Targets: Using Machine Learning Classification Models to Identify Salient Predictors of Cannabis Arrests in New York City, 2006-2018

Project: Capstone Project 1: In-depth Analysis and Machine Learning

Author: Daniel Loew

Introduction and Overview

It has been reported that there is a great racial disparity in cannabis arrests in New York City, which is a component of the larger problem that minority groups in America have long borne the greatest negative impact of the Drug War. It has been reported that 9 out of 10 cannabis arrests made in New York City were of African-Americans and Latinos during Mayor Bloomberg's controversial "stop and frisk" era of arrests which continued into the mayoralty of Bill DeBlasio, even though the Substance Abuse and Mental Health Services Administration (SAMHSA, a branch of the U.S. Department of Health and Human Services) consistently reports in their surveys that people of different racial and ethnic groups use cannabis at roughly the same rates.

In order to further examine this racial disparity , this project focuses on the NYPD's self-reported crime data between 2006 and 2018. Generally, this project created machine learning classification models with Logistic Regression and Random Forest classifiers in order to predict cannabis arrests in New York City from the pool of all crime types, as well as to predict several sub-types of cannabis arrests from the pool of all cannabis arrests in New York City.

These models were created not only to predict cannabis crime and its subtypes in New York City between 2006 and 2018, but to also identify the most salient predictors of these crimes by looking at the coefficients of the best performing Logistic Regression classifier. By identifying the most salient predictors, the biases in arrests can be scientifically identified in order to serve as a deep resource for future research in the areas of criminology and sociology. Therefore, this project does not utilize machine learning classification to test one specific hypothesis, but instead uses it to create a descriptive landscape of cannabis crime in New York City in order to highlight arrest bias in all of the available data features provided by the NYPD in their dataset. Of central importance is the fact that identifying the most salient predictors of cannabis arrests, and therefore biases in making these arrests, cannot elucidate the causes of these biases. As a logistical aside, an analysis of the coefficients that can point to the most salient predictors of the different types of cannabis crime will not be covered in this Machine Learning report, but will be saved for the final project report. Random Forest was also used as a way to nonlinearly predict the cannabis crime classes, but because the Random Forest method does not produce coefficients, it was not used to identify the most salient predictors of cannabis arrests.

The cleaned DataFrames built off of the "NYPD Complaint Data Historic" dataset (https://data.cityofnewyork.us/Public-Safety/NYPD-Complaint-Data-Historic/qgea-i56i) were used to build a series of seven classification models. The target features of the models were designed as categorical binary features that contained information on the categorical crime type for each case in the cleaned DataFrame, so classification was chosen in opposition to linear regression. To create the predictive feature set, the cleaned DataFrames utilized all features native to the NYPD's dataset in addition to a series of derived features (as detailed in the Data Cleaning notebooks). As mentioned above, the central classification method used was Logistic Regression, although Random Forest was also utilized to investigate whether a non-linear classification method could predict cannabis crime and its subtypes more accurately. Random Forest was partially chosen as it has been shown to be very effective with data sets that have a high number of both rows and features.

The seven different classification models classified:

- cannabis crimes (class 1) from all other crimes (class 0),
- cannabis possession crimes (class 1) from cannabis sales crimes (class 0),
- misdemeanor cannabis possession crimes (class 1) from all other cannabis crimes (class 0),
- violation cannabis possession crimes (class 1) from all other cannabis crimes (class 0),
- felony cannabis possession crimes (class 1) from all other cannabis crimes (class 0),
- misdemeanor sales crimes (class 1) from all other cannabis crimes (class 0), and
- felony sales crimes (class 1) from all other cannabis crimes (class 0)

The first model identifying the strongest predictors of cannabis arrests in contrast to all other crimes used the cleaned universe of 220,304 NYC cannabis crimes committed between 2006 and 2018 combined with a random sample of 220,304 non-cannabis crimes committed during the same time period. The second model identifying the strongest predictors of cannabis possession in contrast to cannabis sales used the cleaned universe of NYC cannabis crimes during this time period. The third through seventh models identifying the strongest predictors that differentiate each of the five types of cannabis crimes listed above again used the cleaned universe of cannabis crimes.

The best Logistic Regression classifier for each of the seven models was identified by evaluating the accuracy, precision, recall, and F1 scores of a set of Logistic Regression classifiers. These classifiers were run on scaled and unscaled data, and had different hyperparameter values for the solver, penalty, and regularization hyperparameter ‘C’. These classifiers and their accuracy, precision, recall, and F1 scores were stored in a hyperparameter tuning table. The best scoring Logistic Regression classifier was then identified, and further evaluated by examining the Receiver Operating Characteristic (ROC) curve and Precision Recall Curves (PRC).

The Jupyter notebooks for the seven classification models can be found at the following GitHub locations:

Cannabis and Non-Cannabis crimes:

https://github.com/danloew/SpringboardFirstCapstone/blob/master/ML_cann_v_ncann_final.ipynb

Cannabis possession and sales crimes:

https://github.com/danloew/SpringboardFirstCapstone/blob/master/ML_poss_v_sales_final.ipynb

Misdemeanor cannabis possession from all other cannabis crimes:

https://github.com/danloew/SpringboardFirstCapstone/blob/master/ML_misd_poss_final.ipynb

Violation cannabis possession from all other cannabis crimes:

https://github.com/danloew/SpringboardFirstCapstone/blob/master/ML_viol_poss_final.ipynb

Felony cannabis possession from all other cannabis crimes:

https://github.com/danloew/SpringboardFirstCapstone/blob/master/ML_felony_poss_final.ipynb

Misdemeanor cannabis sales from all other cannabis crimes:

https://github.com/danloew/SpringboardFirstCapstone/blob/master/ML_misd_sales_final.ipynb

Felony cannabis sales from all other cannabis crimes:

https://github.com/danloew/SpringboardFirstCapstone/blob/master/ML_felony_sales_final.ipynb

Initial Hypothesis Tests

As an initial step that was reported in the Statistical Analysis report and that serves as a backdrop for the coefficient analysis in the Final Project Report, hypothesis testing using t-tests uncovered the following findings:

1. The difference seen between the percentage of cannabis crimes where the suspect's race was reported and the percentage of non-cannabis crimes where the suspect's race was reported was not due to chance, and that there was some mediating factor behind this difference. The mediating factor is beyond the scope of this analysis and is an area for future research.
2. African-Americans, Whites, Hispanic Whites, Hispanic African-Americans, and Asians and Pacific Islanders arrested for a crime were not equally likely to be arrested for cannabis crimes as they were for non-cannabis crimes, and that they were more likely to be arrested for non-cannabis crimes.
3. For only those crimes where the suspect's race was reported, African-Americans arrested for a crime were equally likely to be arrested for cannabis crimes as they are for non-cannabis crimes. Again, the reason for this finding is beyond the scope of this analysis.
4. African-Americans arrested for a crime were not equally likely to be charged for cannabis crimes as White people arrested for a crime; they were more likely.
5. Hispanic Whites arrested for a crime were not equally likely to be charged for cannabis crimes as White people arrested for a crime; they were more likely (but to a lesser degree than with African-Americans).
6. Hispanic African-Americans arrested for a crime were not equally likely to be charged for cannabis crimes as White people arrested for a crime; they were more likely (but to a lesser degree than with African-Americans and Hispanic Whites).
7. Asians arrested for a crime were not equally likely to be charged for cannabis crimes as Whites arrested for a crime; they were less likely.
8. African-Americans arrested for a cannabis crime were not equally likely to be charged for misdemeanor cannabis possession as they were for violation cannabis possession; they

were more likely to be arrested for misdemeanor possession, which carries more severe legal consequences than a violation possession charge.

9. Whites arrested for a cannabis crime were equally likely to be charged for misdemeanor cannabis possession as they were for violation cannabis possession.
10. African-Americans arrested for a cannabis crime were equally likely to be arrested for violation possession as were Whites arrested for a cannabis crime. This suggests that violation possession charges were not charged differently among African-American and White suspects. It bears re-mentioning that violation possession charges were the least severe cannabis crime charges in New York City.
11. African-Americans arrested for a cannabis crime were not equally likely to be arrested for misdemeanor possession as they were for felony possession; they were more likely to be arrested for misdemeanor possession.
12. African-Americans arrested for a cannabis crime were not equally likely to be arrested for misdemeanor sales as they were for felony sales; they were more likely to be arrested for misdemeanor sales.
13. Cannabis arrests were not equally as likely to happen in the five boroughs. The Bronx was the most likely, Brooklyn was the second most likely, Manhattan was the third, Queens was the fourth, and Staten Island was the fifth.

The Jupyter notebook running these hypothesis tests can be found here:

https://github.com/danloew/SpringboardFirstCapstone/blob/master/DataStory_HT_Final.ipynb

Machine Learning Pipelines

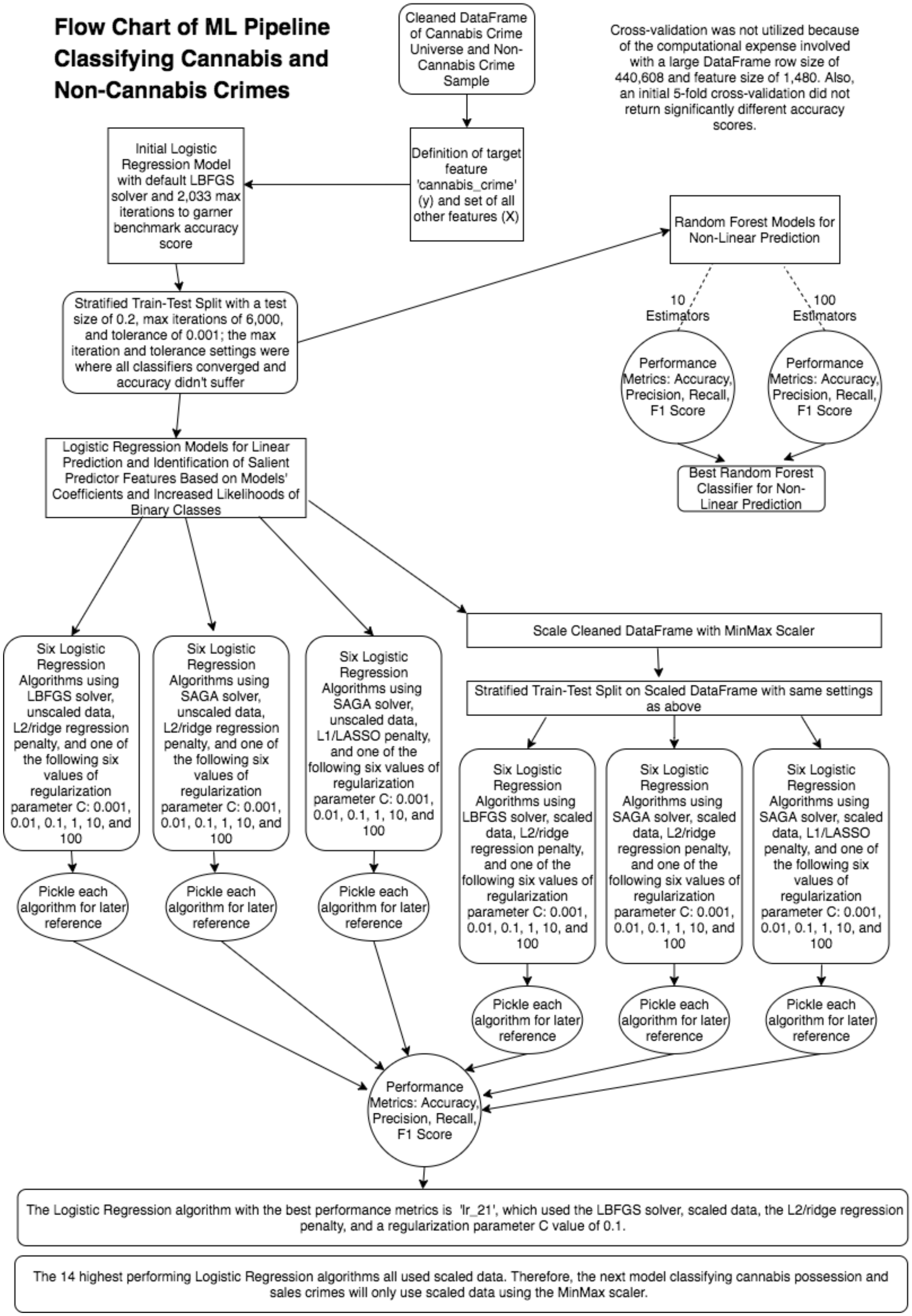
The machine learning classification processes for the seven models is described and evaluated in this section. The coefficient analysis of these models is carried out in the Final Report.

Creation and Evaluation of Models Classifying Cannabis Crime and Non-Cannabis Crime

The machine learning classification pipeline of predicting cannabis crime as differentiated from non-cannabis crime followed a logical series of steps, as shown in the following visualization. This visualization helps to summarize the pipeline, and is offered in lieu of a narrative explanation.

Flow Chart of ML Pipeline Classifying Cannabis and Non-Cannabis Crimes

Cross-validation was not utilized because of the computational expense involved with a large DataFrame row size of 440,608 and feature size of 1,480. Also, an initial 5-fold cross-validation did not return significantly different accuracy scores.


