An Exploration of Classification Model Building:

Predicting Recidivism Within Three Years of Release from Prison

Kayla Strunk

Department of Computer Science, Denver University

COMP4448: Data Science Tools 2

Dr. Neba Nfonsang

November 14, 2022

Introduction

Purpose and Significance of Project

According to the United States Department of Health and Human Services every year over 600,000 individuals are released from both state and federal prisons; more than two-thirds of those released from prison are rearrested and half are reincarcerated (USDHHS, 2022). This project aims to build a classification model that can accurately predict if an individual will be rearrested for a new crime within three years of release with supervision. The models that will be explored include Decision Tree Classification, Random Forest Classification, Support Vector Classification, and K-Nearest Neighbors Classification. With the ability to predict recidivism within three years of release more effort can be put into rehabilitation efforts for at risk persons.

Research Question

How does the performance of a Decision Tree Classifier compare to a Random Forest Classifier, Support Vector Classifier and K-Nearest Neighbors Classifier in predicting recidivism rates within three years of release?

Description of Data set

The data set used for this investigation was obtained from the United States Department of Justice website and contains data on the recidivism rates for approximately 26,000 individuals in Georgia between 2013 and 2015. The U.S. Department of Justice defines recidivism in the scope of this data set as an individual being arrested for a new felony or misdemeanor crime within three years of supervision start date (USDOJ, 2021). The original data set contains 53 feature variables and 1 target variable. Of the 53 feature variables there were 23 categorical features and 20 numeric features which include: individual's age at release, gender, race, prior

arrests, prior convictions, and education level. The target variable is a binary variable noting if the individual was arrested for a crime within three years of release.

Data Preprocessing

Data Preparation

To ensure that the data would be sufficient to train the models of interest, the data was cleaned to remove any instances where the individual was missing information in at least one of the input variables. Additionally, the following variables would be removed from the data set due to the percentage of missingness being greater than 20%: Avg Days per Drug Test, Drug Tests THC Positive, Drug Tests Cocaine Positive, Drug Tests Meth Positive, Drug Tests Other Positive. Because these variables all related to drug screening, the impact of biasing the data by removing them was minimal.

After removing missing data and ensuring the variables were all the correct data type, dummy variables were created for the seven input variables of categorical data type that were not already binary and the 20 numeric input variables were scaled using a standard scaler. Standard scaling ensures that there would be no introduction of biases based on the largeness or smallness of the numbers of different scale. Given the large number of feature variables, a principal component analysis was run for comparison to see if model performance would improve by reducing the dimensions of the data.

Exploratory Data Analysis, Visualization, and Data Splitting

The distributions of the numeric values prior to scaling were explored using histograms (Appendix Figure A1).For features such as ‘prior arrest episodes felony’ and ‘prior arrest episodes misdemeanors’ a large distribution of the population lies on the right-hand side of the histograms. This is due to the data being capped at a certain value during the data cleaning

process. Descriptive statistics were calculated as well to observe features such as the skew of the input data points.

Count plots were created for each of the categorical variables to visualize the distribution across these features. (Figure A2, A3, A4). It was found that for the gender category all the individuals were male. Because there is no variability within this feature, it will likely not provide any useful insight when building the models; therefore, it was removed.

The data was split with 70% of the data being in the training set and 30% of the data being in the test set. This train-test split follows the split outlined originally by the US DOJ. A random state of 42 was initiated to allow the results of this project to be reproduceable. Once the data was split into training and test sets, the distribution of the output variables were graphed to ensure that the distributions of the populations were similar.

Model Building and Evaluation

Model Building and Optimization

To build the models they were first initialized and fit to the data prior to dimension reduction and then the models were fit using the data after dimension reduction using principal component analysis(PCA) with two principal components. Four predictions were run in total, one for each of the classification models: predictions on the training set, predictions on the test set, predictions on the PCA training set, and predictions on the PCA test set (Appendix Figures B1 and B2). After predictions were ran an accuracy score was obtained for each of the models' training and test sets for both non-PCA and PCA data.

Based on accuracy scores obtained from the models when building, the non-PCA data performed better; therefore, model tuning was done solely on the models using the non-PCA data. Grid search cross validation was used to tune the hyperparameters of each model. For the

Decision Tree Classifier, the parameters splitter, max features, and max depth were tuned; for Random Forest Classifier the parameters bootstrap, max features, and n-estimators were tuned; for the SVC model the parameters kernel and gamma were tuned; for the KNN model the n-neighbors parameter was tuned. After tuning, an accuracy score was again obtained for each model based on the best parameters after cross validation.

