**Predicting Criminal Recidivism Using Machine Learning Algorithms**

**Capstone Project 2**

**Springboard Data Science Career Track**

**By**

**Shishay Bisrat**

**February 2018**

**Predicting Criminal Recidivism Using Machine Learning**

1. **Abstract**

Recidivism is a major societal issue that costs the US lives and resources. An accurate recidivism prediction would have a major impact to individuals as well as the society. I use Logistic Regression and Random Forest to build models that predict criminal recidivism using 3-year recidivism data from the state of Iowa. I experimented with various resampling techniques to balance the data. Balancing the data improved the performance of both logistic regression and random forest models. Generally, the random forest performed better than the logistic regression model. In both random forest and logistic regression, the models trained on data resampled by SMOTEENN are the best models (based on performance metrics].

**2. Background Information**

Criminal recidivism refers to the act of a person repeating a criminal offense after being released from prison for a similar offense. Statistics show that recidivism is a major issue in the US: The Bureau of Justice Statistics (BJS) reported 67 percent and 76.6 percent of state prisoners released in 2005 were re-arrested with three and five years of release respectively (Durose et al, 2014). The justice system in the US predicts an individual’s recidivism potential when making key decisions such as parole, probation, and sentencing. The court systems rely on a software package called Correctional Offender Management Profiling for Alternative Sanctions (COMPAS) to predict the likelihood of recidivism. The recidivism prediction made by an individual could be biased depending on one’s life experiences and perspectives. Software packages could also yield biased results if the weight of features is inexplicably fixed and provided by biased individuals. Machine Learning algorithms can be useful in limiting biases and improving recidivism predictions (Teich, 2018). Some justice system institutions such as the Pennsylvania Board of Probation and Parole have begun using machine-learning predictions to help in parole release decisions. It is expected that more states will use the promising field of machine learning to limit biases and make useful contribution in accurately forecasting recidivism (Berk, 2016).

**3. Motivation**

Recidivism is a major societal issue that costs lives and resources. The per capita cost of incarceration in the US was $30,621 in 2014 (James, 2016). An accurate recidivism prediction would have a major impact to individuals as well as the society. Criminal institutions would target their stretched resources on dangerous offenders increasing the safety of citizens. Non-dangerous, one-time offenders, with little likelihood of recidivism would be treated fairly, encouraging them to integrate to society with little supervision. This would free resources and ensure sensible use of resources strictly on dangerous offenders. For example, the time probation supervision officers assign to released offenders is the same irrespective of the risk of recidivism. Supervision officers can get better results of reducing recidivism by focusing their efforts on offenders of high-risk recidivism.

**4. Data and Data Wrangling**

The original data set consisted of 21,646 rows and 17 features. Rows that do not have any value for all the features were removed from the dataset. An “unknown” group was added as a possible value for each feature and any entries with “NaN” values were assigned to the “unknown” group. The dataset is reasonably clean and the number of records in the “unknown group” is few. Eight out of the 17 features have no missing values and 3 other features have less than 35 missing values. The exception is the feature “Main Supervising District” which has 8,470 missing values.

The cleaned dataset consists of 21,611 rows and 17 columns (Table 1). Each row represents a prisoner released from prison in the five-year period between 2010 and 2014. One of the columns, whether an individual returned to prison within 3 years of release, is used as the label [target feature]. Five of the attributes [*'Recidivism Type'*, *'Days to Recidivism'*, *'New Conviction* *Offense Classification'*, *'New Conviction Offense Type'*, *'New Conviction Offense Sub* T*ype*'] represent information collected for recidivist prisoners only and hence are not considered with the predictor features.

|  |  |  |
| --- | --- | --- |
| **Features** | **Possible Values** | **NaNs** |
| Year Released | 2010, 2011, 2012, 2013, 2014 | 0 |
| Recidivism Year | 2013, 2014, 2015, 2016, 2017 | 0 |
| Race/ Ethnicity | White (Non-Hispanic), Black (Non-Hispanic), White, American Indian/ Alaska native, Asian or Pacific Islander, Black (Hispanic) | 35 |
| Sex | Male, Female | 3 |
| Age at release | Under 25, 25-34, 35-44, 45-54, over 55 | 3 |
| Convicting Offense Classification/ New Convicting Offense Classification | D Felony, C Felony, Aggravated Misdemeanor, B Felony, Felony – Enhancement to original penalty, Felony – enhanced, Serious Misdemeanor, Special Sentence 2005, Felony – Mandatory Minimum, Other Felony, A Felony, Simple Misdemeanor, Sexual Predator Community Supervision, Other Misdemeanor | 0/1963 |
| Convicting Offense Type/ New Convicting Offense Type | Drug, Property, Violent, Public Order, Other | 0/1963 |
| Convicting Offense Subtype/ New Convicting Offense Subtype | Trafficking, Assault, Burglary, Theft, OWI, Sex, Forgery/Fraud, Drug Possession, Other Criminal, Other Violent, Traffic, Murder/ Manslaughter, Weapons, Alcohol, Vandalism, Robbery, Other Drug, Other Public Order, Arson, Sex Offender Registry/ Residency, Flight/Escape, Special Sentence Revocation, Kidnap, Prostitution/Pimping, Stolen Property, Animals | 0/1982 |
| Release Type | Parole, Discharged-End of Sentence, Parole Granted, Discharged – Expiration of Sentence, Special Sentence, Paroled with Immediate Discharge, Released to Special Sentence, Paroled to Detainer (INS, Out of State, Iowa, U.S. Marshall), Interstate Compact Parole | 1762 |
| Main Supervising District | 1JD, 2JD, 3JD, 4JD, 5JD, 6JD, 7JD, 8JD, ISC | 8470 |
| Recidivism Return | Yes, No | 0 |
| Recidivism Type | New Recidivism, Tech, New | 2963 |
| Days to Recidivism | 0 to 3 years | 0 |
| Target Population | Yes, No | 0 |