Cleaned DataFrame of Cannabis Crime Universe and Non-Cannabis Crime Sample
Definition of target feature 'cannabis_crime' (y) and set of all other features (X)
Initial Logistic Regression Model with default LBFGS solver and 2,033 max iterations to garner benchmark accuracy score
Stratified Train-Test Split with a test size of 0.2, max iterations of 6,000, and tolerance of 0.001; the max iteration and tolerance settings were where all classifiers converged and accuracy didn't suffer
Random Forest Models for Non-Linear Prediction
10 Estimators
100 Estimators
Performance Metrics: Accuracy, Precision, Recall, F1 Score
Performance Metrics: Accuracy, Precision, Recall, F1 Score
Best Random Forest Classifier for Non-Linear Prediction
Logistic Regression Models for Linear Prediction and Identification of Salient Predictor Features Based on Models' Coefficients and Increased Likelihoods of Binary Classes
Six Logistic Regression Algorithms using LBFGS solver, unscaled data, L2/ridge regression penalty, and one of the following six values of regularization parameter C: 0.001, 0.01, 0.1, 1, 10, and 100
Six Logistic Regression Algorithms using SAGA solver, unscaled data, L2/ridge regression penalty, and one of the following six values of regularization parameter C: 0.001, 0.01, 0.1, 1, 10, and 100
Six Logistic Regression Algorithms using SAGA solver, unscaled data, L1/LASSO penalty, and one of the following six values of regularization parameter C: 0.001, 0.01, 0.1, 1, 10, and 100
Scale Cleaned DataFrame with MinMax Scaler
Stratified Train-Test Split on Scaled DataFrame with same settings as above
Six Logistic Regression Algorithms using LBFGS solver, scaled data, L2/ridge regression penalty, and one of the following six values of regularization parameter C: 0.001, 0.01, 0.1, 1, 10, and 100
Six Logistic Regression Algorithms using SAGA solver, scaled data, L2/ridge regression penalty, and one of the following six values of regularization parameter C: 0.001, 0.01, 0.1, 1, 10, and 100
Six Logistic Regression Algorithms using SAGA solver, scaled data, L1/LASSO penalty, and one of the following six values of regularization parameter C: 0.001, 0.01, 0.1, 1, 10, and 100
Pickle each algorithm for later reference
Pickle each algorithm for later reference
Pickle each algorithm for later reference
Pickle each algorithm for later reference
Pickle each algorithm for later reference
Pickle each algorithm for later reference
Performance Metrics: Accuracy, Precision, Recall, F1 Score

The Logistic Regression algorithm with the best performance metrics is 'lr_21', which used the LBFGS solver, scaled data, the L2/ridge regression penalty, and a regularization parameter C value of 0.1.

The 14 highest performing Logistic Regression algorithms all used scaled data. Therefore, the next model classifying cannabis possession and sales crimes will only use scaled data using the MinMax scaler.

Identification of Best Performing Logistic Regression and Random Forest Algorithms By Evaluating Accuracy, Precision, Recall, and F1 Scores

The best performing Logistic Regression model for the classification of cannabis crimes and non-cannabis crimes ('lr_21') had an accuracy score of 0.843, showing that it made correct predictions on roughly 84.3% of the data points in the DataFrame.

The model's precision was 0.82 for the cannabis crime class (the 1 class), 0.88 for the non-cannabis crime class (the 0 class), and 0.84 on average, showing that 82% of predicted cannabis crimes were actual cannabis crimes, 88% of predicted non-cannabis crimes were actual non-cannabis crimes, and 84% of crimes on weighted average were predicted correctly.

The model's recall is 0.89 for the cannabis crime class (the 1 class), 0.80 for the non-cannabis crime class (the 0 class), and 0.84 on average, showing that 89% of actual cannabis crimes were predicted cannabis crimes, 80% of actual non-cannabis crimes were predicted non-cannabis crimes, and again that 84% of crimes on weighted average were predicted correctly.

The model's F1 score, or harmonic mean of precision and recall, was 0.85 for the cannabis crime class, 0.84 for the non-cannabis crime class, and 0.84 on average. This metric is more informative than precision or recall alone, and shows that 85% of cannabis crimes and 84% of non-cannabis crimes were predicted correctly, for an average of 84%.

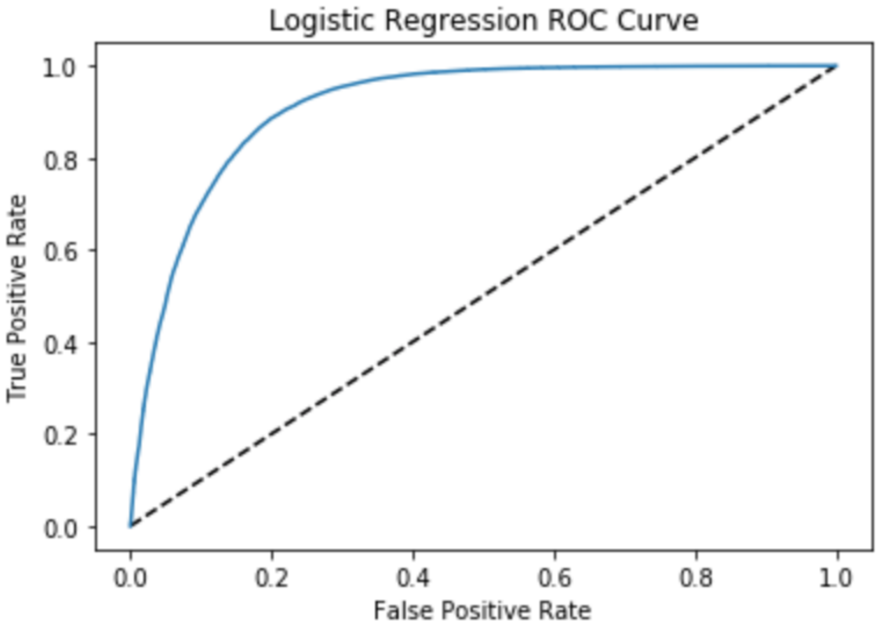
The RandomForest model with 100 estimators had an accuracy of 0.856. It had a 0.85 precision score for the cannabis crime class, a 0.87 precision score for the non-cannabis crime class, and an average precision score of 0.86. It had a recall score of 0.87 for the cannabis crime class, a 0.84 recall score for the non-cannabis crime class, and an average recall score of 0.86. It has a 0.86 F1 score for the cannabis crime class, a 0.85 F1 score for the non-cannabis crime class, and an average F1 score of 0.86.

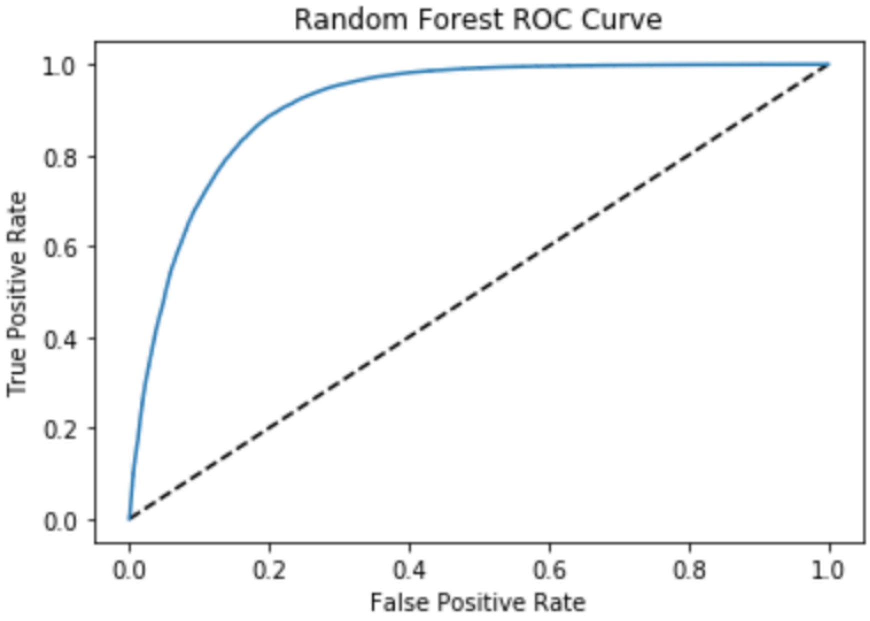
Further Evaluation of Best Performing Logistic Regression and Random Forest Algorithms via their Receiver Operating Characteristic (ROC) Curves

The ROC curves and Areas Under the Curve (AUCs) for the Random Forest classifier with 100 estimators and the best Logistic Regression classifier ('lr_21') were created and calculated to further evaluate these algorithms.

ROC curves help to compare models that predict probabilities for two-class problems. They use the predicted probabilities of each crime as being a cannabis crime (the 1 class) or a non-cannabis crime (the 0 class), while calibrating the threshold of how to interpret these predicted probabilities as belonging to the 1 class or the 0 class, while also reducing false positive or false negative errors in prediction of these classes.

As shown below, ROC curves plot the false positive rate on the X-axis and the true positive rate on the Y-axis for a series of candidate threshold values between 0 and 1, so small values on the X-axis indicate lower false positives and higher true negatives, while larger values on the y-axis indicate higher true positives and lower false negatives. In this case, the area under the curve (AUC) summarizes the skill of the model in predicting cannabis crimes or non-cannabis crimes. A skillful model assigns a higher probability to a randomly chosen real positive occurrence than a negative occurrence on average. This explanation of ROC curves and their AUC is provided for clarification purposes, and is a distillation of Dr. Jason Brownlee's article on the subject at https://machinelearningmastery.com/roc-curves-and-precision-recall-curves-for-classification-in-python/ .


Logistic Regression ROC Curve


Random Forest ROC Curve

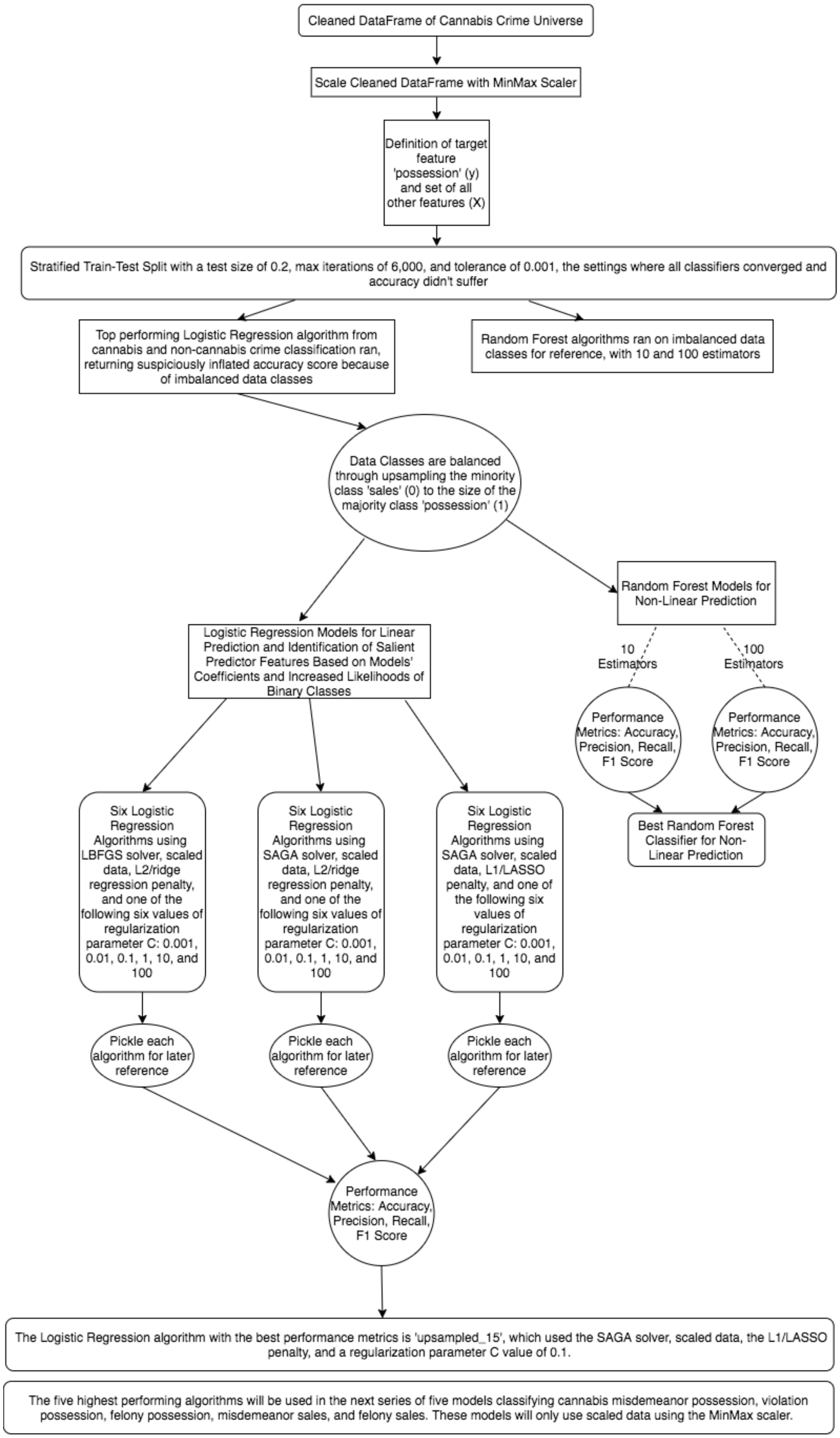
The ROC curve plotted for the best performing LogisticRegression algorithm ('lr_21') has an AUC score of 91.3%, showing that it is a very skillful model in predicting cannabis crimes at a much higher rate than random. Using this model's coefficients can reliably show the features in the NYPD's dataset that have the strongest statistical relationship with cannabis crimes as differentiated from all other crimes for the time period of 2006-2018.