Model Comparison

Prior to model tuning both the Decision Tree model and Random Forest model heavily overfit the training set data; this was not improved by dimensionality reduction through PCA. After tuning, the Decision Tree and Random Forest models did not perform great with the Random Forest model still overfitting the training set. The SVC model performed the best with a 74% accuracy on the training set and 72% accuracy on the test set (Appendix Table B1).

Conclusion

Lessons Learned and Recommendations

Based on the tuning of the models the SVC model with a gamma equal to 0.01 and kernel equal to ‘rbf’ performed the best. The SVC model also performed the best overall pre-tuning with an 80% accuracy on the training set and 72% accuracy on the test set. Although only obtaining a 72% accuracy on the test set post-tuning, this model could be a good starting point in predicting an individual’s chance of recidivism within three years of release. Additional data preparation such as different dimension reduction techniques or scaling may be of use to explore as well to improve the accuracy of the model.

Overall, this data set does pose limitations as it does not report those individuals who may have committed a crime outside of the three-year window, and it may or may not be generalizable to populations outside of Georgia where the data was obtained. The best

recommendation for further research to improve the SVC model would be to obtain a wider scope of data. Data for a wider range of the United States' population would make the model more generalizable and collecting data for a larger time post release may also improve the model's accuracy.

Appendix A

Appendix A contains figures and tables relevant to the exploratory data analysis

Figure A1. Histograms depicting the distribution of the numeric variables

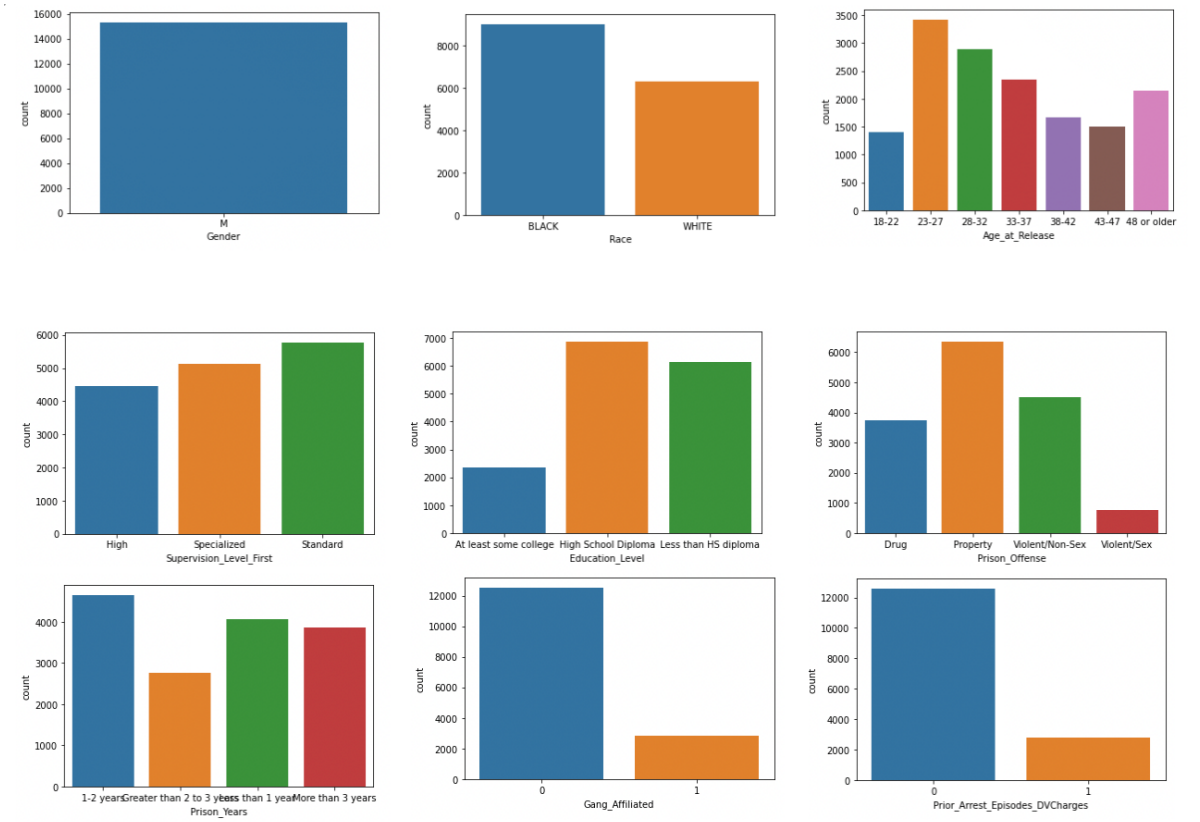

Figure A2: Count plots for categorical variables 1-9

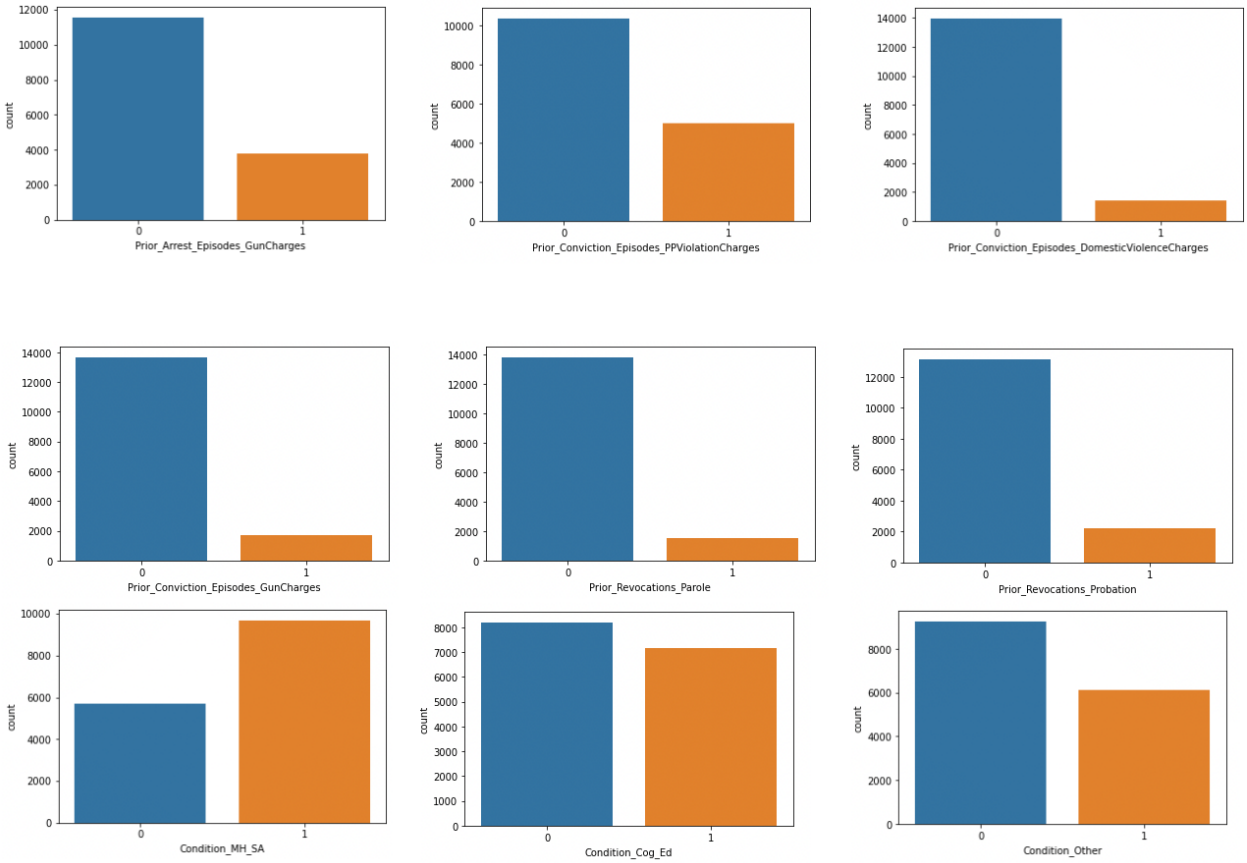

Figure A3: Count plots for categorical variables 10-18

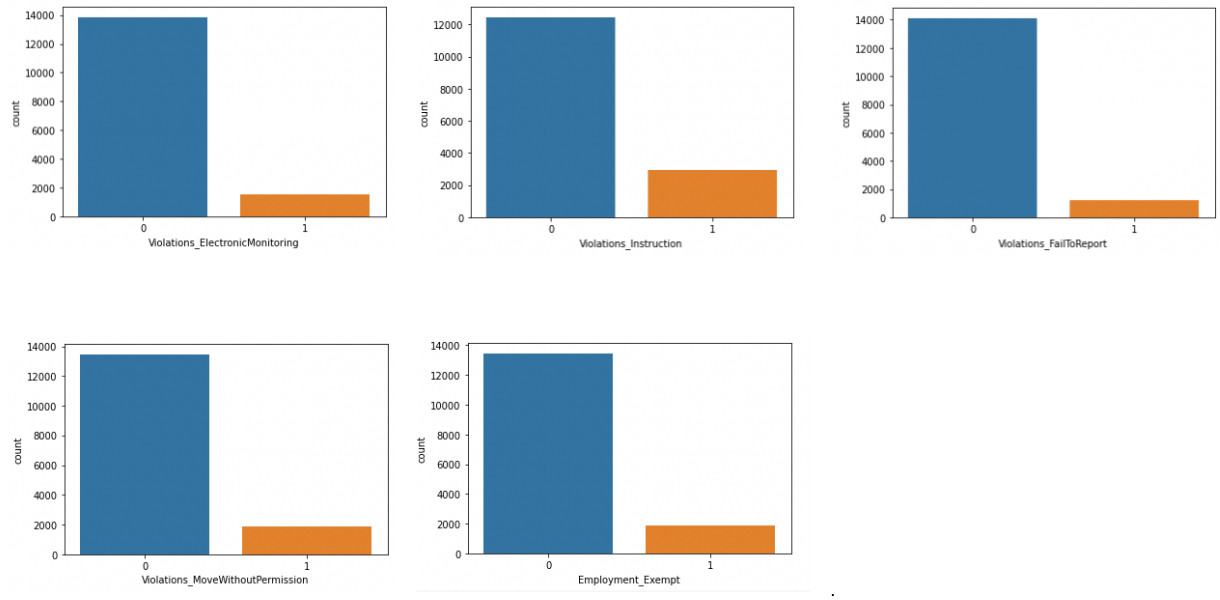

Figure A4: Count plots for categorical variables 19-23

Appendix B

Appendix B contains figures and tables relevant to the building of the classification models as well as their respective accuracy scores.

```
#initialize models
tree_class = DecisionTreeClassifier()
forest_class = RandomForestClassifier()
SVC = SVC()
KNN = KNeighborsClassifier()

#fit models
tree_class.fit(X_train, y_train)
forest_class.fit(X_train, y_train)
SVC.fit(X_train, y_train)
KNN.fit(X_train, y_train)

#get train set predictions
tree_train_pred = tree_class.predict(X_train)
forest_train_pred = forest_class.predict(X_train)
SVC_train_pred = SVC.predict(X_train)
KNN_train_pred = KNN.predict(X_train)

#get test set predictions
tree_test_pred = tree_class.predict(X_test)
forest_test_pred = forest_class.predict(X_test)
SVC_test_pred = SVC.predict(X_test)
KNN_test_pred = KNN.predict(X_test)
```

Figure B1. Code for initializing, fitting, and predicting the non-PCA training and test sets on the classification models.

```
#fit PCA models
pca_tree = tree_class.fit(pca_Train, y_train)
pca_forest=forest_class.fit(pca_Train, y_train)
pca_SVC = SVC.fit(pca_Train, y_train)
pca_KNN = KNN.fit(pca_Train, y_train)

#get train set predictions
pca_tree_train_pred = pca_tree.predict(pca_Train)