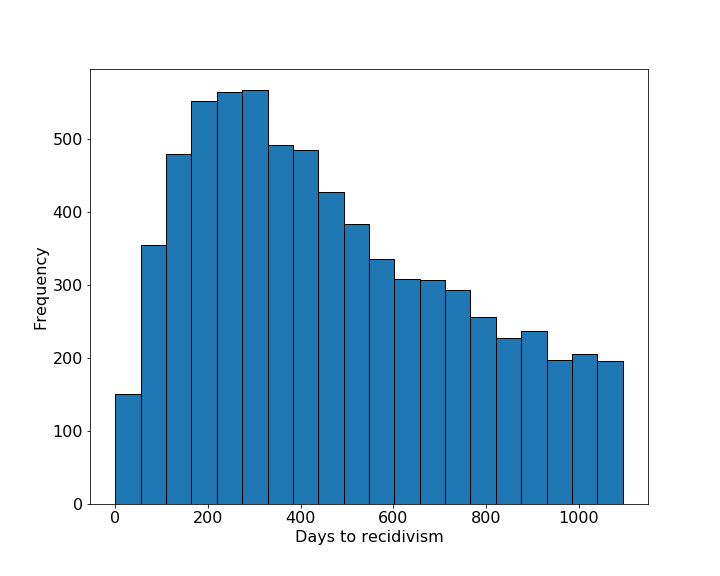
**Table 1:** The attributes, the possible values for each attribute, and the number of missing values for each attribute.

**5. Exploratory Data Analysis (EDA)**

The data shows 7,027 individuals returned to prison; therefore the positive to negative class ratio is ~1: 2 (Figure 1). Although not as extreme as in problems in areas such as fraud detection and infant mortality, there is imbalance in the data that requires balancing the data before applying machine-learning algorithms.



**Figure 1:** The ratio of recidivism to non-recidivism is 1:2.

The prisoners in the study were released in the five-year period from 2010 to 2014. Each prisoner was monitored for 3 years after release as studies suggest most of recidivism usually occurs with in 3 years (Figure 2). 

**Figure 2:** The distribution of “days to recidivism” shows most recidivism occurs between 150 to 400 days. Most recidivism studies focus in the first 3 years after prisoner release since most prisoners, if they are ever to recidivate, they recidivate within three years.

The recidivism rate by year ranged from 29.7 (for prisoners released in 2011) to 35.4% (for prisoners released in 2014). The average recidivism for the five-year period is 32.5% (Figure 3). Iowa’s recidivism rate is less than the national average.

****

**Figure 3:** The ratio of recidivism to non-recidivism is highest for the year 2014. The average recidivism for the 5-year period is 32.5%.

The recidivism data consists of a column that indicates the race/ethnicity of the prisoners. The data shows that the majority of the prisoners are White (non-Hispanic) followed by Black (non-Hispanic), White (Hispanic) and American Indian/ Alaska native. This roughly mirrors the racial composition of Iowa, which reported a population of White (91.4%), Black (3.7%), American Indian/Alaska Native (0.5%), Asia (2.5%), and Hispanic/ Latino (5.8%) for the year 2017. The recidivism rate for whites (non-Hispanic), blacks, whites (Hispanic), and American Indians/Alaska natives is 33.4, 32.8, 20.9, and 39.4% respectively (Figure 4). This is in contrast to many findings that showed recidivism is highly correlated to race and African Americans are more likely than Whites to be rearrested, reconvicted, or returned to prison with or with out a new sentence (Langan and Levin, 2002). However, the dataset here is limited to 5-year period and state of Iowa only and it is not prudent to make such comparisons.

****

**Figure 4:** The recidivism rate for the 5-year period grouped by race.

The distribution of prisoners by age shows that the 36.6% of the total released prisoners belong to that age group of 25-34. The representation of the other age groups in the released prisoners is 35-44 (23.9%), under 25 (17.8%), 45-54 (16.8%), and over 55 (4.9%). The recidivism rate for the age groups under 25, 25-34, 35-44, 45-54, and over 55 is 37.1, 34.6, 31.8, 28.0, and 19.1% respectively (Figure 5). This suggests recidivism is highest for the age group under 25 and the lowest for age group over 55 years. This agrees with studies that showed age is a predictor for recidivism, and the older the prisoner the more the likelihood of recidivism (Langan and Levin, 2002).



**Figure 5:** The recidivism rate for the 5-year period grouped by age group.

The gender distribution of the released prisoners shows males and females make up 87.4 and 12.6% respectively. The recidivism rate for males and females is 33.2 and 27.7% respectively (Figure 6). Previous studies (Durose et. al., 2014) have shown that males recidivate at higher rate than females. This is true across many offenses. Piquero et. al. (2015) found that males are more likely to recidivate than females for violent offenses. Similarly, Stahler et. al. (2013) concluded that male drug offenders are more than twice as likely to recidivate than females.

