The author(s) shown below used Federal funding provided by the U.S. Department of Justice to prepare the following resource:

Document Title:	NIJ Recidivism Challenge Report: Team Smith
Author(s):	Andy J. Smith
Document Number:	305051
Date Received:	July 2022
Award Number:	NIJ Recidivism Forecasting Challenge Winning Paper

NIJ Recidivism Challenge Report

Team Smith

Andy J. Smith – sgt.a.smith@gmail..com

1. Introduction

During this challenge, contestants were provided with a training data set and three test sets with the objective of using the training data to develop ad train a machine learning (ML) model that can predict the recidivism of the individuals in the test data set. The training data provided was intended to represent the data a parole officer would have at the time the individual was released on parole. The data provided covered a wide range of inputs from the education and mental health to previous arrest and conviction information.

Using the training data, I developed and tested a variety of traditional ML models to predict the recidivism of each person. To support this I brought in a variety of geographic data to inform on the environment each person was returning to; though they provided little significance to the final model. The final model selected was an ensemble method of four traditional ML models.

2. Variables

In addition to the data provided by the NIJ for this challenge, I added additional data taken from the PUMA of the individual as part of the normalization process. I had initially assumed that bringing in additional data about the PUMA zone would inform on the individual's recidivism. The data that was brought in included the average income, average lot size, etc. but none of these variables were statistically significant. It is believed that they were not statistically significant because of the similarity between the PUMAs included in this challenge and the ways the PUMAs were grouped together. The statistically insignificant variables were included in the model as removing them did not change performance; however, in hindsight retesting the removed models on the DNN may have increased its performance but this was not tried. The variables added were:

- **Education / Dependents** – Each category was normalized as a percent of the population in its PUMA zone providing a value 0-1
    - Toddlers
    - K-3
    - G4-6
    - G7-11
    - 12th grade - no diploma
    - Regular High School Diploma
    - GED or alternative
    - Some college, no degree
    - Associate's degree
    - Bachelor's Degree
    - Grad Degree

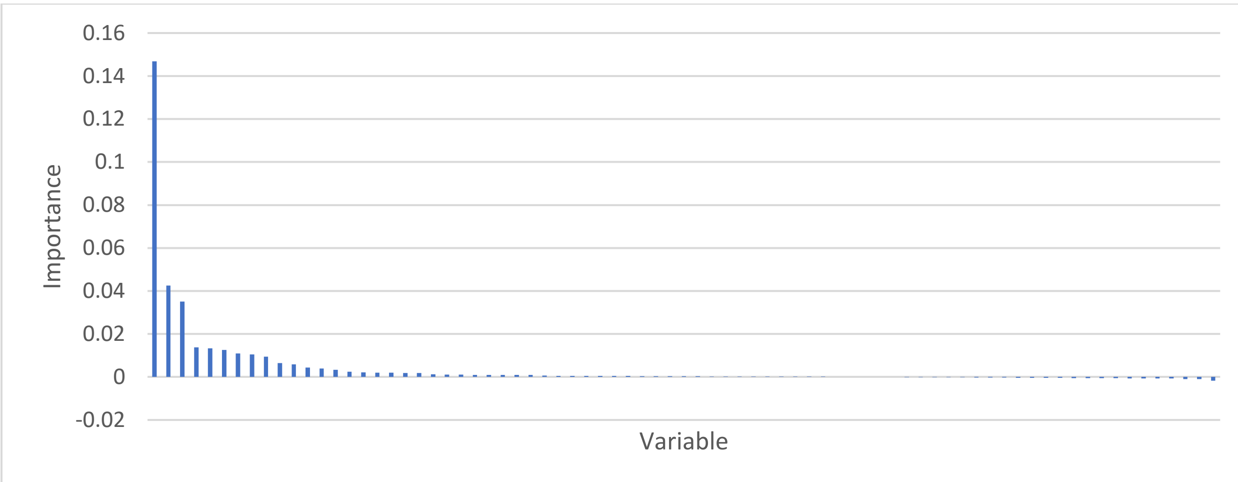
- **Family** – Each category was normalized as a percent of the population in its PUMA zone providing a value 0-1
    - N/A (GQ/vacant/not a family/same-sex married-couple families)
    - Married-couple family: Husband and wife in LF
    - Married-couple family: Husband in labor force, wife not in LF