The ROC curve plotted for the best performing Random Forest algorithm has an AUC score of 92.4%. Therefore, this is a very skillful model and one that is slightly more skillful than the best LogisticRegression model detailed above. It can be used to predict future crimes as being cannabis related or not, but because of its non-linear nature cannot uncover the features in the NYPD's dataset that have the strongest relationship with cannabis crimes.

Creation and Evaluation of Models Classifying Cannabis Possession and Sales Crimes

The machine learning classification pipeline of predicting cannabis possession and sales crimes followed a logical series of steps, as shown in the visualization on the next page.

Flow Chart of ML Pipeline Classifying Cannabis Possession & Sales Crimes


Cleaned DataFrame of Cannabis Crime Universe
Scale Cleaned DataFrame with MinMax Scaler
Definition of target feature 'possession' (y) and set of all other features (X)
Stratified Train-Test Split with a test size of 0.2, max iterations of 6,000, and tolerance of 0.001, the settings where all classifiers converged and accuracy didn't suffer
Top performing Logistic Regression algorithm from cannabis and non-cannabis crime classification ran, returning suspiciously inflated accuracy score because of imbalanced data classes
Random Forest algorithms ran on imbalanced data classes for reference, with 10 and 100 estimators
Data Classes are balanced through upsampling the minority class 'sales' (0) to the size of the majority class 'possession' (1)
Random Forest Models for Non-Linear Prediction
10 Estimators
100 Estimators
Performance Metrics: Accuracy, Precision, Recall, F1 Score
Performance Metrics: Accuracy, Precision, Recall, F1 Score
Best Random Forest Classifier for Non-Linear Prediction
Logistic Regression Models for Linear Prediction and Identification of Salient Predictor Features Based on Models' Coefficients and Increased Likelihoods of Binary Classes
Six Logistic Regression Algorithms using LBFGS solver, scaled data, L2/ridge regression penalty, and one of the following six values of regularization parameter C: 0.001, 0.01, 0.1, 1, 10, and 100
Six Logistic Regression Algorithms using SAGA solver, scaled data, L2/ridge regression penalty, and one of the following six values of regularization parameter C: 0.001, 0.01, 0.1, 1, 10, and 100
Six Logistic Regression Algorithms using SAGA solver, scaled data, L1/LASSO penalty, and one of the following six values of regularization parameter C: 0.001, 0.01, 0.1, 1, 10, and 100
Pickle each algorithm for later reference
Pickle each algorithm for later reference
Pickle each algorithm for later reference
Performance Metrics: Accuracy, Precision, Recall, F1 Score
The Logistic Regression algorithm with the best performance metrics is 'upsampled_15', which used the SAGA solver, scaled data, the L1/LASSO penalty, and a regularization parameter C value of 0.1.
The five highest performing algorithms will be used in the next series of five models classifying cannabis misdemeanor possession, violation possession, felony possession, misdemeanor sales, and felony sales. These models will only use scaled data using the MinMax scaler.

Identification of Best Performing Logistic Regression and Random Forest Algorithms By Using Accuracy, Precision, Recall, and F1 Scores

The best performing Logistic Regression model for the classification of cannabis crimes and non-cannabis crimes ('upsampled_15') had an accuracy score of 0.682, showing that it made correct predictions on roughly 68.2% of the data points in the DataFrame. This is quite a bit lower than the .843 accuracy of the model classifying cannabis crimes from non-cannabis crimes, suggesting that the model's coefficients may not be as illustrative of the true relationship between the feature set and the target classes of cannabis possession and sales.

The model's precision was 0.97 for the cannabis possession class (the 1 class), 0.11 for the cannabis sales class (the 0 class), and 0.92 on weighted average, showing that 97% of predicted cannabis possession crimes were actual cannabis possession crimes, 11% of predicted cannabis sales crimes were actual cannabis sales crimes, and 92% of crimes on weighted average were predicted correctly.

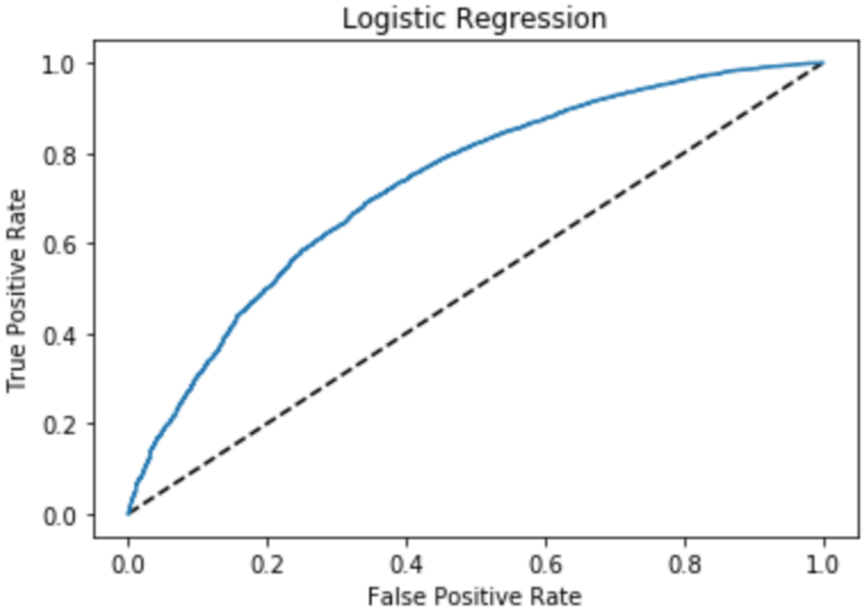
The model's recall was 0.68 for the cannabis possession class (the 1 class), 0.66 for the cannabis sales class (0), and 0.68 on weighted average, showing that 68% of actual cannabis possession crimes were predicted as cannabis possession crimes, 66% of actual cannabis sales crimes were predicted as cannabis sales crimes, and that 68% of crimes on weighted average were predicted correctly.

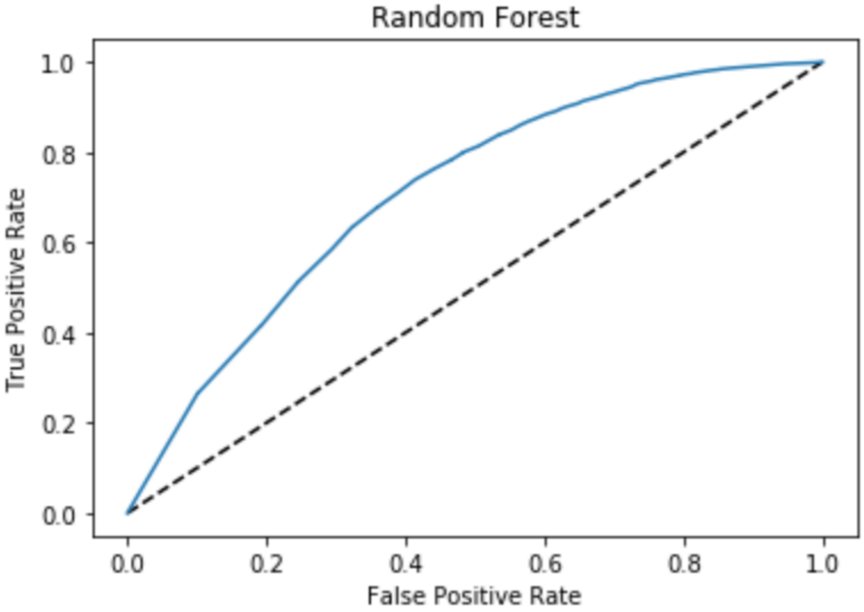
The model's F1 score, or harmonic mean of precision and recall, was 0.80 for the cannabis possession class, 0.19 for the cannabis sales class, and 0.77 on weighted average. This metric was more informative than precision or recall alone, and shows that 80% of cannabis possession crimes and 19% of cannabis sales crimes were being predicted correctly, for a weighted average of 77%.

The RandomForest model with 100 estimators has an accuracy of 0.931. It has a 0.95 precision score for the cannabis possession class, a 0.32 precision score for the cannabis sales class, and a weighted average precision score of 0.92. It has a recall score of 0.98 for the cannabis possession class, a 0.19 recall score for the cannabis sales class, and a weighted average recall score of 0.93. It has a 0.96 F1 score for the cannabis possession class, a 0.23 F1 score for the cannabis sales class, and a weighted average F1 score of 0.92.

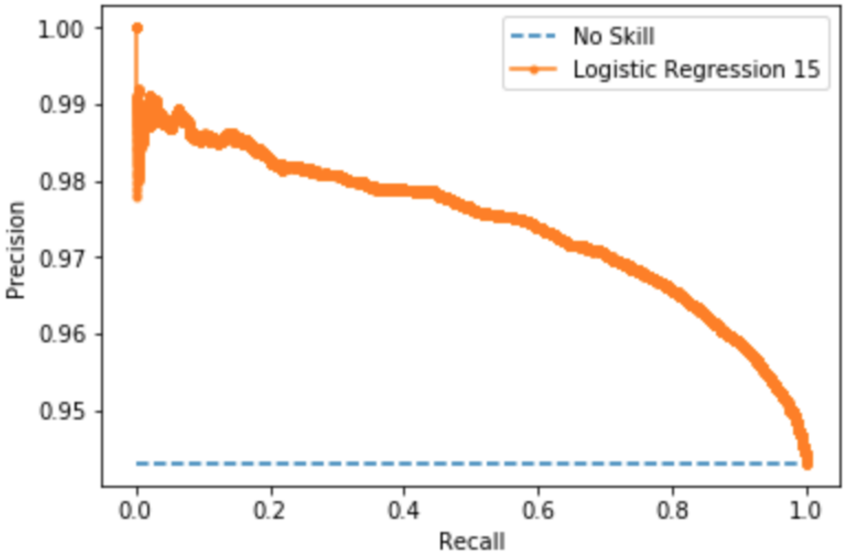
Further Evaluation of Best Performing Logistic Regression and Random Forest Algorithms via their Receiver Operating Characteristic (ROC) and Precision-Recall Curves (PRC)

The ROC curves and Areas Under the Curve (AUCs) for the best Logistic Regression algorithm ('upsampled_15') and the Random Forest algorithm with 100 estimators, as well as the Precision-Recall curve for the Logistic Regression algorithm, were created and calculated for further evaluation.


Logistic Regression


Random Forest

Because of the imbalanced class count, the 15th LogisticRegression algorithm is evaluated below with a Precision-Recall Curve, as recommended by this article by Jason Brownlee, PhD: https://machinelearningmastery.com/roc-curves-and-precision-recall-curves-for-classification-in-python/. As Dr. Brownlee says, the future performance of a model with imbalanced class counts is better evaluated with a Precision-Recall Curve than an ROC curve.



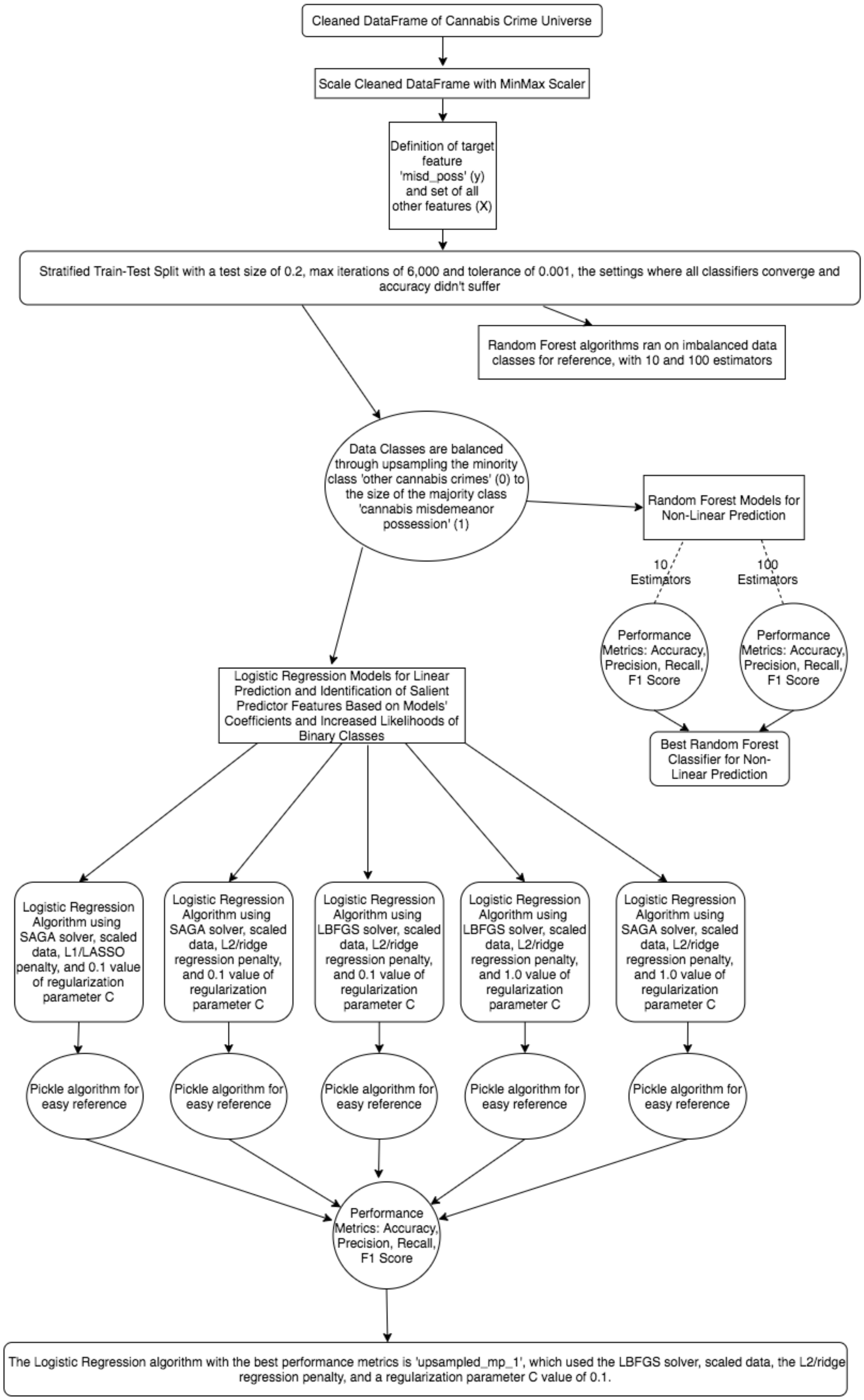
The ROC curves plotted above show that the best LogisticRegression model ('upsampled_15') has an ROC AUC score of 72.9%, showing that it is a moderately skillful model in predicting cannabis possession crimes at a rate higher than random. The Precision Recall Curve has an AUC score of 97.4%. The higher Precision Recall AUC score shows that the model is skilled at differentiating true positives from false positives and false negatives, especially for the cannabis possession class. However, as is shown above, it is not as skillful at predicting cannabis sales crimes.

