_{b=1}^{B} f_b(x)$$

The idea behind this is that the mode of our regressions would capture a stronger prediction than would any single regression; moreover, as regressions may overfit to the training data, this ensemble algorithm corrects for such overfitting.

B. Unsupervised Learning

k-Means Classifier - The k-Means algorithm classifies each training example into one of a few clusters based off of minimum distance to the cluster center

$$C^{(i)} = argmin_j ||x^i - \mu_j||^2$$

.

The cluster centers, μ_j, start as random but are then updated according to the samples placed in them with:

$$\mu_j = \frac{\sum_i 1(c^i = j)x^i}{\sum i 1(c^i = j)}$$

C. Feature Inference

Mutual Information - In order to look at the similarity between features and the outcomes, we used the mutual information statistic. In particular, we looked at the similarities between the values our features x_i took on and the positive (reoffending) outcome $y = 1$, as follows:

$$MI(x_i, y=1) = \sum_{x_i \in \chi}\sum_{y=1} p(x_i, y=1) log\left(\frac{p(x_i, y=1)}{p(x_i)p(y=1)}\right)$$

V. Results and Discussion

We first began by establishing naive assumptions as predictors and figuring out baseline test errors associated with using these predictors. The five we looked at were to assume that everyone recommits (87.69% error), assume no-one recommits (12.31% error), assume that all people under the age of 30 will recommit (56.49% error), assume all violent criminals recommit (30.10% error), and assume that all with previous records recommit (24.67% error). Our subsequent results would be compared to these baselines. However, it is immediately apparent that there is quite a low error associated with predicting no one recommits given how imbalanced our training set is. As a result, our initial estimated test errors could barely exceed the predictive power of the "null hypothesis" that no one recommits a crime. Given this, we split our subsequent analyses into two main sections: one in which we trained on the full, imbalanced data set, and another in which we randomly sampled both positive and negative training examples in order to create a set with even numbers of the two.

Full Dataset	% Error
Baseline	12.32 (Assume No-One Recommits)
All Features	12.21 (Polynomial SVM)
5 Features	13.14 (Logistic Regression)
After Clustering	8.59 (8 clusters, Polynomial SVM)

TABLE I
FULL DATASET RESULTS: THE BEST TEST ERRORS WHEN TRAINING AND TESTING ON THE FULL DATASET. CLEARLY, THE TEST ERROR REMAINED MORE OR LESS CONSTANT UNTIL THE HYBRID MODEL OF K-MEANS CLUSTERING AND SUPERVISED LEARNING WAS EMPLOYED.

A. Full Data set Analysis

Our first step in deriving a model was to train the four supervised learning algorithms. From our set of 11,712 examples, we performed k-fold cross-validation with 10 folds and calculated the average test error. Using all of the features in our set, we found that the errors we could get barely exceeded the predictive power of the "null-hypothesis" that no one recommits a crime. There were slight variations for the various models, but the best that each could do was 12.28% error for logistic regression, 12.21% for the SVM's (best with polynomial Kernel), and 12.28% for the random forest classifier.

While this showed some improvement over many of our baseline tests, it was clear that the prediction was very skewed towards giving a negative result and in regards to that baseline, we saw no improvement.

B. Balanced Subset Analysis

Seeing how skewed our model was towards the negative prediction and realizing that close to 90% of the data used for both training and testing were negative examples, we attempted to create a more even subset of the data. We took a random sampling of negative examples until the number of negative examples matched the number of positive examples in our data. We then concatenated all of these examples into a matrix that we will refer to from now on as our "balanced set".

We found baselines and employed learning algorithms to this balanced set in the exact same way that we had in the full, unbalanced set. For this set, if you assume that everyone recommits or no-one recommits the test error is 50%, if you assume that all criminals under 30 will recommit it is 44.80%, if you assume all violent criminals recommit it is 61.86% and if you assume all with a previous record recommit it is 50.72%.

Our results for the test error of the four unsupervised algorithms in the balanced set was significantly improved

from these baselines. When using all of the features in our data, logistic regression achieved 25.12% error, the SVM with Guassian Kernel gave 25.19% error, and the random forest classifier resulted in 21.16% error.

Balanced Dataset	% Error
Baseline	44.80 (Assume Only Young Criminals Recommit)
All Features	21.16 (Random Forest Classifier)
3 Features	21.12 (Random Forest Classifier)
After Clustering	33.45 (4 clusters, Polynomial SVM)

TABLE II

FULL DATASET RESULTS: THE BEST TEST ERRORS WHEN TRAINING AND TESTING ON THE BALANCED DATASET. WHILE THE OVERALL ACCURACY OF THESE MODELS ARE NOT AS GOOD AS WITH THE FULL TRAINING SET, WE HAVE IMPROVED PERFORMANCE OVER THE BASELINES. THE 3 FEATURE MODEL HAS AS MUCH (AND EVEN SLIGHTLY MORE) ACCURACY AS THE MODEL USING ALL OF OUR FEATURES.

