Crimes in Philadelphia

Team 3

**Data Science Capstone Project   
Predictive Modeling Report**

Date:

03/02/21

Team Members:

Name: Raj Patel

Name: Hong Son

Name: Kunal Sharma

[The purpose of this report is to describe the predictive modeling on the data that you have acquired, pre-processed, and explored in DSCI 591.]

1. Define the Predictive Modeling Problem

1. Input: What are the input data and define the input data clearly?
   1. The input data consists of a combination between the following columns ['Lon', 'Lat', 'Hour', 'Dc\_Dist']. Lon and Lat are the longitude and latitude points. Hour is the hour that the crime occured. Dc\_Dist is a two character field that names the District boundary. These fields were selected from feature selection methods described in the Appendix.
2. Data Representation: What is the data representation?
   1. The overall data representation is in the form of two-dimensional tables with all attributes denoted as numeric values.
3. Output: What are you trying to predict? Define the output clearly.
   1. Using machine learning, we aim to predict the type of crime based on grouping of similar crimes and likelihood of crime that would occur based on a series of results. The exact output would be the name of grouped crimes likely to occur and the respective probabilities.

2. Predictive Models

1. What are the methods? Give a general introduction of the methods with references
   1. KNN
      1. “K Nearest Neighbour is a simple algorithm that stores all the available cases and classifies the new data or case based on a similarity measure. It is mostly used to classify a data point based on how its neighbours are classified.”
      2. <https://towardsdatascience.com/a-simple-introduction-to-k-nearest-neighbors-algorithm-b3519ed98e>
   2. Decision Trees
      1. “A trees classifier implements a meta estimator on various sub-samples of the dataset and uses averaging to improve the predictive accuracy and control over-fitting.”
      2. <https://scikit-learn.org/stable/modules/generated/sklearn.tree.DecisionTreeClassifier.html>.
   3. Random Forest
      1. “Random Forest is a robust machine learning algorithm that can be used for a variety of tasks including regression and classification. It is an ensemble method, meaning that a random forest model is made up of a large number of small decision trees, called estimators, which each produce their own predictions. The random forest model combines the predictions of the estimators to produce a more accurate prediction.”
      2. <https://deepai.org/machine-learning-glossary-and-terms/random-forest>
   4. Logistic Regression
      1. “Logistic regression is the appropriate regression analysis to conduct when the dependent variable is dichotomous (binary). Like all regression analyses, the logistic regression is a predictive analysis. Logistic regression is used to describe data and to explain the relationship between one dependent binary variable and one or more nominal, ordinal, interval or ratio-level independent variables.”
      2. <https://www.statisticssolutions.com/what-is-logistic-regression/>
2. Describe the methods with a pseudo code using the definitions in Section 1.
   1. KNN

1. Load the training and test data

2. Choose the value of K

3. For each point in test data:

- find the Euclidean distance to all training data points

- store the Euclidean distances in a list and sort it

- choose the first k points

- assign a class to the test point based on the majority of classes present in the chosen points

* 1. Decision Tree Classifier

1. D[ecision tree](https://en.wikipedia.org/wiki/Decision_tree) classifiers make observations about an item (represented in the branches) to conclusions about the item's target value (represented in the leaves). Tree models where the target variable can take a discrete set of values are called [classification](https://en.wikipedia.org/wiki/Classification) [trees](https://en.wikipedia.org/wiki/Decision_tree); in these tree structures, [leaves](https://en.wikipedia.org/wiki/Leaf_node) represent class labels and branches represent [conjunctions](https://en.wikipedia.org/wiki/Logical_conjunction) of features that lead to those class labels.

1. Create a X number of unpruned decision trees

* Calculate entropy of the target.
* The dataset is then split on the different attributes. The entropy for each branch is calculated. Then it is added proportionally, to get total entropy for the split. The resulting entropy is subtracted from the entropy before the split. The result is the Information Gain, or decrease in entropy.
* Choose attribute with the largest information gain as the decision node, divide the dataset by its branches and repeat the same process on every branch.

2. Predictions are made by averaging the prediction of the decision trees in the case of regression or using majority voting in the case of classification

* 1. Random Forest

1. Randomly select “k” features from total “m” features.

* Where k << m

2. Among the “k” features, calculate the node “d” using the best split point.

3. Split the node into daughter nodes using the best split.

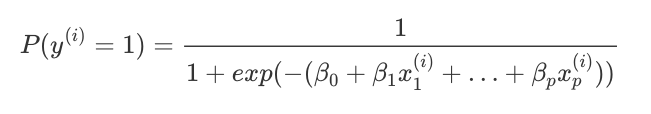
4. Repeat 1 to 3 steps until the “l” number of nodes has been reached.