The ROC curve plotted above shows that the Random Forest algorithm with 100 estimators has an AUC score of 71.2%. Therefore this is not a very skillful model, and because of its non-linear nature cannot uncover the features in the NYPD's dataset that have the strongest relationship with cannabis possession and sales crimes.

Creation and Evaluation of Model Classifying Misdemeanor Cannabis Possession Crimes

The machine learning classification pipeline of predicting misdemeanor possession crimes followed a logical series of steps, as shown in the visualization on the following page.

Flow Chart of ML Pipeline Classifying Misdemeanor Possession Crimes


Cleaned DataFrame of Cannabis Crime Universe
Scale Cleaned DataFrame with MinMax Scaler
Definition of target feature 'misd_poss' (y) and set of all other features (X)
Stratified Train-Test Split with a test size of 0.2, max iterations of 6,000 and tolerance of 0.001, the settings where all classifiers converge and accuracy didn't suffer
Random Forest algorithms ran on imbalanced data classes for reference, with 10 and 100 estimators
Data Classes are balanced through upsampling the minority class 'other cannabis crimes' (0) to the size of the majority class 'cannabis misdemeanor possession' (1)
Random Forest Models for Non-Linear Prediction
10 Estimators
100 Estimators
Performance Metrics: Accuracy, Precision, Recall, F1 Score
Performance Metrics: Accuracy, Precision, Recall, F1 Score
Best Random Forest Classifier for Non-Linear Prediction
Logistic Regression Models for Linear Prediction and Identification of Salient Predictor Features Based on Models' Coefficients and Increased Likelihoods of Binary Classes
Logistic Regression Algorithm using SAGA solver, scaled data, L1/LASSO penalty, and 0.1 value of regularization parameter C
Logistic Regression Algorithm using SAGA solver, scaled data, L2/ridge regression penalty, and 0.1 value of regularization parameter C
Logistic Regression Algorithm using LBFGS solver, scaled data, L2/ridge regression penalty, and 0.1 value of regularization parameter C
Logistic Regression Algorithm using LBFGS solver, scaled data, L2/ridge regression penalty, and 1.0 value of regularization parameter C
Logistic Regression Algorithm using SAGA solver, scaled data, L2/ridge regression penalty, and 1.0 value of regularization parameter C
Pickle algorithm for easy reference
Pickle algorithm for easy reference
Pickle algorithm for easy reference
Pickle algorithm for easy reference
Pickle algorithm for easy reference
Performance Metrics: Accuracy, Precision, Recall, F1 Score
The Logistic Regression algorithm with the best performance metrics is 'upsampled_mp_1', which used the LBFGS solver, scaled data, the L2/ridge regression penalty, and a regularization parameter C value of 0.1.

Identification of Best Performing Logistic Regression and Random Forest Algorithms By Using Accuracy, Precision, Recall, and F1 Scores

The best Logistic Regression model ('upsampled_mp_1') has an accuracy of 0.716, showing that it makes correct predictions on roughly 71.6% of the data points in the DataFrame.

The model's precision was 0.94 for the misdemeanor possession class (the 1 class), 0.20 for the other cannabis crime types class (the 0 class), and 0.86 on weighted average, showing that 94% of predicted misdemeanor possession crimes are actual misdemeanor possession crimes, 20% of predicted other cannabis crimes are actual other cannabis crimes, and 86% of cannabis crimes on average are predicted correctly.

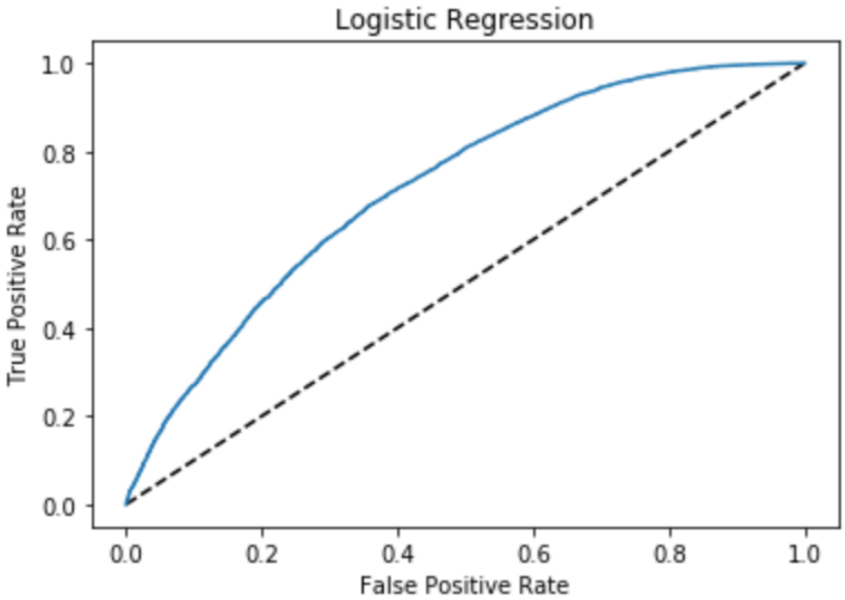
The model's recall was 0.73 for the misdemeanor possession class (the 1 class), 0.58 for the other cannabis crime types class (0), and 0.72 on weighted average, showing that 73% of actual misdemeanor possession crimes are predicted as misdemeanor possession crimes, 58% of actual other cannabis crimes are predicted as other cannabis crimes, and that 72% of cannabis crimes on average are predicted correctly.

The model's F1 score, or harmonic mean of precision and recall, was 0.82 for the misdemeanor possession class, 0.30 for the other cannabis crimes class, and 0.77 on weighted average. This metric is more informative than precision or recall alone, and shows that 82% of misdemeanor possession crimes and 30% of other cannabis crimes are being predicted correctly, for a weighted average of 77%.

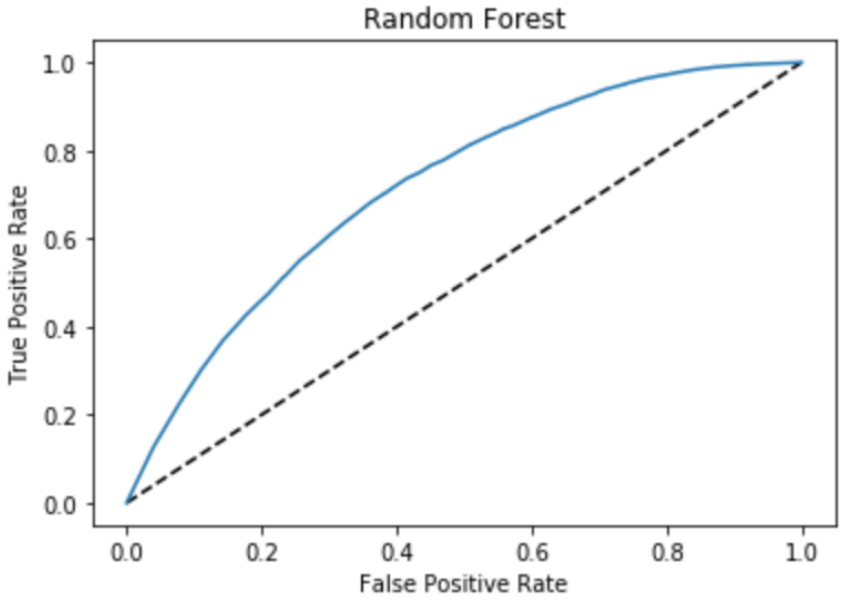
The RandomForest model with 100 estimators has an accuracy of 0.88. It has a 0.92 precision score for the misdemeanor possession class, a 0.39 precision score for the other crime types class, and a weighted average precision score of 0.86. The model has a recall score of 0.95 for the misdemeanor possession class, a 0.26 recall score for the other crime types class, and a weighted average recall score of 0.88. It has a 0.93 F1 score for the misdemeanor possession class, a 0.31 F1 score for the other crime types class, and a weighted average F1 score of 0.87.

Further Evaluation of Best Performing Logistic Regression and Random Forest Algorithms via their Receiver Operating Characteristic (ROC) and Precision-Recall Curves

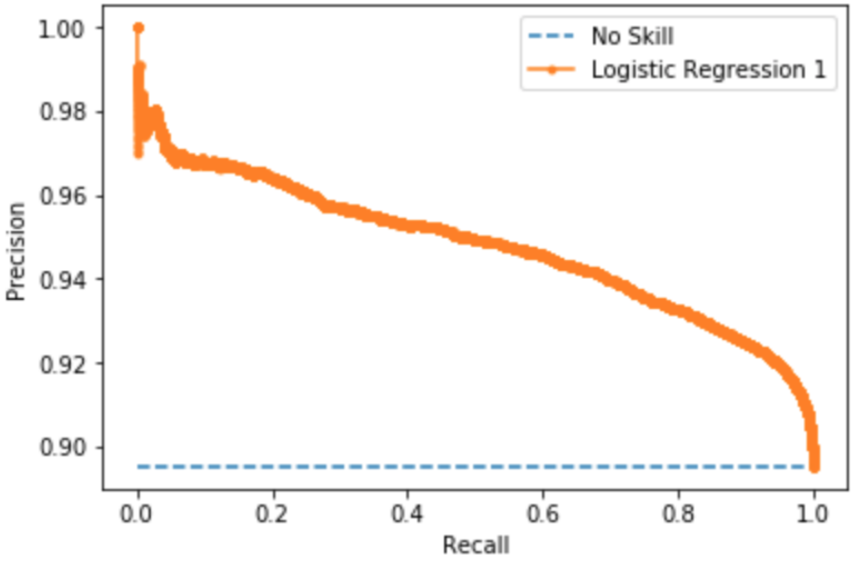
The ROC curves and Areas Under the Curve (AUCs) for the best Logistic Regression algorithm ('upsampled_mp_1’) and the Random Forest algorithm with 100 estimators, as well as the Precision-Recall curve for the Logistic Regression algorithm, were created and calculated for further evaluation.



Logistic Regression



Random Forest



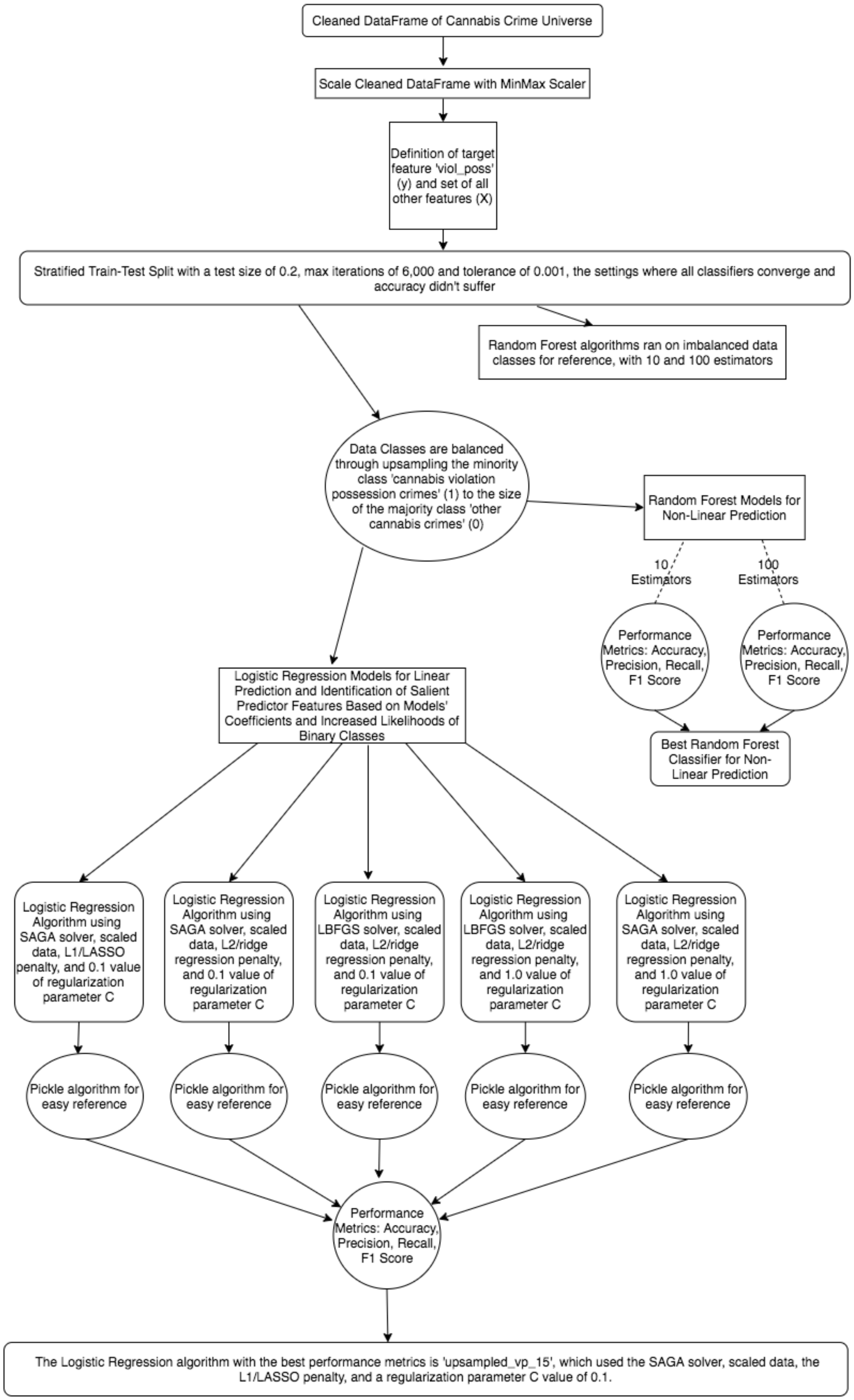
The ROC curves plotted above show that the best LogisticRegression model ('upsampled_mp_1') has an ROC AUC score of 71.9%, showing that it is a moderately skillful model in predicting misdemeanor possession crimes at a rate higher than random. The Precision Recall Curve has an AUC score of 94.8%. The higher Precision Recall AUC score shows that the model is skilled at differentiating true positives from false positives and false negatives, especially for the cannabis misdemeanor possession class. However, as is shown above, it is not as skillful at predicting the other cannabis crime types as a group.