- Married-couple family: Husband not in LF, wife in LF
- Married-couple family: Neither husband nor wife in LF
- Other family: Male householder, no wife present, in LF
- Other family: Male householder, no wife present, not in LF
- Other family: Female householder, no husband present, in LF
- Other family: Female householder, no husband present, not in LF

- **Wealth** – Each category was normalized across all PUMAs so that the PUMA with the highest value provides an input of 1 and the PUMA with lowest value has an input of 0.
  - Family income (past 12 months, use ADJINC to adjust FINCP to constant dollars)
  - Wages or salary income past 12 months (use ADJINC to adjust WAGP to constant dollars)
  - Property value
  - Income-to-poverty ratio recode

- **Geography** – Each category was normalized across all PUMAs so that the PUMA with the highest value provides an input of 1 and the PUMA with lowest value has an input of 0.
  - Total Lot Size (ACR)
  - N/A (GQ/not a one-family house or mobile home)
  - House on less than one acre
  - House on one to less than ten acres
  - House on ten or more acres

During the variable evaluation it became clear that a majority of the variables used were not statistically significant, instead, there were a few variables that heavily contributed to the predicted recidivism of each individual as shown below.



Variable	Importance
Percent_Days_Employed	0.327906
Jobs_Per_Year	0.146774
Delinquency_Reports	0.042524
Residence_Changes	0.035046
Prior_Arrest_Episodes_Felony	0.013777
Age_at_Release	0.013256
Avg_Days_per_DrugTest	0.012549
Prior_Arrest_Episodes_PPViolationCharges	0.010899
Prior_Arrest_Episodes_Misd	0.010444

Program_Attendances	0.009509
Gang_Affiliated	0.006419
Gender	0.005934
Prior_Arrest_Episodes_Property	0.004345
Prior_Revocations_Parole	0.003896
Supervision_Level_First	0.00332
Supervision_Risk_Score	0.002506
Prison_Years	0.002125
Condition_MH_SA	0.001994
Violations_Instruction_col	0.001981
Prior_Conviction_Episodes_Felony	0.001875
Education_Level	0.001782
Prior_Conviction_Episodes_GunCharges	0.001249
Prior_Conviction_Episodes_PPViolationCharges	0.001157
Dependents	0.001056
Toddlers	0.001015
Condition_Cog_Ed	0.001
Bachelor's degree	0.000962
Prior_Arrest_Episodes_GunCharges	0.000899
Race	0.000878
Prior_Conviction_Episodes_DomesticViolenceCharges	0.000648
Associate's degree	0.000576

Family income (past 12 months, use ADJINC to adjust FINCP to constant dollars)	0.000501
Married-couple family: Husband not in LF, wife in LF	0.000473
Other family: Male householder, no wife present, not in LF	0.000447
Violations_FailToReport_col	0.000439
House on one to less than ten acres	0.000436
GED or alternative credential	0.000414
Violations_MoveWithoutPermission	0.000393
Program_UnexcusedAbsences	0.000384
Other family: Female householder, no husband present, not in LF	0.000357
Employment_Exempt	0.000321
Property value	0.000243
House on less than one acre	0.000233
Prior_Arrest_Episodes_DVCharges	0.000225
k-3	8.96E-05
G7-11	8.79E-05
Married-couple family: Husband and wife in LF	8.05E-05
Married-couple family: Neither husband nor wife in LF	7.96E-05
Prison_Offense	1.16E-05
House on ten or more acres	6.61E-06
	0
Prior_Conviction_Episodes_Misd	0
Prior_Revocations_Probation	0

Prior_Conviction_Episodes_Drug	0
Prior_Conviction_Episodes_Prop	0
Condition_Other	-1.18E-05
Regular high school diploma	-2.93E-05
Total Lot Size (ACR)	-5.41E-05
N/A (GQ/not a one-family house or mobile home)	-8.24E-05
Wages or salary income past 12 months (use ADJINC to adjust WAGP to constant dollars)	-0.00013
DrugTests_Meth_Positive	-0.00019
Other family: Female householder, no husband present, in LF	-0.0002
Residence_PUMA	-0.00031
DrugTests_Other_Positive	-0.00038
Other family: Male householder, no wife present, in LF	-0.00038
Income-to-poverty ratio recode	-0.00044
Prior_Arrest_Episodes_Drug	-0.00044
Violations_ElectronicMonitoring	-0.00048
DrugTests_Cocaine_Positive	-0.00052
Married-couple family: Husband in labor force, wife not in LF	-0.00053
G4-6	-0.00053
Grad Degree	-0.00063
Some college, no degree	-0.00066
N/A (GQ/vacant/not a family/same-sex married-couple families)	-0.00068
12th grade - no diploma	-0.00075