pca_forest_train_pred = pca_forest.predict(pca_Train)
pca_SVC_train_pred = pca_SVC.predict(pca_Train)
pca_KNN_train_pred = pca_KNN.predict(pca_Train)

#get test set predictions
pca_tree_test_pred = pca_tree.predict(pca_Test)
pca_forest_test_pred = pca_forest.predict(pca_Test)
pca_SVC_test_pred = pca_SVC.predict(pca_Test)
pca_KNN_test_pred = pca_KNN.predict(pca_Test)
```

Figure B2. Code for fitting and predicting the PCA training and test sets on the classification models.

Accuracy Scores			
	Non-PCA Classification Models	PCA Classification Models	Non-PCA Classification Models Post Tuning
Decision Tree Classifier	Train Set Accuracy: 1.0 Test Set Accuracy: 0.63	Train Set Accuracy: 1.0 Test Set Accuracy: 0.55	Train Set Accuracy: 0.75 Test Set Accuracy: 0.66
Random Forest Classifier	Train Set Accuracy:1.0 Test Set Accuracy: 0.72	Train Set Accuracy: 1.0 Test Set Accuracy: 0.59	Train Set Accuracy: 1.0 Test Set Accuracy: 0.73
SVC	Train Set Accuracy: 0.80 Test Set Accuracy: 0.72	Train Set Accuracy: 0.64 Test Set Accuracy: 0.65	Train Set Accuracy: 0.74 Test Set Accuracy: 0.72
KNN	Train Set Accuracy: 0.76 Test Set Accuracy: 0.64	Train Set Accuracy:0.72 Test Set Accuracy:0.59	Train Set Accuracy: 0.70 Test Set Accuracy: 0.69

Table B1. Summary table for each classification model's accuracy scores on the Non-PCA data set, PCA data set, and tuned models.

Works Cited

U.S. Department of Health and Human Services. (2022, November 11). *Incarceration and Reentry.* https://aspe.hhs.gov/topics/human-services/incarceration-reentry-0

Office of Justice Programs. (2022, November 11). *NIJ Recidivism Challenge*. https://data.ojp.usdoj.gov/stories/s/daxx-hznc