****

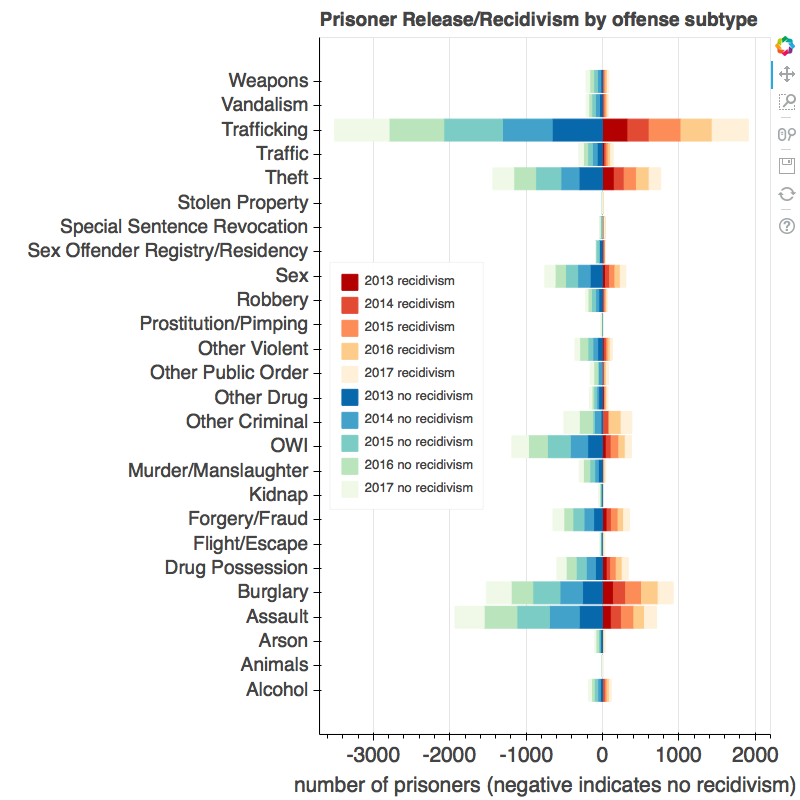
**Figure 6:** The recidivism rate for the 5-year period grouped by gender. Although female prisoners make up only 12.6% of the total number of prisoners, the recidivism rate is 27.7%.

Based on the offense class of the prison system (Figure 7), the prisoners can be grouped to D Felony (40.5%), C Felony (26.2%), aggravated misdemeanor (18.8%), B Felony (6.9%), Felony – enhanced to original penalty (5.6%), and Felony – enhanced (1%). Other classifications, such as serious misdemeanor, special sentence, A Felony, Other Felony, simple misdemeanor made up less than 1% of the total. The recidivism for each offense class are D Felony (30.2%), C Felony (40.0%), aggravated misdemeanor (33.1%), B Felony (32.0%), Felony – enhanced to original penalty (40.4%), and Felony – enhanced (35.0%).



**Figure 7:** The recidivism rate for the 5-year period grouped by offense classification.

Trafficking related offenses make up 25% of the total offenses. The top list of offenses include assault (12.2%), burglary (11.3%), theft 10.1%), OWI (7.2%), sex (4.9%), forgery/ fraud (4.6%), drug possession (4.3%), other criminal (4.1%), other violent (2.3%). Murder/manslaughter, weapons, alcohol, vandalism, and robbery each account for 1 to 2% of total offenses (Figure 8). The highest rate of recidivism is observed in special sentence revocation (46.0%), flight/escape (44.9%), other criminal (43.1%), alcohol (38.6%), burglary (37.8%), stolen property (37.5%), and drug possession (36%).



**Figure 8:** The recidivism rate for the 5-year period grouped by offense class subtype. Trafficking, assault, burglary, OWI, sex, and forgery/fraud consistently make the top of convicting offenses.

**5.1. Summary**

Exploring the recidivism data show that there is possible correlation between demographic (age group, gender, race, etc) and criminal history (offense classification, offense type, etc) and the risk of recidivism. Although including gender and race in predicting recidivism has been controversial, Berk et. al. (2012) argues for including them to solve recidivism and its consequences. Taxman et. al. (2013) supports for inclusion of gender and race data arguing demographic-neutral models forego important information in the fight against recidivism.

**6. Machine Learning Methods**

**6.1. Resampling**

The data shows 7,027 individuals returned to prison; therefore the positive to negative class ratio is ~1:2. Although the class imbalance is not as extreme as in problems in areas such as fraud detection and infant mortality, there is imbalance in the data that requires balancing the data before applying machine-learning algorithms. I have used under-sampling, over-sampling and combined (under-sampling followed by over-sampling) to balance the data. The resampling step can be applied before or after splitting the dataset into train and test datasets.

**6.1.1 Under-sampling methods**

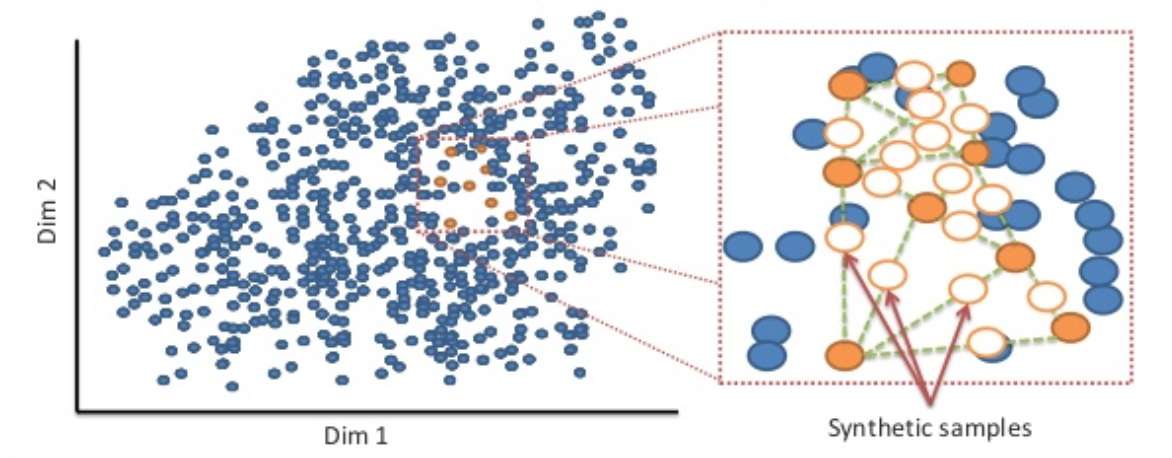
I used Random Under-Sampling (RUS), RUS/TomekLinks, and Edited Nearest Neighbor (ENN) to sample the subset of the majority class in order to reduce the size of the majority class and balance the dataset.