The ROC curve for the Random Forest algorithm with 100 estimators plotted above shows that it has an AUC score of 71.8%. This is a skillful model at predicting misdemeanor possession crimes, but because of its non-linear nature cannot uncover the features in the NYPD's dataset that have the strongest relationship with misdemeanor possession crimes.

Creation and Evaluation of Model Classifying Violation Cannabis Possession Crimes

The machine learning classification pipeline of predicting violation possession crimes followed a logical series of steps, as shown in the visualization on the following page.

Flow Chart of ML Pipeline Classifying Violation Possession Crimes


Cleaned DataFrame of Cannabis Crime Universe
Scale Cleaned DataFrame with MinMax Scaler
Definition of target feature 'viol_poss' (y) and set of all other features (X)
Stratified Train-Test Split with a test size of 0.2, max iterations of 6,000 and tolerance of 0.001, the settings where all classifiers converge and accuracy didn't suffer
Random Forest algorithms ran on imbalanced data classes for reference, with 10 and 100 estimators
Data Classes are balanced through upsampling the minority class 'cannabis violation possession crimes' (1) to the size of the majority class 'other cannabis crimes' (0)
Random Forest Models for Non-Linear Prediction
10 Estimators
100 Estimators
Performance Metrics: Accuracy, Precision, Recall, F1 Score
Performance Metrics: Accuracy, Precision, Recall, F1 Score
Best Random Forest Classifier for Non-Linear Prediction
Logistic Regression Models for Linear Prediction and Identification of Salient Predictor Features Based on Models' Coefficients and Increased Likelihoods of Binary Classes
Logistic Regression Algorithm using SAGA solver, scaled data, L1/LASSO penalty, and 0.1 value of regularization parameter C
Logistic Regression Algorithm using SAGA solver, scaled data, L2/ridge regression penalty, and 0.1 value of regularization parameter C
Logistic Regression Algorithm using LBFGS solver, scaled data, L2/ridge regression penalty, and 0.1 value of regularization parameter C
Logistic Regression Algorithm using LBFGS solver, scaled data, L2/ridge regression penalty, and 1.0 value of regularization parameter C
Logistic Regression Algorithm using SAGA solver, scaled data, L2/ridge regression penalty, and 1.0 value of regularization parameter C
Pickle algorithm for easy reference
Pickle algorithm for easy reference
Pickle algorithm for easy reference
Pickle algorithm for easy reference
Pickle algorithm for easy reference
Performance Metrics: Accuracy, Precision, Recall, F1 Score
The Logistic Regression algorithm with the best performance metrics is 'upsampled_vp_15', which used the SAGA solver, scaled data, the L1/LASSO penalty, and a regularization parameter C value of 0.1.

Identification of Best Performing Logistic Regression and Random Forest Algorithms By Using Accuracy, Precision, Recall, and F1 Scores

The best Logistic Regression model ('upsampled_vp_15') has an accuracy of 0.813, showing that it makes correct predictions on roughly 81.3% of the data points in the DataFrame.

The model's precision was 0.10 for the violation possession class (the 1 class), 0.99 for the other cannabis crime types class (the 0 class), and 0.96 on weighted average, showing that 10% of predicted violation possession crimes were actual violation possession crimes, 99% of predicted other cannabis crimes were actual other cannabis crimes, and 96% of cannabis crimes on weighted average were predicted correctly.

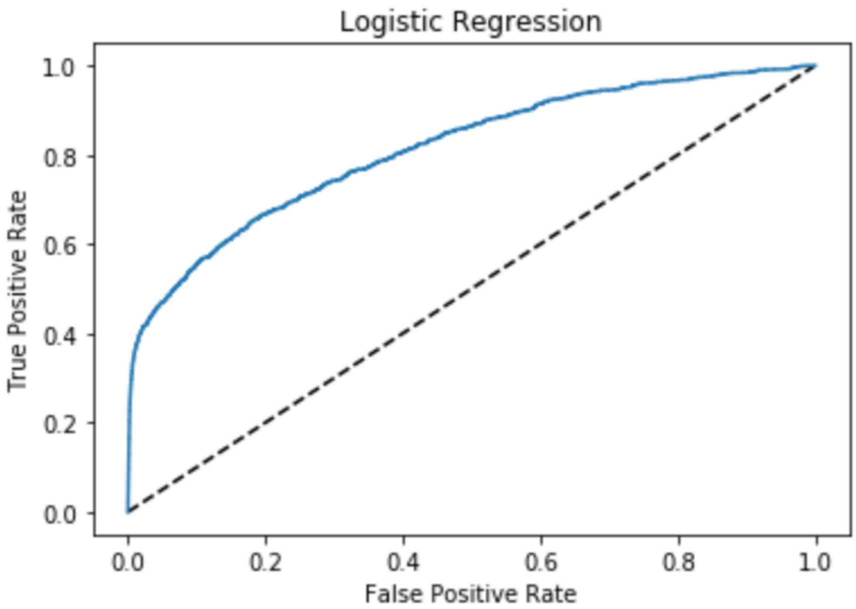
The model's recall was 0.65 for the violation possession class (the 1 class), 0.82 for the other cannabis crime types class (0), and 0.81 on weighted average, showing that 65% of actual violation possession crimes were predicted as violation possession crimes, 82% of actual other cannabis crimes were predicted as other cannabis crimes, and that 81% of cannabis crimes on weighted average were predicted correctly.

The model's F1 score, or harmonic mean of precision and recall, was 0.18 for the violation possession class, 0.89 for the other cannabis crimes class, and 0.87 on weighted average. This metric is more informative than precision or recall alone, and shows that 18% of violation possession crimes and 89% of other cannabis crimes are being predicted correctly, for a weighted average of 87%.

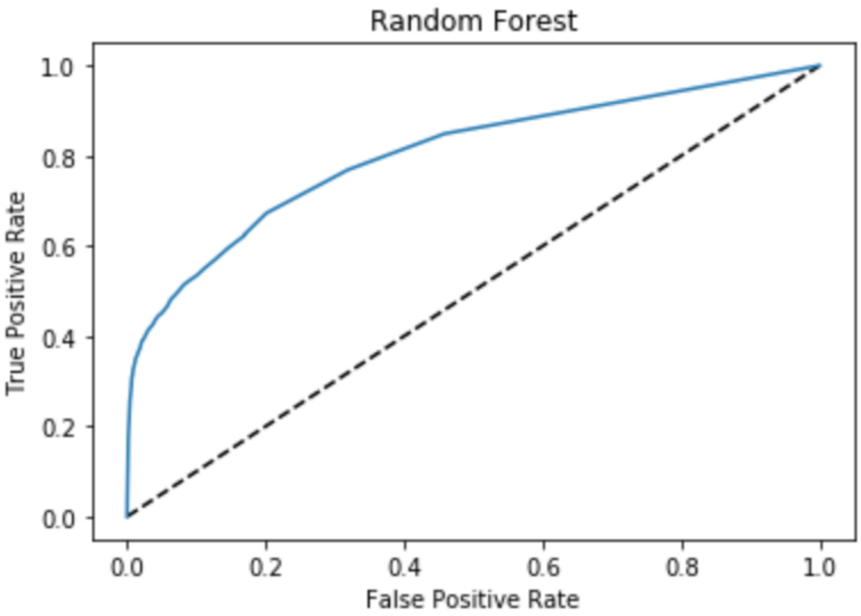
The RandomForest model with 100 estimators has an accuracy of 0.971. It has a 0.55 precision score for the violation possession class, a 0.98 precision score for the other crime types class, and a weighted average precision score of 0.97. The model has a recall score of 0.32 for the violation possession class, a 0.99 recall score for the other crime types class, and a weighted average recall score of 0.97. It has a 0.41 F1 score for the violation possession class, a 0.99 F1 score for the other crime types class, and a weighted average F1 score of 0.97.

Further Evaluation of Best Performing Logistic Regression and Random Forest Algorithms via their Receiver Operating Characteristic (ROC) and Precision-Recall Curves

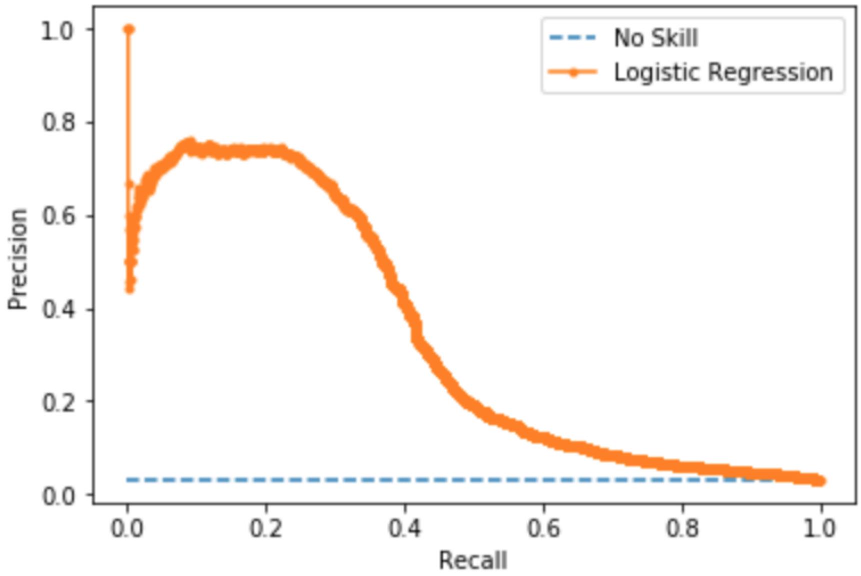
The ROC curves and Areas Under the Curve (AUCs) for the best Logistic Regression algorithm ('upsampled_vp_15') and the Random Forest algorithm with 100 estimators, as well as the Precision-Recall curve for the Logistic Regression algorithm, were created and calculated for further evaluation.