C. Feature Selection

Next, we found that our training error was significantly lower than the test error (as low as 1% in some cases), our models were most likely suffering from high variance. Since we had a limited number of examples, we could not really increase the size of our training data. We looked at using more folds for cross-validation, but had no significant improvement. We then tried to use fewer features.

We performed both forwards and backwards search to find the features with that showed the largest change in test error when added or removed. For both models, backwards search yielded few results, as the model was already so skewed in one direction and removing one feature at a time did not give a significant change in the error.

Using forward search on the full data set, the algorithm effectively identified the lowest occurring features. When averaging across LR, NB, RF, and SVM, the lowest expected test error (using CV) resulted from a 5-feature model which included the four lowest occurring features as well as the crime committed being a weapons offense. Naturally, the features that were most predictive were simply the lowest occurring features, as they were likely to result in the model predicting a negative result, and this procedure clearly revealed nothing surprising. More importantly, however, at no point were we able to ascertain a group of features that gave us a test error below the 12.2% error that we had seen before that seems to predict that almost no one would recommit.

For the balanced set, on the other hand, we could identify 3 features that could get the training error to within a statistical margin of the results when using all of our features (See Figure 1). These features - whether or not the person was Hispanic, whether or not there was an investigation prior to the trial, and whether or not the person was a female - were not necessarily the ones that we expected. When training and testing (using cross validation) on this balanced set using only these three features, the test error converged to near the error when considering all features, and in some cases was better. The best test error came from the Random Forest Classifier with an error of 21.12%, slightly lower than the best with the full set and again significantly improved from our baselines. We could not, however, drastically outperform the model that used all of the features.

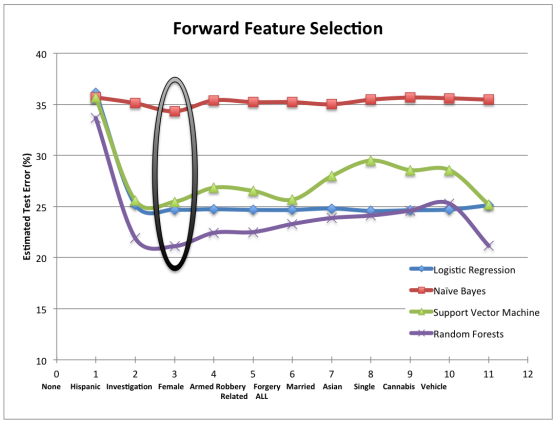

Fig. 1. Forward Feature Selection: The training error from training and testing (with CV) on the balanced data set. The test error for every algorithm (except Naive Bayes) comes down significantly and converges after the first three features added.

Finally, we looked at the Mutual Information statistic to see which features shared the most information with the results vector in hopes that this could shed light on which features to use. The results produced a few dichotomies that follow stereotypes that one may predict: men are positively correlated with recidivism, women are negatively; single people positively, married people negatively; black people positively, white people negatively. Perhaps surprisingly, the other feature that shared the most information with the results was the binary value that indicated a crime was burglary (See Figure 2).

While these dichotomies give some interesting insight, the features are almost definitely correlated with many other factors, some of which we did not have access to. For example, being black is very highly correlated in our data with having a prior criminal record, being single, and being employed less of the time before the crime was committed. Therefore, it is important to acknowledge the inter-dependence of so many features (especially demographic features) rather than jumping to conclusions. Indeed in the *FiveThirtyEight* and *Marshall Project* collaboration, they found that even when they tried to explicitly eliminate certain biases they couldn't. In one analysis, they found that they could remove features on race and income without losing much information as other features, such as being single and unemployed or being male and without a high-school diploma, simply served as proxies for your removed features.

We tested models based on just these features with strong mutual information, but again our models failed to significantly outperform the null-hypothesis for the full set and did not perform better than our model from feature selection for the balanced set.