In random under-sampling (RUS), a subset of the majority class are randomly included creating a smaller, balanced dataset. The loss of information with the removal of data from the majority class is the main disadvantage of this resampling method.

TomekLinks provides a guided undersampling/oversampling of a dataset by identifying and removing “noisy” data (TomekLinks) at the boundary of classes. Elhassan et al. (2016) suggest removing noise observation from majority class followed by RUS improve the performance of classification by reducing the chance of information loss.

In Edited Nearest Neighbor (ENN), the majority class is under-sampled by removing points whose class label differs from a majority of its *k* nearest neighbors.

**6.1.2. Over-sampling methods**

Oversampling methods balance the dataset by increasing the size of the minority class. The two major disadvantages of oversampling are the increased possibility of over-fitting and increased learning time due to increase in learning examples. I use Synthetic Minority Oversampling Technique (SMOTE, Chawla et. al., 2002, Figure 9) and Adaptive Synthetic Sampling Approach (ADAYSN, He et.al, 2008) to oversample the minority class.

**Figure 9:** Synthetic Minority Oversampling Technique (Bahnsen et. al.).

SMOTE oversamples the minority class by creating synthetic examples, rather than the traditional way of oversampling with replacement. ADASYN is an extension of SMOTE that creates more samples at the boundaries between the two classes than in the interior of the interior of the minority class.

**6.1.3. Combined over- and under-sampling methods**

SMOTE can generate noisy samples by interpolating new points between marginal data points. TomekLinks and Edited Nearest Neighborhood (ENN) methods can be used to clean the data.

**6.2. Algorithms**

I used logistic regression and random forests for machine learning, algorithms that are commonly used for classification.

**6.2.1 Logistic Regression**

Logistic regression is the common machine-learning tool in classification problems. For binary classification, logistic regression computes the probability of an outcome [between 0 and 1] by computing fitted probabilities of linear combination of predictors. Sigmoid function is used to constrain the probability between 0 and 1.



**Figure 10:** Sigmoid function.

The procedures below were followed to build the model.

1. Separate the attributes into categorical (nominal), numerical (non-categorical), and binary. Encode the categorical attributes
2. Specify predictor features and target features. The target feature in this project is recidivism of a prisoner..
3. Split the dataset into training and test datasets (80% train and 20% test data set).
4. Use Grid Search and Cross-Validation (using k-fold cross validation with a k= 5) and determine the optimum regularization parameter.
5. Train the best algorithm on the train set and test the algorithm on the test set.
6. Evaluate the classifier using performance metrics. Use classification report (from package imblearn; Lemaire et. al., 2017) to evaluate a classifier trained using the original data.

**6.2.1.1. Parameters**

The logistic regression function in sklearn has three options for the norms used in penalization. I used the default, norm L2 regularization, in this study.

