MODEL FOR RECIDIVISM PREDICTION

Julia Andre, Thomas Trinelle, and Luis Ceferino

Stanford University

CS229 - MACHINE LEARNING PROJECT - STANFORD UNIVERSITY

Motivation

Today, the **United States have one of the highest recidivism rate in the world**: with 2.3 billion people in jail, almost 70% of the prisoner will be re-arrested after their release. This poses a serious problem of safety, and proves that we don't make the decision that really make us safer. Judges, even though they have good intentions make decision subjectively. Studies show that high-risk individuals are being released 50% of the time while low risk individual are being released less often than then should be. With machine learning tools we can produce accurate predictive models based on various factors as age, gender, ethnicity, employment. Detecting patterns in recidivism would **provide supportive arguments** for judges to determine the appropriate sentence, which will decrease safety risks while trying to avoid over-punishment.

Goal

The purpose of this project to use data and analytics to transform the way we do criminal justice. Using **supervised learning** we can design a **predictive model for recidivism**. This decision making tool will help the judges determine wether an individual is dangerous or not. The common concerns addressed by this tool are precedence, biases, access to information, and jurisdictional errors.

Methodology

Our analysis consisted of the following steps :

- Data acquisition.
- Data pre-processing.
- Feature extraction.
- Algorithm implementation using CV Technics
- Feature Engineering
- Analysis and comparison of the different results.

Prediction problem

Input variable*:* We used a total of 16 input variable. Most of the variable are coded as a binary rule of the form *x E (0,1)*. Some other feature (like the age, years in school or time served) consists of an integer.

Outcome variable*:* binary variable if the person went back to jail or not.

To develop the most accurate predictive model, we tested eight different machine learning algorithms and analysed their learning curve in order to mitigate high biases or variance. Algorithm tested: Logistic regression, Gradient Boosting, Random Forest, Perceptron, Linear and Kernelized SVM. To assess the accuracy of each algorithm we used cross-validation method (simple and k-fold) with random selection.

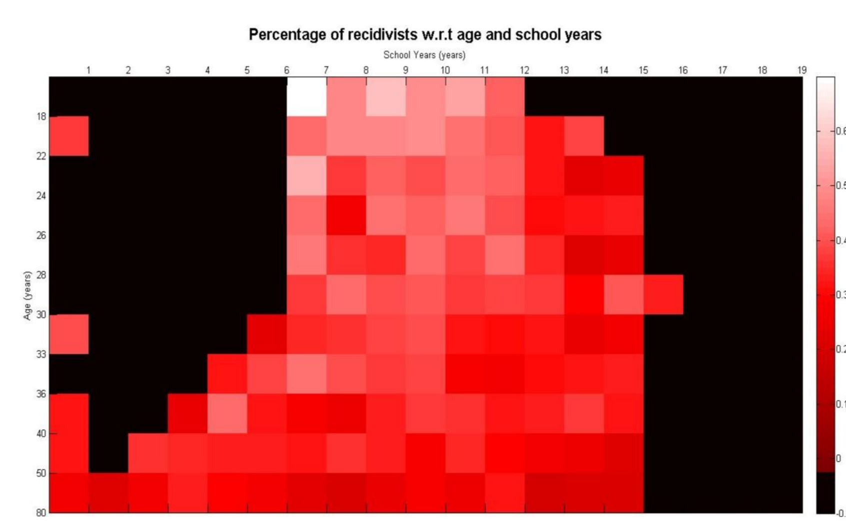
The data

Feature	Description
Ethnicity	Origin of the individual
Age	Age of the individual
Male	Gender of the individual
Alchoolic	Is the individual an alchoolic?
Junki	Is the individual a drug addict?
Married	Is the individual marriedt?
Supervision	Did he receive supervision in Prison
Felony	Was he arrested for felony ?
Work Rel	Was he part of the work release program in jail?
Property	Was he arrested for a crime against property ?
Person	Was he arrested for a crime against a person ?
Time served	How long did he stay in prison?
Follow	How long ago was the release?
Priors	How many time was he arrested before?
School	How many years of education?
Rule	How many rules did he break while in prison?