Logistic Regression



Random Forest



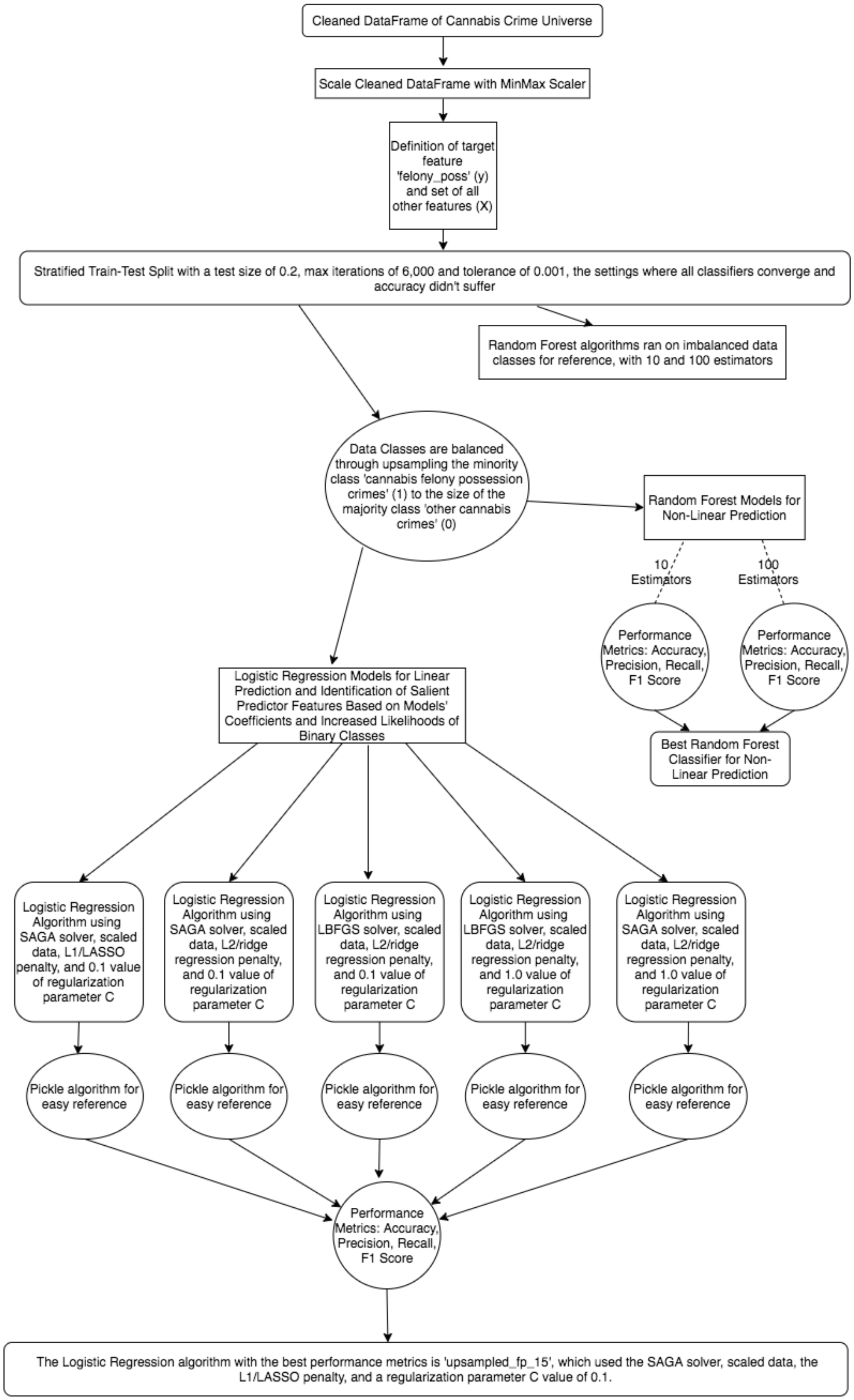
The ROC curves plotted above show that the best LogisticRegression model ('upsampled_vp_15') has an ROC AUC score of 81.3%, showing that it is a moderately skillful model in predicting violation possession crimes at a rate higher than random. The Precision Recall Curve has an AUC score of 33.8%. The lower Precision Recall AUC score shows that the model is not very skilled at differentiating true positives from false positives and false negatives, especially for the violation possession class. However, as is shown above, it is more skillful at predicting the other cannabis crime types as a group.

The ROC curve for the Random Forest algorithm with 100 estimators plotted above shows that it has an AUC score of 80.3%. This is not a very skillful model at predicting violation possession crimes.

Creation and Evaluation of Model Classifying Felony Cannabis Possession Crimes

The machine learning classification pipeline of predicting felony possession crimes followed a logical series of steps, as shown in the visualization on the following page.

Flow Chart of ML Pipeline Classifying Felony Possession Crimes


Cleaned DataFrame of Cannabis Crime Universe
Scale Cleaned DataFrame with MinMax Scaler
Definition of target feature 'felony_poss' (y) and set of all other features (X)
Stratified Train-Test Split with a test size of 0.2, max iterations of 6,000 and tolerance of 0.001, the settings where all classifiers converge and accuracy didn't suffer
Random Forest algorithms ran on imbalanced data classes for reference, with 10 and 100 estimators
Data Classes are balanced through upsampling the minority class 'cannabis felony possession crimes' (1) to the size of the majority class 'other cannabis crimes' (0)
Random Forest Models for Non-Linear Prediction
10 Estimators
100 Estimators
Performance Metrics: Accuracy, Precision, Recall, F1 Score
Performance Metrics: Accuracy, Precision, Recall, F1 Score
Best Random Forest Classifier for Non-Linear Prediction
Logistic Regression Models for Linear Prediction and Identification of Salient Predictor Features Based on Models' Coefficients and Increased Likelihoods of Binary Classes
Logistic Regression Algorithm using SAGA solver, scaled data, L1/LASSO penalty, and 0.1 value of regularization parameter C
Logistic Regression Algorithm using SAGA solver, scaled data, L2/ridge regression penalty, and 0.1 value of regularization parameter C
Logistic Regression Algorithm using LBFGS solver, scaled data, L2/ridge regression penalty, and 0.1 value of regularization parameter C
Logistic Regression Algorithm using LBFGS solver, scaled data, L2/ridge regression penalty, and 1.0 value of regularization parameter C
Logistic Regression Algorithm using SAGA solver, scaled data, L2/ridge regression penalty, and 1.0 value of regularization parameter C
Pickle algorithm for easy reference
Pickle algorithm for easy reference
Pickle algorithm for easy reference
Pickle algorithm for easy reference
Pickle algorithm for easy reference
Performance Metrics: Accuracy, Precision, Recall, F1 Score
The Logistic Regression algorithm with the best performance metrics is 'upsampled_fp_15', which used the SAGA solver, scaled data, the L1/LASSO penalty, and a regularization parameter C value of 0.1.

Identification of Best Performing Logistic Regression and Random Forest Algorithms By Using Accuracy, Precision, Recall, and F1 Scores

The best Logistic Regression model ('upsampled_fp_15') has an accuracy of 0.736, showing that it makes correct predictions on roughly 73.6% of the data points in the DataFrame.

The model's precision was 0.04 for the felony possession class (the 1 class), 0.99 for the other cannabis crime types class (the 0 class), and 0.98 on weighted average, showing that 4% of predicted felony possession crimes are actual felony possession crimes, 99% of predicted other cannabis crimes are actual other cannabis crimes, and 98% of crimes on weighted average are predicted correctly.

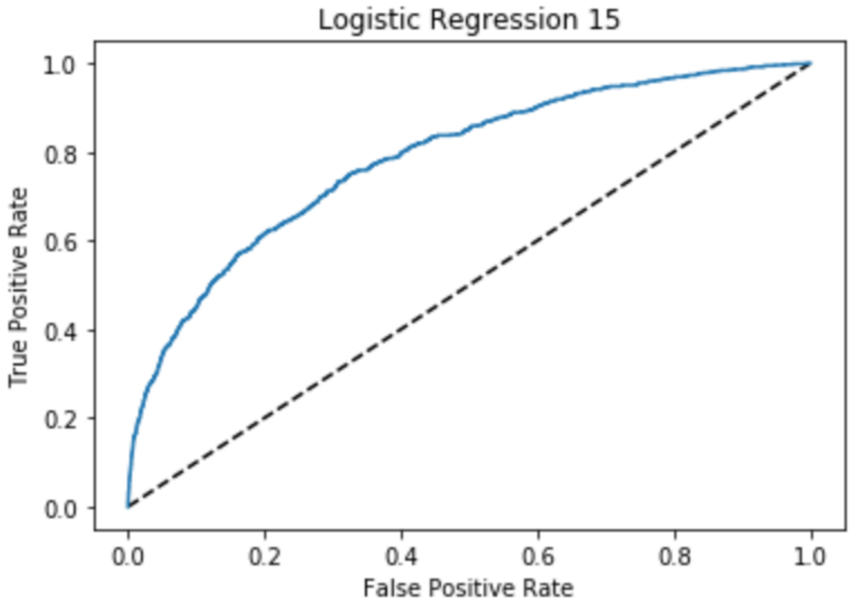
The model's recall was 0.67 for the felony possession class (the 1 class), 0.74 for the other cannabis crime types class (0), and 0.74 on weighted average, showing that 67% of actual felony possession crimes were predicted as felony possession crimes, 74% of actual other cannabis crimes were predicted as other cannabis crimes, and that 74% of crimes on weighted average were predicted correctly.

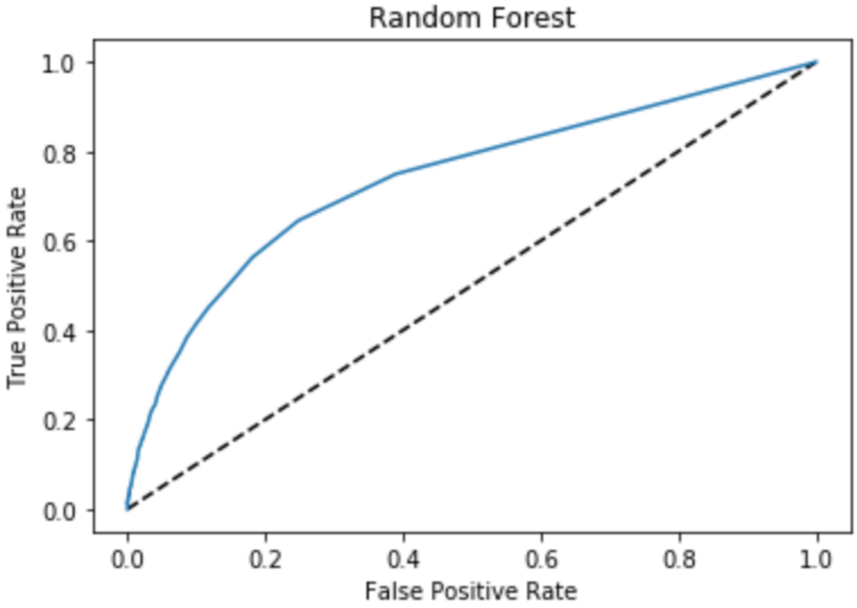
The model's F1 score, or harmonic mean of precision and recall, was 0.08 for the felony possession class, 0.85 for the other cannabis crimes class, and 0.83 on weighted average. This metric is more informative than precision or recall alone, and shows that only 8% of felony possession crimes and 85% of other cannabis crimes were being predicted correctly, for a weighted average of 83%.

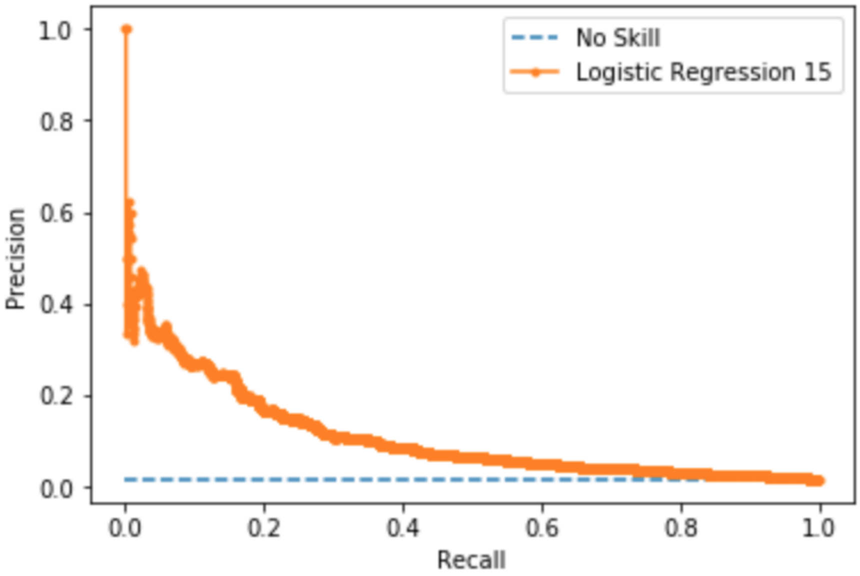
The RandomForest model with 100 estimators has an accuracy of 0.980. It has a 0.19 precision score for the felony possession class, a 0.98 precision score for the other crime types class, and a weighted average precision score of 0.97. The model has a recall score of 0.04 for the felony possession class, a 1.0 recall score for the other crime types class, and a weighted average recall score of 0.98. It has a 0.07 F1 score for the felony possession class, a 0.99 F1 score for the other crime types class, and a weighted average F1 score of 0.97.

Further Evaluation of Best Performing Logistic Regression and Random Forest Algorithms via their Receiver Operating Characteristic (ROC) and Precision-Recall Curves

The ROC curves and Areas Under the Curve (AUCs) for the best Logistic Regression algorithm ('upsampled_fp_15’) and the Random Forest algorithm with 100 estimators, as well as the Precision-Recall curve for the Logistic Regression algorithm, were created and calculated for further evaluation.