**6.2.1.2. Results**

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | | PP | NP | SS | Accuracy | Precision | Recall | TNR |
| LR\_NoResample | Train | 0.32 | 0.68 | 17288 | 0.684 | 0.546 | 0.141 | 0.944 |
| Test | 0.33 | 0.67 | 4323 | 0.675 | 0.539 | 0.140 | 0.940 |
| LR\_ENN\_Split | Train | 0.605 | 0.395 | 9300 | 0.725 | 0.734 | 0.856 | 0.525 |
| Test | 0.601 | 0.399 | 2326 | 0.718 | 0.726 | 0.850 | 0.518 |
| LR\_Split\_ENN | Train | 0.558 | 0.442 | 10085 | 0.721 | 0.731 | 0.790 | 0.633 |
| Test | 0.325 | 0.675 | 4323 | 0.559 | 0.407 | 0.790 | 0.447 |
| LR\_ENN\_selMODE\_Split | Train | 0.432 | 0.568 | 12979 | 0.667 | 0.625 | 0.571 | 0.740 |
| Test | 0.439 | 0.561 | 3245 | 0.667 | 0.635 | 0.570 | 0.744 |
| LR\_Split\_ENN\_selMODE | Train | 0.380 | 0.620 | 14876 | 0.68 | 0.608 | 0.444 | 0.824 |
|  | Test | 0.318 | 0.682 | 4323 | 0.677 | 0.490 | 0.446 | 0.784 |
| LR\_RUS\_Split | Train | 0.499 | 0.501 | 12648 | 0.626 | 0.617 | 0.656 | 0.595 |
| Test | 0.512 | 0.488 | 1406 | 0.617 | 0.619 | 0.656 | 0.577 |
| LR\_Split\_RUS | Train | 0.5 | 0.5 | 12584 | 0.625 | 0.619 | 0.649 | 0.601 |
| Test | 0.34 | 0.66 | 2162 | 0.621 | 0.46 | 0.654 | 0.604 |
| LR\_RUS\_Tomek\_Split | Train | 0.502 | 0.498 | 11243 | 0.635 | 0.628 | 0.666 | 0.603 |
| Test | 0.493 | 0.507 | 2811 | 0.604 | 0.589 | 0.650 | 0.559 |
| LR\_Split\_RUS\_Tomek | Train | 0.500 | 0.500 | 11276 | 0.627 | 0.619 | 0.663 | 0.591 |
| Test | 0.321 | 0.679 | 4323 | 0.598 | 0.421 | 0.664 | 0.567 |
| LR\_CNN | Train | 0.528 | 0.472 | 10702 | 0.616 | 0.611 | 0.75 | 0.466 |
| Test | 0.514 | 0.486 | 2676 | 0.604 | 0.593 | 0.727 | 0.473 |
| LR\_Split\_SMOTE | Train | 0. 5 | 0. 5 | 23334 | 0.628 | 0.619 | 0.666 | 0.589 |
| Test | 0.325 | 0.675 | 4323 | 0.615 | 0.439 | 0.668 | 0.589 |
| LR\_SMOTE\_Split | Train | 0.503 | 0.497 | 23334 | 0.627 | 0.619 | 0.673 | 0.581 |
| Test | 0.487 | 0.513 | 5834 | 0.636 | 0.612 | 0.691 | 0.583 |
| LR\_Split\_ADASYN | Train | 0.509 | 0.491 | 23768 | 0.594 | 0.59 | 0.661 | 0.524 |
| Test | 0.325 | 0.675 | 4323 | 0.581 | 0.417 | 0.723 | 0.513 |
| LR\_ADASYN\_Split | Train | 0.511 | 0.489 | 23809 | 0.601 | 0.596 | 0.684 | 0.515 |
| Test | 0.505 | 0.495 | 5953 | 0.614 | 0.602 | 0.694 | 0.532 |
| LR\_Split\_SMOTEENN | Train | 0.554 | 0.446 | 8196 | 0.798 | 0.794 | 0.857 | 0.724 |
| Test | 0.325 | 0.675 | 4323 | 0.558 | 0.4 | 0.718 | 0.481 |
| LR\_SMOTEENN\_Split | Train | 0.471 | 0.529 | 11135 | 0.869 | 0.913 | 0.798 | 0.932 |
| Test | 0.470 | 0.530 | 2784 | 0.871 | 0.912 | 0.804 | 0.931 |

**Table 2:** The results of logistic regression classifier trained on data resampled using different techniques.

**6.2.1.3. Performance evaluation**

I used performance metrics values (accuracy, precision, recall, TNR, AUC) and curves (confusion matrix, precision-recall curve, receiver operating characteristic) to evaluate and visualize the performance of the classifiers.

True Positive (TP, also called hits) refers to a result that correctly identifies a condition that is present. A recidivism prediction result is a true positive when a prisoner is identified as recidivistic and he/she commits an offense after release from prison.

True Negative (TN) refers to a result that correctly identifies the absence of a condition. A recidivism prediction result is a true negative when a prisoner is identified as *not* recidivistic and he/she *does not* commit an offense after release from prison.

False Positive (FP) refers to a result that incorrectly identifies an absent condition as present. A recidivism prediction result is a false positive when a prisoner is identified as recidivistic even though he/she does not commit an offense after release from prison. A false positive represents a waste of limited resources on people who would not recidivate. It can also be viewed as injustice since a prisoner can be negatively affected by decisions made based on inaccurate predictions.

False Negative (FN) refers to a result that fails to correctly identify a present condition. A recidivism prediction result is a false negative when a prisoner who is not identified as recidivistic commits an offense after release from prison. Therefore false negative represents a crime that could have been prevented with accurate prediction.

The TP, TN, FP, and FN are normally given in a table format called confusion matrix (Table 3).

|  |  |  |
| --- | --- | --- |
| **Predicted Condition**  **True Condition** | **Condition absent** | **Condition present** |
| **Condition absent** | True Negative (TN) | False Positive  (FP) |
| **Condition present** | False Negative (FN) | True Positive  (TP) |

**Table 3:** Confusion matrix.

The major indicators of performance can easily be derived from the confusion matrix.

**6.2.1.4. Best Logistic Regression Model**

The performance indicators show the best model is the model built by using SMOTEENN resampling technique (Table 2, Figure 11), followed by splitting the dataset into train and test sets. The accuracy, precision, recall, and TNR for the model are 0.87, 0.91, 0.80, and 0.93 respectively. High precision implies there are few false positives (i.e. there is high chance that the predicted recidivists are actually recidivist and would commit similar offense if released). Recall refers to the percentage of recidivists who are correctly identified as recidivists. High recall implies there are few false negatives and the model would identify a high percentage of recidivists (i.e. there is high chance that the model would correctly identify a recidivist). High TNR indicates the number of false positives is relatively small (i.e. if the model predicts non-recidivism, there is high chance that the prisoner would not recidivate). The high precision and recall values show that the model would correctly identify most of the recidivists and there is high chance that those who are identified as recidivists turn out to be recidivists. The high TNR indicates there is small number of false positives and therefore there is high confidence of identifying most of the non-recidivists as non-recidivists.