Data were retrieved from a database in the Inter-Univ. Consortium for Political and Social Research (ICPSR) and collected by Smith and Witte (1984). It has information on two cohorts of inmates that were released in 1978 and 1980 from the prison of North Carolina. This data set contains around 18000 data points.

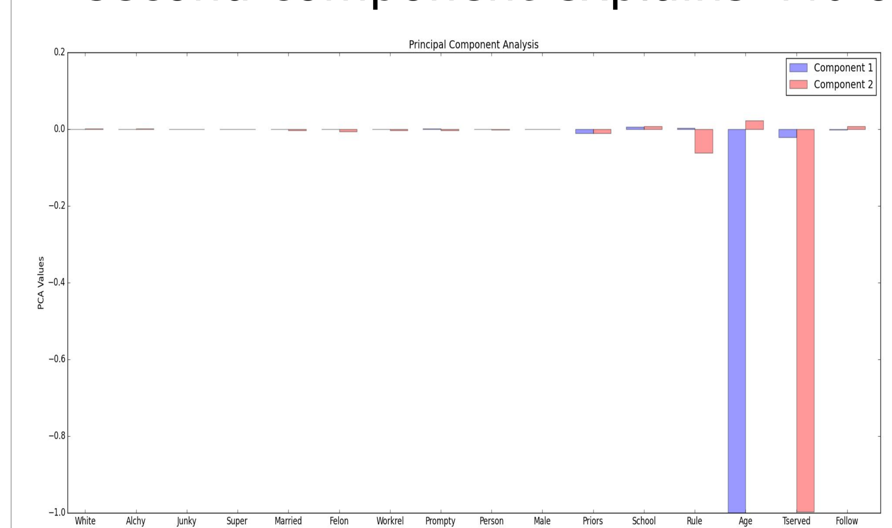
Feature Engineering

Statistical Exploration



PCA

We ran the PCA to determine which feature is most important to reduce the dimensionality of our data. Results show that the first component explains nearly 95% of the total variance, whereas the second component explains 4% of the total variance.