Logistic Regression 15
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
0.0
0.2
0.4
0.6
0.8
1.0


Random Forest
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
0.0
0.2
0.4
0.6
0.8
1.0



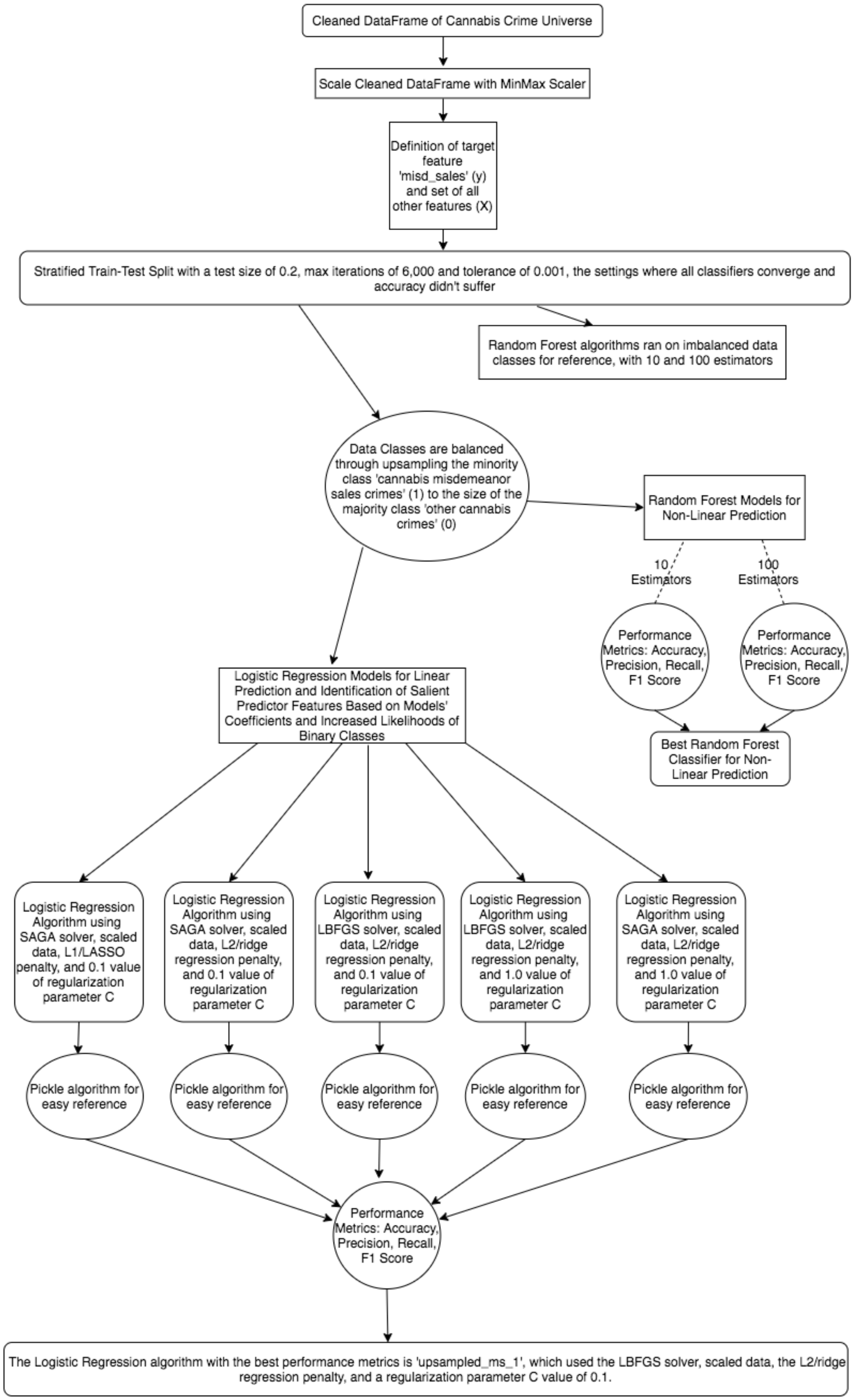
The ROC curves plotted above show that the best LogisticRegression model ('upsampled_fp_15') has an ROC AUC score of 78.4%, showing that it is a moderately skillful model in predicting felony possession crimes at a rate higher than random. The Precision Recall Curve has an AUC score of 11.1%. The very low Precision Recall AUC score shows that the model is not very skilled at differentiating true positives from false positives and false negatives for the felony possession class. However, as is shown above, it is more skillful at predicting the other cannabis crime types as a group.

The ROC curve for the Random Forest algorithm with 100 estimators plotted above shows that it has an AUC score of 73.9%. This is a very unskillful model at predicting felony possession crimes.

Creation and Evaluation of Model Classifying Misdemeanor Sales Crimes

The machine learning classification pipeline of predicting misdemeanor sales crimes followed a logical series of steps, as shown in the visualization on the following page.

Flow Chart of ML Pipeline Classifying Misdemeanor Sales Crimes


Cleaned DataFrame of Cannabis Crime Universe
Scale Cleaned DataFrame with MinMax Scaler
Definition of target feature 'misd_sales' (y) and set of all other features (X)
Stratified Train-Test Split with a test size of 0.2, max iterations of 6,000 and tolerance of 0.001, the settings where all classifiers converge and accuracy didn't suffer
Random Forest algorithms ran on imbalanced data classes for reference, with 10 and 100 estimators
Data Classes are balanced through upsampling the minority class 'cannabis misdemeanor sales crimes' (1) to the size of the majority class 'other cannabis crimes' (0)
Random Forest Models for Non-Linear Prediction
10 Estimators
100 Estimators
Performance Metrics: Accuracy, Precision, Recall, F1 Score
Performance Metrics: Accuracy, Precision, Recall, F1 Score
Best Random Forest Classifier for Non-Linear Prediction
Logistic Regression Models for Linear Prediction and Identification of Salient Predictor Features Based on Models' Coefficients and Increased Likelihoods of Binary Classes
Logistic Regression Algorithm using SAGA solver, scaled data, L1/LASSO penalty, and 0.1 value of regularization parameter C
Logistic Regression Algorithm using SAGA solver, scaled data, L2/ridge regression penalty, and 0.1 value of regularization parameter C
Logistic Regression Algorithm using LBFGS solver, scaled data, L2/ridge regression penalty, and 0.1 value of regularization parameter C
Logistic Regression Algorithm using LBFGS solver, scaled data, L2/ridge regression penalty, and 1.0 value of regularization parameter C
Logistic Regression Algorithm using SAGA solver, scaled data, L2/ridge regression penalty, and 1.0 value of regularization parameter C
Pickle algorithm for easy reference
Pickle algorithm for easy reference
Pickle algorithm for easy reference
Pickle algorithm for easy reference
Pickle algorithm for easy reference
Performance Metrics: Accuracy, Precision, Recall, F1 Score
The Logistic Regression algorithm with the best performance metrics is 'upsampled_ms_1', which used the LBFGS solver, scaled data, the L2/ridge regression penalty, and a regularization parameter C value of 0.1.

Identification of Best Performing Logistic Regression and Random Forest Algorithms By Using Accuracy, Precision, Recall, and F1 Scores

The best LogisticRegression model ('upsampled_ms_1') has an accuracy of 0.688, showing that it makes correct predictions on roughly 68.8% of the data points in the DataFrame.

The model's precision was 0.10 for the misdemeanor sales class (the 1 class), 0.97 for the other cannabis crime types class (the 0 class), and 0.93 on weighted average, showing that 10% of predicted misdemeanor sales crimes were actual misdemeanor sales crimes, 97% of predicted other cannabis crimes were actual other cannabis crimes, and 93% of crimes on weighted average were predicted correctly.

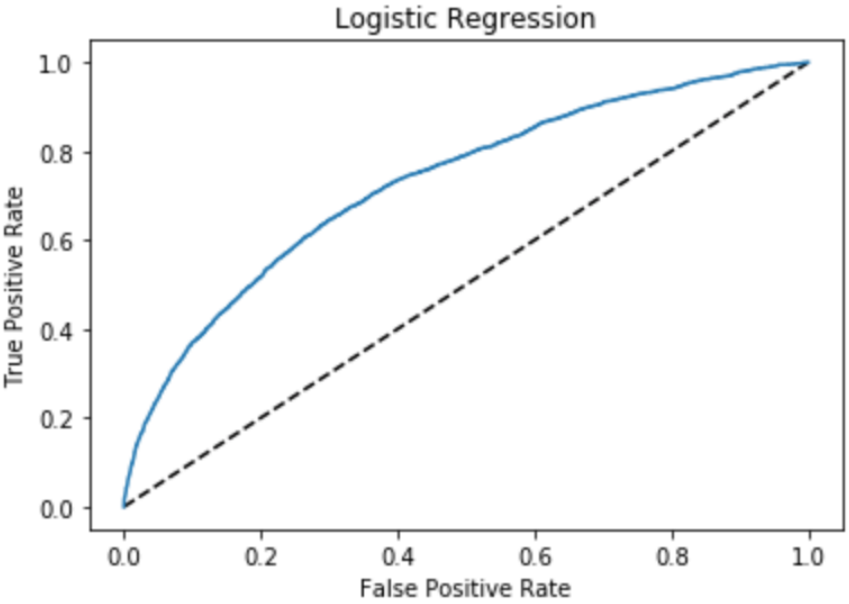
The model's recall was 0.66 for the misdemeanor sales class (the 1 class), 0.69 for the other cannabis crime types class (0), and 0.69 on weighted average, showing that 66% of actual misdemeanor sales crimes were predicted as misdemeanor sales crimes, 69% of actual other cannabis crimes were predicted as other cannabis crimes, and that 69% of crimes on weighted average were predicted correctly.

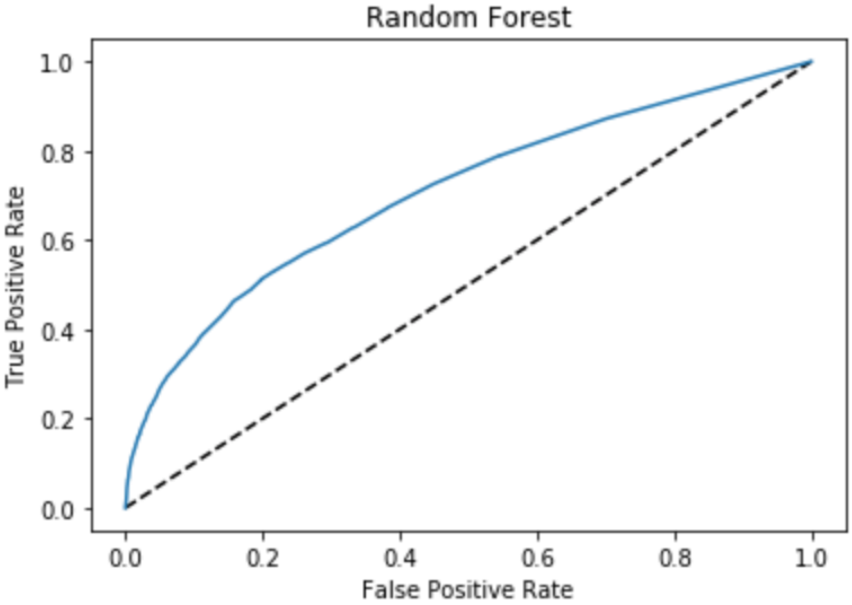
The model's F1 score, or harmonic mean of precision and recall, was 0.18 for the misdemeanor sales class, 0.81 for the other cannabis crimes class, and 0.77 on weighted average. This metric is more informative than precision or recall alone, and shows that only 18% of misdemeanor sales crimes and 81% of other cannabis crimes are being predicted correctly, for a weighted average of 77%.

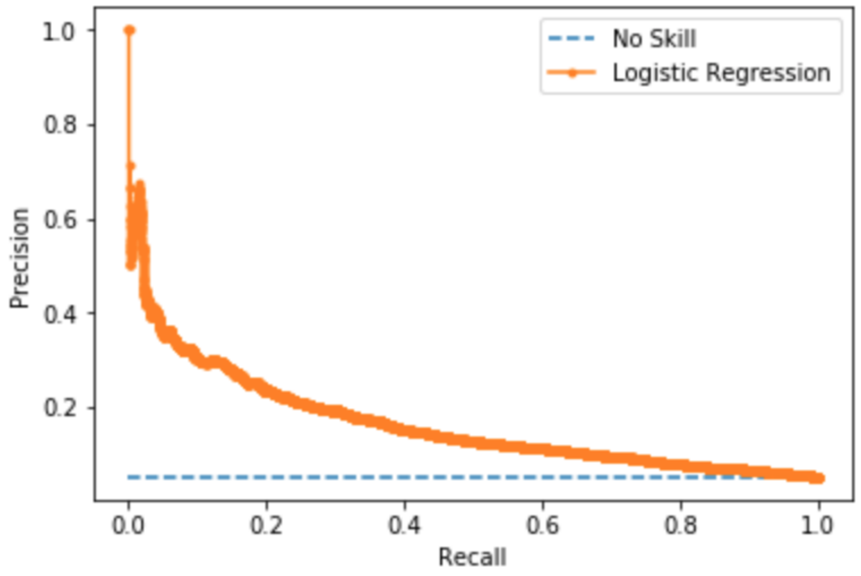
The RandomForest model with 100 estimators has an accuracy of 0.706. It has a 0.30 precision score for the misdemeanor sales class, a 0.96 precision score for the other crime types class, and a weighted average precision score of 0.92. The model has a recall score of 0.17 for the misdemeanor sales class, a 0.98 recall score for the other crime types class, and a weighted average recall score of 0.94. It has a 0.22 F1 score for the misdemeanor sales class, a 0.97 F1 score for the other crime types class, and a weighted average F1 score of 0.93.

Further Evaluation of Best Performing Logistic Regression and Random Forest Algorithms via their Receiver Operating Characteristic (ROC) and Precision-Recall Curves

The ROC curves and Areas Under the Curve (AUCs) for the best Logistic Regression algorithm ('upsampled_ms_1’) and the Random Forest algorithm with 100 estimators, as well as the Precision-Recall curve for the Logistic Regression algorithm, were created and calculated for further evaluation.