The accuracy of the model on the train and test sets is 0.87. This shows there is no issue of over-fitting with the model.

|  |  |  |
| --- | --- | --- |
|  | **Train Set** | **Test Set** |
| **Confusion Matrix** |  |  |
| **Precision-Recall Curve** |  |  |
| **ROC** |  |  |

**Figure 11:** Confusion matrix, precision-recall curve, and ROC for train and test set. The data used is resampled using SMOTEENN.

**6.2.2 Random Forests**

Decision trees are generated from the binary splitting of variables at nodes with similar cases grouped in the same branches (Breiman, 2001). Random Forest grows many classification trees with each tree giving a classification (which is commonly called a vote for that class). The random forest chooses the classification by choosing the most votes over all the trees in the forest or averaging the trees’ probabilistic prediction. Random Forests are used widely for classification problems due to their high accuracy and efficiency when dealing with large databases.

I use the Random Forest Classifier from the sklearn.ensemble module. In this classifier, the split chosen at each node when constructing the trees is not the best among all features; instead, it is the best among subset of features randomly chosen for each tree. Decision trees are sensitive to the train data and have high variance. The randomness slightly increases the bias. However, the averaging decreases the inherently high variance of decision trees, resulting in a better model.

**6.2.2.1. Parameters**

The two main parameters that need adjustment when using Random Forest Classifier are the number of trees in the forest (n\_estimators) and the size of the subset of features randomly selected at each split (max\_features). Increasing the number of trees improves the model although the improvement becomes less significant beyond some critical number of trees. For classification problems, empirical studies show that the square root of the total number of features is the optimal number to of features to use at each split.. I use OOB score to choose the best number of trees and subset of features [OOB score is a measure of accuracy of a model when it is fed samples not included when training the model]. The average OOB score versus number of estimators plot shows that there is not significant difference in performance whether the number of estimators used is log2 or square root of number of features.



**Figure 12:** Plot showing the grid search results for maximum features and number of estimators used at split for random forest classifier.

**6.2.2.2. Results**

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | | **PP** | **NP** | **SS** | **Accuracy** | **Precision** | **Recall** | **TNR** |
| RF\_NoResample\_Param | Train | 0.324 | 0.676 | 17288 | OOB Score = 0.6426 | | | |
| Test | 0.331 | 0.669 | 4323 | 0.653 | 0.468 | 0.333 | 0.812 |
| RF\_NoResample\_Auto | Train | 0.324 | 0.676 | 17288 | OOB Score = 0.6441 | | | |
| Test | 0.331 | 0.669 | 4323 | 0.647 | 0.454 | 0.326 | 0.806 |
| RF\_ENN\_Split\_Param | Train | 0.648 | 0.352 | 8645 | OOB Score = 0.89 | | | |
| Test | 0.661 | 0.339 | 2162 | 0.895 | 0.932 | 0.907 | 0.870 |
| RF\_ENN\_Split\_Auto | Train | 0.648 | 0.352 | 8645 | OOB Score = 0.85 | | | |
| Test | 0.66 | 0.34 | 2162 | 0.877 | 0.93 | 0.879 | 0.872 |
| RF\_ENN\_Ratio\_Split\_Param | Train | 0.443 | 0.557 | 5375 | OOB Score = 0.95 | | | |
| Test | 0.414 | 0.586 | 1344 | 0.962 | 0.965 | 0.943 | 0.976 |
| RF\_ENN\_Ratio\_Split\_Auto | Train | 0.439 | 0.561 | 5375 | OOB Score = 0.95 | | | |
| Test | 0.43 | 0.57 | 1344 | 0.954 | 0.966 | 0.926 | 0.975 |
| RF\_Split\_ENN\_Param | Train | 0.536 | 0.464 | 10448 | OOB Score = 0.84 | | | |
| Test | 0.33 | 0.67 | 4323 | 0.516 | 0.382 | 0.759 | 0.396 |
| RF\_Split\_ENN\_Auto | Train | 0.537 | 0.463 | 10443 | OOB Score = 0.80 | | | |
| Test | 0.327 | 0.673 | 4323 | 0.532 | 0.389 | 0.758 | 0.421 |
| RF\_Split\_ENN\_Ratio\_Param | Train | 0.542 | 0.458 | 10413 | OOB Score = 0.85 | | | |
| Test | 0.32 | 0.68 | 4323 | 0.519 | 0.378 | 0.78 | 0.396 |
| RF\_Split\_ENN\_Ratio\_Auto | Train | 0.534 | 0.466 | 10529 | OOB Score = 0.80 | | | |
| Test | 0.325 | 0.675 | 4323 | 0.527 | 0.378 | 0.699 | 0.444 |
| RF\_Split\_SMOTE\_Param | Train | 0.5 | 0.5 | 23334 | OOB Score = 0.73 | | | |
| Test | 0.325 | 0.675 | 4323 | 0.639 | 0.439 | 0.4 | 0.754 |
| RF\_Split\_SMOTE\_Auto | Train | 0.5 | 0.5 | 23334 | OOB Score = 0.70 | | | |
| Test | 0.325 | 0.675 | 4323 | 0.635 | 0.432 | 0.392 | 0.752 |
| RF\_SMOTE\_Split\_Param | Train | 0.499 | 0.501 | 23334 | OOB Score = 0.71 | | | |
| Test | 0.505 | 0.495 | 5834 | 0.732 | 0.74 | 0.715 | 0.756 |
| RF\_SMOTE\_Split\_Auto | Train | 0.499 | 0.501 | 23334 | OOB Score = 0.69 | | | |
| Test | 0.503 | 0.497 | 5834 | 0.712 | 0.73 | 0.682 | 0.742 |
| RF\_Split\_ADASYN\_Param | Train | 0.511 | 0.489 | 23858 | OOB Score = 0.75 | | | |
| Test | 0.325 | 0.675 | 4323 | 0.641 | 0.432 | 0.328 | 0.792 |
| RF\_Split\_ADASYN\_Auto | Train | 0.508 | 0.492 | 23734 | OOB Score = 0.73 | | | |
| Test | 0.325 | 0.675 | 4323 | 0.642 | 0.435 | 0.333 | 0.791 |
| RF\_ADASYN\_Split\_Param | Train | 0.51 | 0.49 | 23809 | OOB Score = 0.73 | | | |
| Test | 0.508 | 0.492 | 5953 | 0.742 | 0.783 | 0.682 | 0.805 |
| RF\_ADASYN\_Split\_Auto | Train | 0.508 | 0.492 | 23809 | OOB Score = 0.73 | | | |
| Test | 0.517 | 0.483 | 5953 | 0.738 | 0.792 | 0.67 | 0.811 |
| RF\_Split\_TOMEK\_RUS\_Param | Train | 0.5 | 0.5 | 11292 | OOB Score = 0.59 | | | |
| Test | 0.319 | 0.681 | 4323 | 0.575 | 0.383 | 0.559 | 0.583 |
| RF\_Split\_TOMEK\_RUS\_Auto | Train | 0.5 | 0.5 | 11050 | OOB Score = 0.57 | | | |
| Test | 0.347 | 0.653 | 4323 | 0.587 | 0.428 | 0.561 | 0.6 |
| RF\_TOMEK\_RUS\_Split\_Param | Train | 0.499 | 0.501 | 11243 | OOB Score = 0.56 | | | |
| Test | 0.506 | 0.494 | 2811 | 0.573 | 0.579 | 0.565 | 0.581 |
| RF\_TOMEK\_RUS\_Split\_Auto | Train | 0.502 | 0.498 | 11243 | OOB Score = 0.57 | | | |
| Test | 0.493 | 0.507 | 2811 | 0.568 | 0.565 | 0.543 | 0.594 |
| RF\_Split\_SMOTEENN\_Param | Train | 0.566 | 0.434 | 8202 | OOB Score = 0.95 | | | |
| Test | 0.325 | 0.675 | 4323 | 0.585 | 0.412 | 0.651 | 0.552 |
| RF\_Split\_SMOTEENN\_Auto | Train | 0.556 | 0.444 | 8242 | OOB Score = 0.92 | | | |
| Test | 0.325 | 0.675 | 4323 | 0.595 | 0.415 | 0.599 | 0.593 |
| RF\_SMOTEENN\_Split\_Param | Train | 0.574 | 0.426 | 10156 | OOB Score = 0.96 | | | |
| Test | 0.58 | 0.42 | 2540 | 0.957 | 0.963 | 0.963 | 0.949 |
| RF\_SMOTEENN\_Split\_Auto | Train | 0.575 | 0.425 | 10101 | OOB Score = 0.92 | | | |
| Test | 0.578 | 0.422 | 2526 | 0.946 | 0.969 | 0.937 | 0.959 |