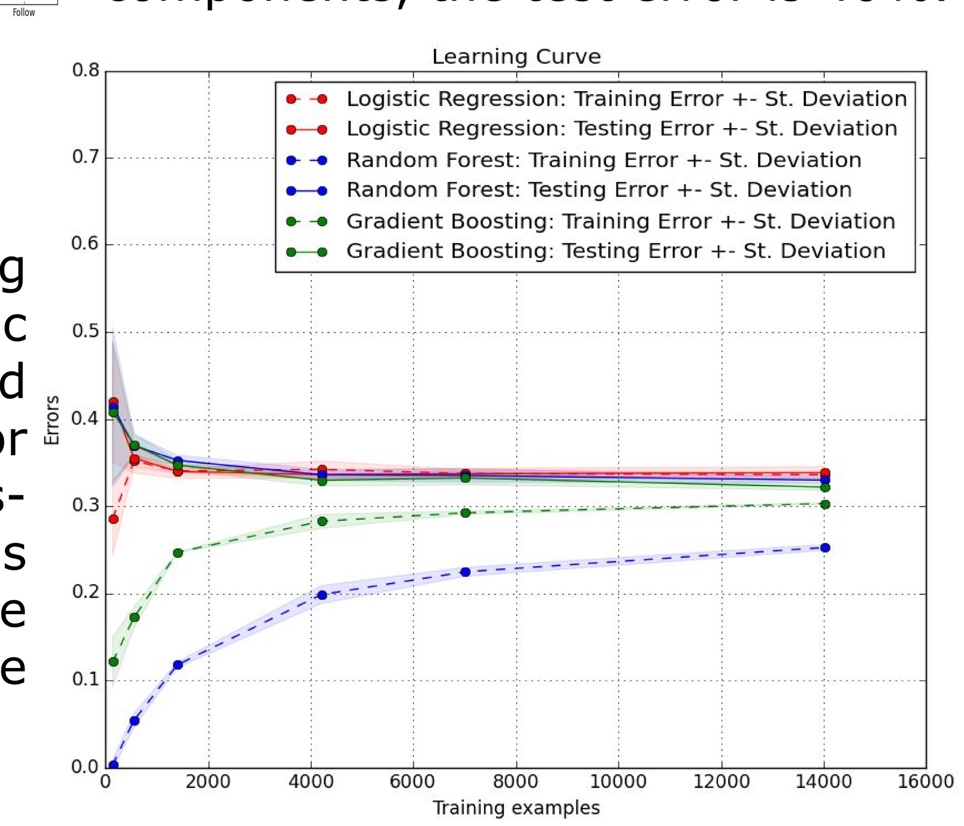


We performed a linear transformation of the feature space into the first component subspace. The test error found is 37%. Using the first two components, the test error found is 40%, and for the first three components, the test error is 40%.

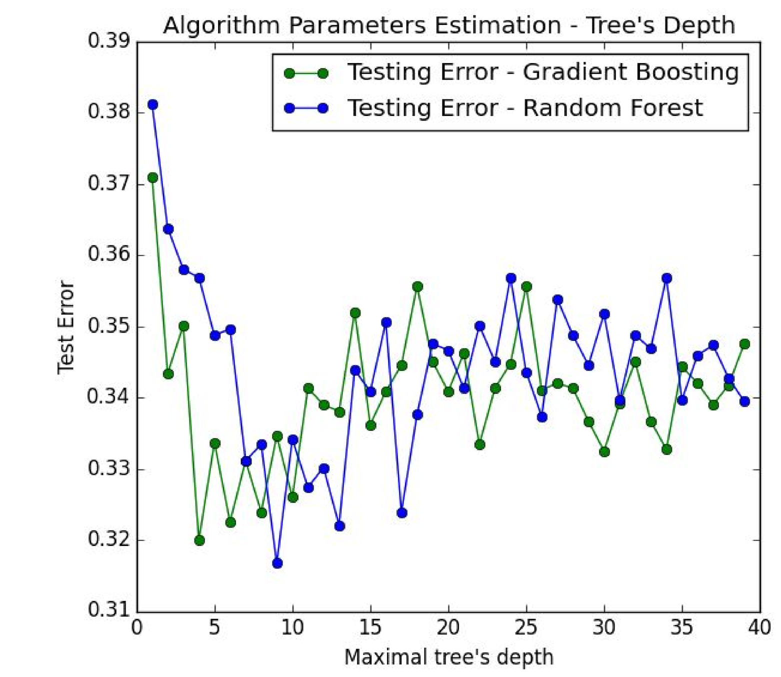
Results

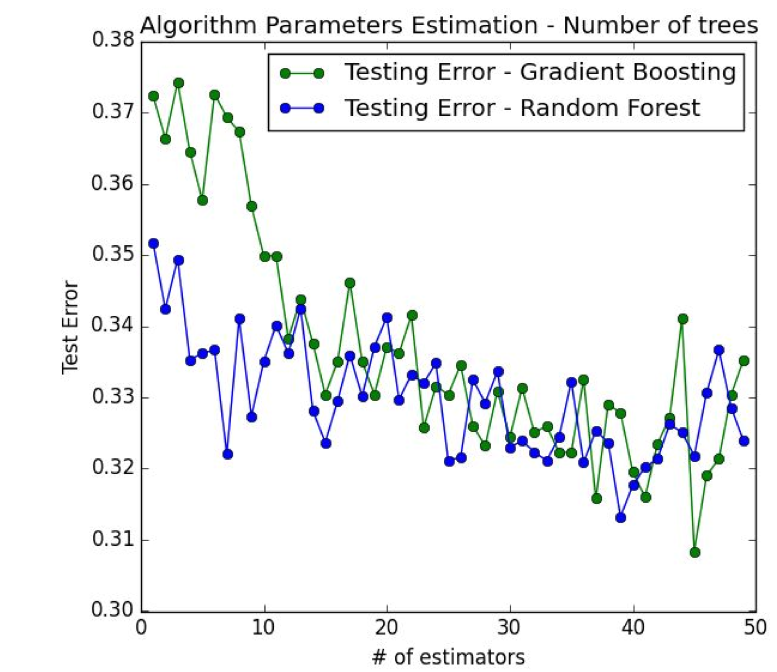
Learning curve

Our team assessed the learning curves using 3 methods: Logistic Regression, Random Forest and Gradient Boosting. The test error was found using 5-fold cross-validation. The objective of this analysis was to diagnose bias/variance issues, and the convergence of the methods.