Logistic Regression
True Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0


Random Forest
True Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0



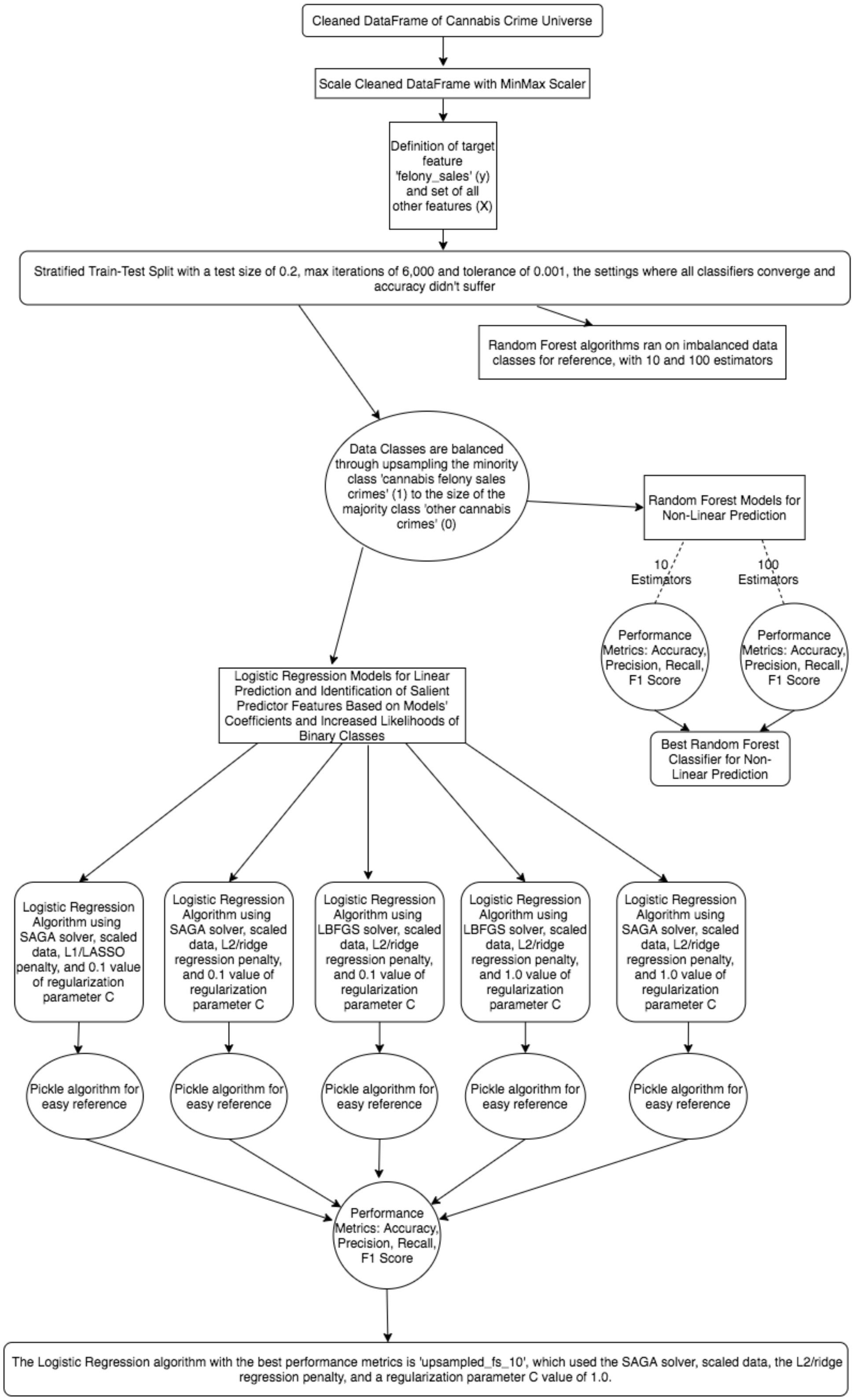
The ROC curves plotted above show that the best LogisticRegression model ('upsampled_ms_1') has an ROC AUC score of 72.9%, showing that it is a moderately skillful model in predicting misdemeanor sales crimes at a rate higher than random. The Precision Recall Curve has an AUC score of 11.6%. The very low Precision Recall AUC score shows that the model is not very skilled at differentiating true positives from false positives and false negatives for the misdemeanor sales class. However, as is shown above, it is more skillful at predicting the other cannabis crime types as a group.

The ROC curve for the Random Forest algorithm with 100 estimators plotted above shows that it has an AUC score of 70.6%. This is an unskillful model at predicting misdemeanor sales crimes.

Creation and Evaluation of Model Classifying Felony Sales Crimes

The machine learning classification pipeline of predicting felony sales crimes followed a logical series of steps, as shown in the visualization on the following page.

Flow Chart of ML Pipeline Classifying Felony Sales Crimes


Cleaned DataFrame of Cannabis Crime Universe
Scale Cleaned DataFrame with MinMax Scaler
Definition of target feature 'felony_sales' (y) and set of all other features (X)
Stratified Train-Test Split with a test size of 0.2, max iterations of 6,000 and tolerance of 0.001, the settings where all classifiers converge and accuracy didn't suffer
Random Forest algorithms ran on imbalanced data classes for reference, with 10 and 100 estimators
Data Classes are balanced through upsampling the minority class 'cannabis felony sales crimes' (1) to the size of the majority class 'other cannabis crimes' (0)
Random Forest Models for Non-Linear Prediction
10 Estimators
100 Estimators
Performance Metrics: Accuracy, Precision, Recall, F1 Score
Performance Metrics: Accuracy, Precision, Recall, F1 Score
Best Random Forest Classifier for Non-Linear Prediction
Logistic Regression Models for Linear Prediction and Identification of Salient Predictor Features Based on Models' Coefficients and Increased Likelihoods of Binary Classes
Logistic Regression Algorithm using SAGA solver, scaled data, L1/LASSO penalty, and 0.1 value of regularization parameter C
Logistic Regression Algorithm using SAGA solver, scaled data, L2/ridge regression penalty, and 0.1 value of regularization parameter C
Logistic Regression Algorithm using LBFGS solver, scaled data, L2/ridge regression penalty, and 0.1 value of regularization parameter C
Logistic Regression Algorithm using LBFGS solver, scaled data, L2/ridge regression penalty, and 1.0 value of regularization parameter C
Logistic Regression Algorithm using SAGA solver, scaled data, L2/ridge regression penalty, and 1.0 value of regularization parameter C
Pickle algorithm for easy reference
Pickle algorithm for easy reference
Pickle algorithm for easy reference
Pickle algorithm for easy reference
Pickle algorithm for easy reference
Performance Metrics: Accuracy, Precision, Recall, F1 Score
The Logistic Regression algorithm with the best performance metrics is 'upsampled_fs_10', which used the SAGA solver, scaled data, the L2/ridge regression penalty, and a regularization parameter C value of 1.0.

Identification of Best Performing Logistic Regression and Random Forest Algorithms By Using Accuracy, Precision, Recall, and F1 Scores

The best LogisticRegression model ('upsampled_fs_1') has an accuracy of 0.708, showing that it makes correct predictions on roughly 70.8% of the data points in the DataFrame.

The model's precision was 0.01 for the felony sales class (the 1 class), 1.0 for the other cannabis crime types class (the 0 class), and 0.99 on weighted average, showing that 1% of predicted felony sales crimes were actual felony sales crimes, 100% of predicted other cannabis crimes were actual other cannabis crimes, and 99% of crimes on weighted average were predicted correctly.

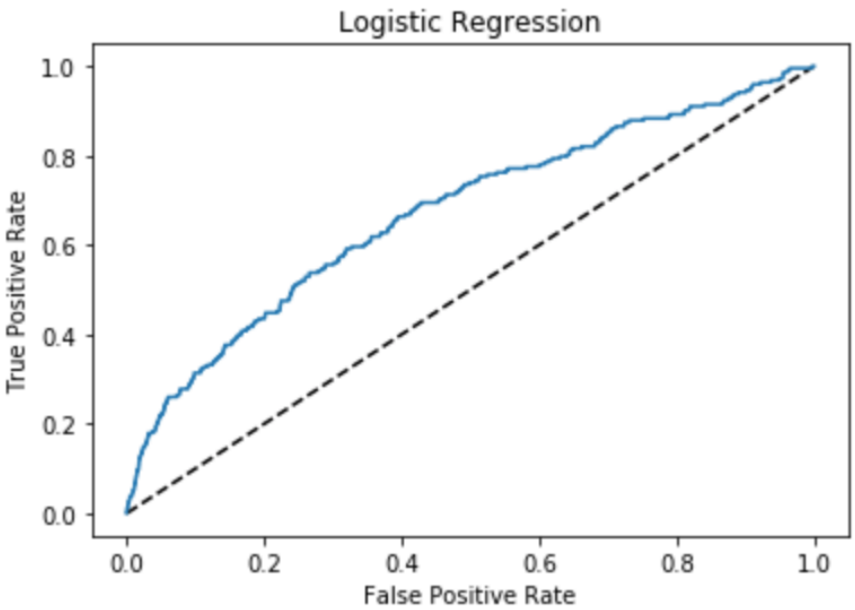
The model's recall was 0.56 for the felony sales class (the 1 class), 0.71 for the other cannabis crime types class (0), and 0.71 on weighted average, showing that 56% of actual felony sales crimes were predicted as felony sales crimes, 71% of actual other cannabis crimes were predicted as other cannabis crimes, and that 56% of crimes on weighted average were predicted correctly.

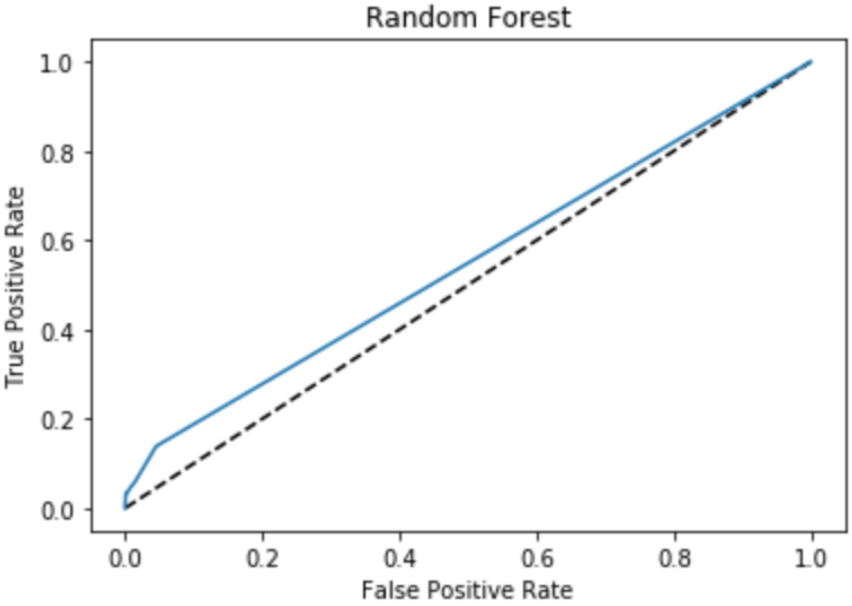
The model's F1 score, or harmonic mean of precision and recall, was 0.02 for the felony sales class, 0.83 for the other cannabis crimes class, and 0.82 on weighted average. This metric was more informative than precision or recall alone, and shows that only 2% of felony sales crimes and 83% of other cannabis crimes were being predicted correctly, for a weighted average of 77%.

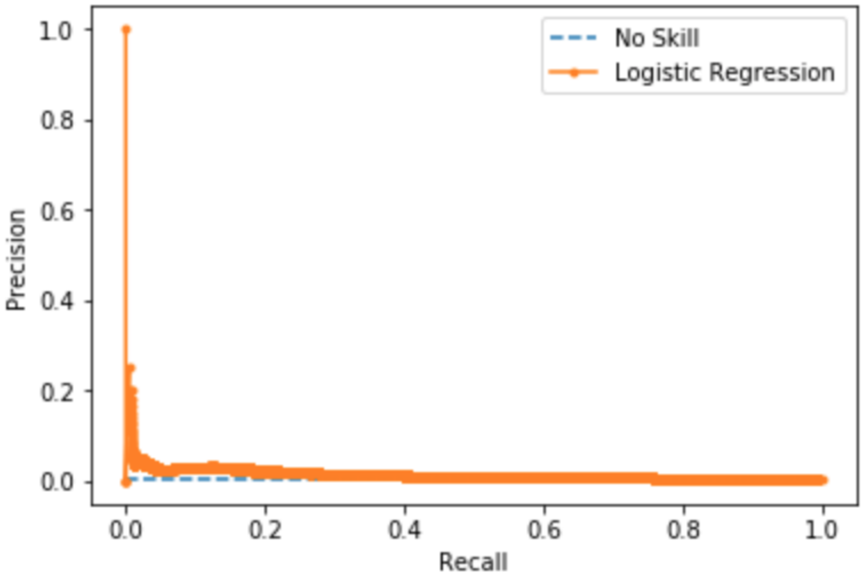
The RandomForest model with 100 estimators has an accuracy of 0.99396291504959. It has a 0.06 precision score for the felony sales class, a 1.0 precision score for the other crime types class, and a weighted average precision score of 0.99. The model has a recall score of 0.01 for the felony sales class, a 1.0 recall score for the other crime types class, and a weighted average recall score of 0.99. It has a 0.02 F1 score for the felony sales class, a 1.0 F1 score for the other crime types class, and a weighted average F1 score of 0.99.

Further Evaluation of Best Performing Logistic Regression and Random Forest Algorithms via their Receiver Operating Characteristic (ROC) and Precision-Recall Curves

The ROC curves and Areas Under the Curve (AUCs) for the best Logistic Regression algorithm ('upsampled_fs_1') and the Random Forest algorithm with 100 estimators, as well as the Precision-Recall curve for the Logistic Regression algorithm, were created and calculated for further evaluation.


Logistic Regression


Random Forest



The ROC curves plotted above show that the best LogisticRegression model ('upsampled_fs_10') has an ROC AUC score of 67.3%, showing that it is not a very skillful model in predicting felony sales crimes at a rate higher than random. The Precision Recall Curve has an AUC score of 1.5%. The extremely low Precision Recall AUC score shows that the model is not skilled at all at differentiating true positives from false positives and false negatives for the felony sales class. However, as is shown above, it is more skillful at predicting the other cannabis crime types as a group.

The ROC curve for the Random Forest algorithm with 100 estimators plotted above shows that it has an AUC score of 54.7%. This is an unskillful model at predicting felony sales crimes.