**Table 4:** The results of random forest classifier trained on data resampled using different techniques.

**6.2.2.3. Best Random Forest Model**

The performance evaluation table shows the best model is the random forest model (just like the logistic regression model) built by using SMOTEENN resampling technique, followed by splitting the dataset into train and test sets (Figure 13).

|  |  |  |
| --- | --- | --- |
|  | **Train Set** | **Test Set** |
| **Confusion Matrix** |  |  |

**Figure 13a:** Confusion matrix for both train and test sets.

|  |  |  |
| --- | --- | --- |
| **Confusion Matrix** |  |  |

**Figure 13b:** Precision-recall curve (left) and ROC (right) for random forest trained data resampled by SMOTEENN.

**6.2.2.4. Feature Importance**

Many machine-learning applications have to deal with thousands or tens/hundreds of thousands of variables. In such cases feature selection based on importance is critical to deep understanding of the data and build fast and cost-effective prediction models (Guyon and Elisseeff, 2003). Even for applications that have few features, it is desirable to evaluate the importance of each feature. The feature importance can be positive or negative. The most important factor is the magnitude of the coefficients of the variables. A feature of high importance can significantly sway the prediction of recidivism. A feature of low importance has negligible effect on the prediction of recidivism.

Random forest algorithm in sklearn has an attribute (feature\_importances) that provides the relative importance of each feature used in building the model. The two most important features for classification are race and age (Figure 14). The third most important feature is the ‘Main Supervising District”. It is very difficult, with limited information, to conclusively declare why “Main Supervising District” is important in recidivism. However, recidivism can differ from district to district. For the year 2016, for example, the Sixth Judicial District, reported recidivism rate of 23.1% although the statewide recidivism rate was 31.5% (Sixth Judicial District Department of Corrections Services Annual Report, 2016).



**Figure 14:** Features used in classification ranked by their importance. Black lines for each feature show the standard deviation.

**7. Recommendation**

1. The justice system uses recidivism prediction to make decisions on probation, parole, and sentencing. The predictions based on software or individuals can be biased. Some states such as Pennsylvania have started using machine learning to predict recidivism. Understandably, there is some concern by many people, who want to know what goes “under the hood” before embracing machine learning based predictions. The excellent performance of machine learning algorithms in predicting recidivism should convince more states to take advantage of such advances. In this study, random forest model trained on SMOTEENN resampled data is the best model. Random forest models are intuitive and can easily be explained to relevant personnel.