D. Hybrid Model

It was not until we combined the unsupervised learning with the supervised learning tool of k-means clustering that we saw improvement in the test error. We ran the k-means algorithm

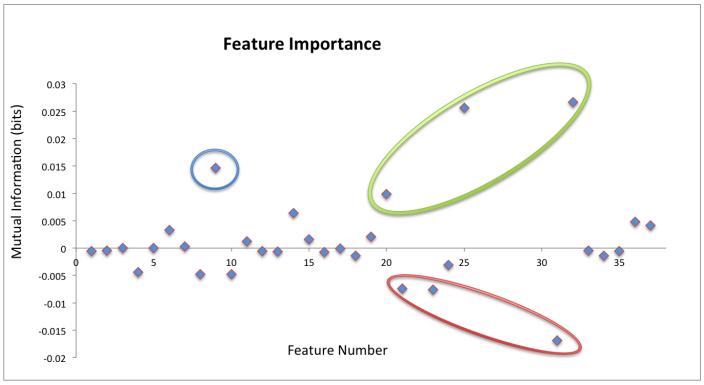

Fig. 2. Graph of Mutual Information: Mutual Information between each feature and the outcome vector. The features with the strongest MI are circled. The blue circle represents the crime burglary, the green circle (from left to right) is being a man, being single, and identifying as black, and the red circle (from L to R) has the counterparts, being a woman, being married, and identifying as white.

with random initial cluster centers and then once the data was classified, trained and tested using 10-fold cross-validation on the samples in each individual cluster. This was repeated 10 times to get an average error. The best test error that we found in all of our tests was 8.59% when the data was sorted into 8 clusters and we classified the data in each cluster using an SVM with a polynomial Kernel.

We ran a similar protocol for the balanced set, but found actually slightly worse results after clustering. There are a number of possible explanations for this, but perhaps this clustering can not do as good a job of separating the data into very distinct groups due to the even distribution of positive and negative examples that you start with.

It is important to note that every time that the tests were performed, the training and testing was done on within the same cluster. Therefore, it is important that the cluster centers that are chosen for training are held fixed when clustering a new example. Only then can this clustering improve your accuracy of prediction.

VI. Conclusion

The questions and controversies surrounding risk-assessment as a tool in criminal sentencing cannot be overstated. Statistics, after all, allow us to infer generalizations about groups of people. To use the actions of previous collections of what we deemed ”similar” people in deciding the fate of an individual carries implicit moral/philosophical assumptions and value-judgements which we do not necessarily agree with. Still, independent of this, the concept of predicting recidivism poses an interesting academic challenge.

Given the data available to us and the time-scale of this project, we could not develop a model that predicted recidivism with accuracy significantly different than making a naive prediction that no one recommits. Based off of this, and combined with the gravity of a false positive result (someone serving longer jail sentences), our model shows no added value of using statistics to look for individuals who will recommit crimes. When we had a contrived set with even numbers of positive and negative examples, we could outperform this naive prediction by almost two fold. However, attaining this even distribution may never be feasible and a 25% error is still far too high in our minds to base sentencing based off of this model.

In the future, if a similar predictive model is to be built and used, a far more in depth and larger data set is needed. Not only would we want to look at more examples, but we see merit in both tracking criminals over a longer period of time and having more complete information about someone’s background in order to control for correlated features. We believe that recidivism prediction is possible, but given the resources we currently have, we are not able to make an adequately performing model.

VII. Acknowledgments

Many thanks to Professor Andrew Ng and the entire CS229 teaching staff for all of their support on this project.

References

[1] *http://www.legis.state.pa.us/.*
[2] The Marshall Project. *The New Science of Sentencing*. FiveThirtyEight, Oct. 2014.
[3] Patrick A. Langan, PhD, David J. Levin, PhD. *Recidivism of Prisoners Released in 1994.* Federal Sentencing Reporter, Vol. 15, No. 1, Recent State Reforms II: The Impact of New Fiscal and Political Realities (October 2002), pp. 58-65.
[4] Alexia D. Cooper, Ph.D., Matthew R. Durose, Howard N. Snyder, Ph.D. *Multistate Criminal History Patterns Of Prisoners Released In 30 States.* September 24, 2015 NCJ 248942.
[5] Alexia D. Cooper, Ph.D., Matthew R. Durose, Howard N. Snyder, Ph.D. *Recidivism Of Prisoners Released In 30 States In 2005: Patterns From 2005 To 2010.* April 22, 2014 NCJ 244205.
[6] Allen J. Beck, Ph.D., Bernard Shipley. *Recidivism Of Prisoners Released In 1983.* April 1, 1989 NCJ 116261.
[7] *Recidivism of Felons on Probation, 1986-1989*. United States Department of Justice. Office of Justice Programs. Bureau of Justice Statistics.
[8] *Risk Assessment*. Pennsylvania Commission of Sentencing. PCS, Pennsylvania State University.