DrugTests_THC_Positive	-0.00097
Prior_Conviction_Episodes_Viol	-0.00104
Prior_Arrest_Episodes_Violent	-0.00171

Importance analysis was performed using permutation analysis where the value of each input was varied to determine the changes small permutations have on each variable. Those variables where small changes result in larger changes of the predicted recidivism are more important to the model's predictions.

Section 2.01 Data Formatting

The design process began by normalizing each data type to a format that can be interpreted by the machine learning regression model. This included formatting strings to numerical values. This typically consisted of two approaches: First, for variables with clear trends, the variables were scaled 0 to 1 (e.g., number of dependents). Second, for variables with no clear linear transform, multiple Boolean variables were created for each possibility. The result of this process was a 1D array of floats ranging from 0 to 1 that represented the inputs for each individual that was provided to the model to predict their probability of recidivism.

3. Models

With the normalized data, the process of selecting a prediction model began. At a high level, two approaches were attempted, first a DNN regressor model and then a variety of traditional ML regression and classification models. The model found to perform best was a voting-based ensemble of multiple traditional ML models. The ensemble included four models that outperformed other model types and outperformed the individual models in combination. A genetic algorithm was used to tune model parameters before the model down selection and on the complete ensemble. The GA used the Briar Score as the evaluation functions for parameter tuning. Parameters included the number of trees in a random forest for example.

To evaluate each model, the training data was split into ten segments. For each evaluation, nine of the ten segments were used to generate the training data set and the holdout segment was used as the test data set. This process was used to initially evaluate all models and perform the genetic algorithm to optimize the model parameters.

The following models were attempted and resulted in the Briar Score noted:

Model	**Briar Score**
*VotingRegressor Ensemble post GA	0.158128
*VotingRegressor Ensemble	0.159285
GradientBoostingRegressor	0.160589
RandomForestRegressor	0.16154
ExtraTreesRegressor	0.168156
RandomForestClassifier_proba	0.17066
ExtraTreesClassifier_proba	0.173432
GradientBoostingClassifier_proba	0.174814
BaggingRegressor	0.174933
BaggingClassifier tree_proba	0.179978
MultinomialNB	0.181583
AdaBoostRegressor	0.182334
*StackingRegressor Ensemble	0.191164

CategoricalNB	0.197266
BernoulliNB	0.201911
DNN Regressor	0.205252
ComplementNB	0.209128
AdaBoostClassifier_proba	0.245182
AdaBoostClassifier	0.247228
ExtraTreesClassifier	0.248337
GaussianNB	0.251246
*VotingClassifier Ensemble	0.251663
RandomForestClassifier	0.252772
GradientBoostingClassifier	0.256098
BaggingClassifier tree	0.272727
Decision Tree Regressor	0.293792
Decision Tree Classifier	0.298226

*The ensemble methods consisted of the following models: Gradient Boosting Regressor, Random Forest Regressor, Extra Trees Regressor, and Random Forest Classifier.

As states, the Briar Score was used as both the official test metric and the metric used to evaluate model performance and evaluation. It is an effective way to measure the performance of the ML model's ability to predict recidivism. The 0.5 threshold did not appear to affect the results. From an ethical perspective, it makes sense to penalize false positives but that is beyond my expertise to determine how to weigh it. Because it was not the primary portion of this challenge, the fairness model was not used to train these results. Instead, every attempt was made to predict each individual’s outcome as accurately as possible with the expectation this would lead to fair predictions.

4. Future Considerations

The largest difference that could have improved this challenge is that the training data set did not include all of the potential inputs of the test data set. This difference resulted in the need to redo the data normalization process for the 2^{nd} and 3^{rd} test sets. This complication ate into the time in these challenges. If not possible to provide all inputs in the provided training set, the ranges could be provided in the data description document.

5. Conclusion

Given the variable performance, it is clear that maintaining employment throughout parole is key to preventing recidivism. It is not obvious if employment is what causes the decrease in recidivism or if those more likely to return to prison are also those less likely to maintain employment. However, it is apparent that this is a valuable indicator for predicting recidivism.