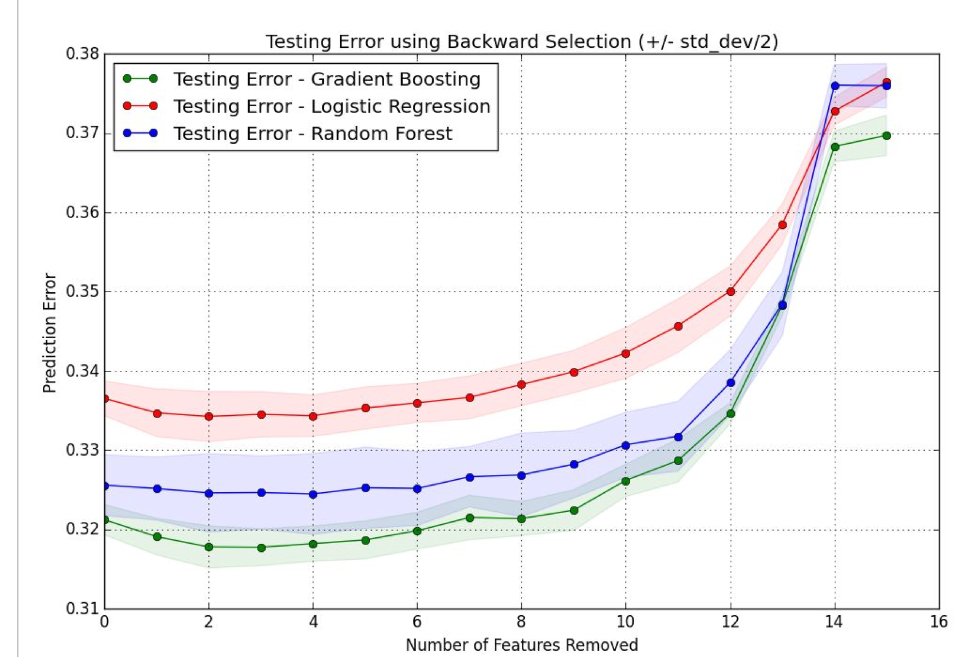


Parameter Estimation





Feature Selection



We have used both **forward & backward** feature selection to measure the importance of the features w.r.t to one another. Without doubts the most crucial features are (in order):

(1) Ethnicity
(12) Time Served
(14) School years
(15) Rule violations

Part of the difficulty was understanding the **which features were the most indicative** of individuals likely to recidive. 'Manually' engineering features (linear & logic combinations) as been explored un-fruitfully. We believe this approach should be pursuid in feature work.

Conclusion

Our work is an attempt to recidivism modelling. We use feature that are easily accessible by the judge and have a significant impact on the probability of recidivism. It was determined that the best predictive model is the gradient boosting algorithm using 13 features (follow, felony & property) with an error of 31.8%.

The error remains pretty high which can be accounted for by 2 major reasons (1) the incompleteness of the data and (2) the possibility of wrong training points. Huge data set with millions of points have already been collected in the US. Yet, we could not access it for this project since an IRB protocol is required for sensitive data on human subjects.

This exploration is not be confused with a willingness to substitute judge by machines. On the contrary, it helps them make better decision to improve the american criminal justice system, to make it more just, objective and fair.

References

[1] Milgram, 2014, "Why smart statistics are the key to fighting crime", Ted Talk.
[2]J. Zheng, B. Ustun, C. Rudin, 2015, Interpretable Classification Models for Recidivism Prediction