2. The best model is generally expected to have high performance metrics such as accuracy, precision, recall, and AUC. Some models can be better at certain metrics but worse at other metrics. The purpose of the model dictates the choice of the best model. For example, for violent offenders, the justice system may want to err on safety side, preferring low False Negative Rate at the cost of incorrectly identifying non-recidivists as recidivists (i.e. higher False Positive Rate). It is recommended to start with the objective (e.g is the interest violent recidivism or general recidivism? Is it preferred to identify as many potential recidivists at the cost of classifying non-recidivists as recidivists?) and then select the best model.

**8. Conclusion**

The high recidivism rate in the US has been a source of debate about crime accountability, and rehabilitation capability of correctional facilities. Many research studies have been dedicated to identifying individuals at greater risk of recidivism and targeting them for focused rehabilitation programs. Prediction of recidivism also plays a great role in decisions related to parole, probation and sentencing. In this study, I explored the predictive capability of two common machine learning classification algorithms namely logistic regression and random forest.

The logistic regression model has good accuracy comparable to accuracy values in the literature. The model has similar performance on both the train and test data sets showing no indication of over-fitting.

The prediction results show that random forests outperformed logistic regression. The results agree with many studies (e.g. Neuilly et. al., 2011) which showed random forests performed better than logistic regression. Stalans et. al. (2004), in their study of prediction of violent recidivism, also showed that the classification accuracy of classification trees was better than logistic regression. This could be explained by the fact that random forest algorithm is appropriate for non-linear data structures which potentially involve complex higher-order interactions (Pflueger et. al., 2015)

**9. Future Work**

1. Use other algorithms (such as SVM and Neural Networks) to see if they perform better.
2. The data used in this study has relatively limited information when compared with other well-known recidivism datasets such as the 1994 Prisoner Recidivism Data released by Bureau of Justice. The 1994 dataset is restricted. I would like to use the models/procedures used here to study the 1994 Prisoner Recidivism dataset.
3. Multiclass classification (e.g. felony recidivism or violent crime recidivism) can be implemented to give detailed and offense class based recidivism. This would give a prediction of risk for specific recidivism in addition to risk for general recidivism.

**References**

Berk, Richard (2016): An impact assessment of machine learning forecasts on parole board decision and recidivism, working paper 2016-4.0, Penn Criminology Working Papers Collection.

Breiman, L. (2001): Statisical modeling: the two cultures, Statistical Science, 16(3), 199-215.

Chawla, N. V., Bowyer K. W., Hall L. O., and W. P. Kegelmeyer (2002): SMOTE: Synthetic Minority Over-sampling Technique, Journal of Artificial Intelligence Research, 321-357.

Durose, M. R., Cooper, A. D., and Snyder, H. N. (2014): Recidivism of prisoners released in 30 states in 2005: Patterns from 2005 to 2010. Washington, DC: Bureau of Justice Statistics.

Elhassan T, Aljurf M, Al-Mohanna F, and Shoukri M (2016): Classification of Imbalance Data Using Tomek Link (T-Link) Combined with Random Under-Sampling (RUS) as a Data Reduction Method, Journal of Informatics and Data Mining.

Guyon, I., and Elisseeff, A. (2003): An introduction to feature and variable selection, Journal of Machine Learning Research, 3, 1157 – 1182.

He, H., Bai, Y., Garcia, E. A., and Li, S. (2008): ADASYN: Adaptive synthetic sampling approach for imbalanced learning, IEEE international Joint Conference on Neural Networks, pp. 1322-1328.

James, N. (2016): The federal prison population buildup: options for congress ([https://fas.org/sgp/crs/misc/R42937.pdf](https://fas.org/sgp/crs/misc/R42937.pdf" \t "_blank))

Langan, P., and Levin, D. (2002): Recidivism of prisoners released in 1994, Special Report, Bureau of Justice Statistics, Washington, DC.

Neuilly. M., Zgoba, K. M., Tita, G. E., and Lee, S. S. (2011): Predicting recidivism in homicide offenders using classification tree analysis, Homicide Studies, 15(2), 154-176.

Picquero, A. R., Jennings, W. G., Diamond, B., and Reingle, J. M. (2015): A systematic review of age, sex, ethnicity, and race as predictors of violent recidivism, International Journal of Offender Therapy and Comaparatic Criminology, 59(1), 5-26.

Stahler, G. J., Mennis, J., Belenko, S., Welsh, W. N., Hiller, M. L., and Zajac, G. (2013): Predicting recidivism for released state prison offenders: examining the influence of individual and neighborhood characteristics and spatial contagion on the likelihood of reincarnation, Criminal Justice and Behavior, 40(6), 690-711.

Stalans, L. J., Yarnolds, P. R., Seng, M., Olson, D. E., and Repp, M. (2004): Identifying three types of violent offenders and predicting violent recidivism while on probation: A classification tree analysis, Law and Human Behavior, 28(3), 253-271.

Taxman, F. S., Pattavina, A., Caudy, M. S., Byrne, J., and Durso, J. (2013): The empirical basis for the RNR model with an updated RNR conceptual framework (in F. S. Taxman and A. Pattavina (eds), Simulation strategies to reduce recidivism: risk need responsivity (RNR) modeling for the criminal justice system (pp. 1 – 315), New York).

Pflueger, M. O., Franke, I., Graf, M., and Hachtel, H. (2015): Predicting general criminal recidivism in mentally disordered offenders using a random forest approach, BMC Psychiatry, 15:62.

Teich, David A. (2018): Management AI: bias, criminal recidivism, and the promise of machine learning.