5. Build forest by repeating steps 1 to 4 for “n” number times to create “n” number of trees.

* 1. Logistic Regression

1. Load the training and test data

2. Fill in the Xn for each variable in the logistic regression function [below]



3. Compute P(y^(i)) (probabilities between 0 and 1)

1. Justify the choice of the method.
   1. KNN

K Nearest Neighbor is the simplest of all the classifiers. It finds the nearing k number of user specified neighbors using euclidean distance to all training data points and returns the first k points. Using the k different points the model then returns the probability of the result for each class.

* 1. Decision Trees

Decision tree classifiers calculate the entropy of the target and split the dataset on different attributes. The results of each attribute is either an information gain or decrease in entropy. The final decision or list of probability from the classifier is based on the attribute with the largest information gain.

* 1. Random Forests

Unlike decision trees, random forest randomly first selects k number of attributes from total number of attributes. It then splits the node after calculating the best splitting point until there is just one node or attribute left. The process keeps repeating n number of times using n number of decision trees. The final decision or list of probability from each of the trees is based on the attribute with the largest information gain.

* 1. Logistic Regression

Logistic Regression calculates and returns a result between 0 and 1 as a probability function where each input will either belong to one class or it won't. The decision is based on a given threshold.

3. Evaluations

1. What metrics do you use for evaluation?
   1. Accuracy
   2. Precision
   3. Recall
   4. F1 Score

In addition to accuracy, we will be using precision, recall, and recall to evaluate the models. Accuracy isn’t always the best indicator as a model could be biased, always predicting one class, etc. By looking at precision and recall, we will be able to access the proportion of relevant instances.

1. What is your ground truth?
   1. Measured for target variable

We found that predicting the crime that might occur to a person is very difficult.

1. Discuss the performance and the limitations of the method.

After finishing our initial model predicting the text\_general\_code, we found the following:

* 1. Multiple classes in predictions is leading to low accuracy.
  2. Same location could have different crimes in the same circumstance.

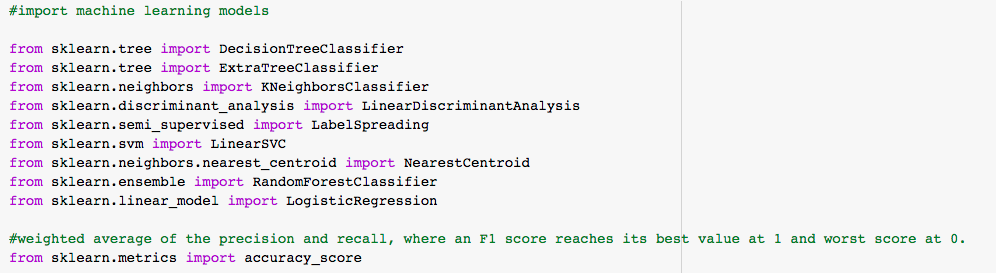
To work off of these problems, we looked at predicting the top 5 crimes, grouping the crimes to similar subsets, and predicting the type of crime (violent and non-violent).

Other limitations of the project included:

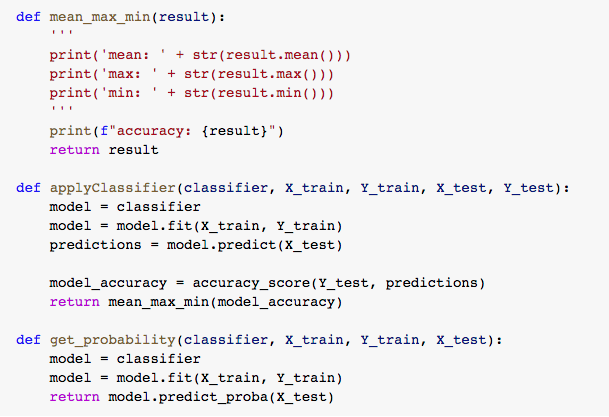
* Lack of helpful attributes to move the needle and improve model performance
* We tried to integrate more datasets, but it was very difficult to integrate other datasets.
* Lack of existing research in this area
* Very little work on optimization of models (class imbalances, bias, ambiguity)

**Appendix**

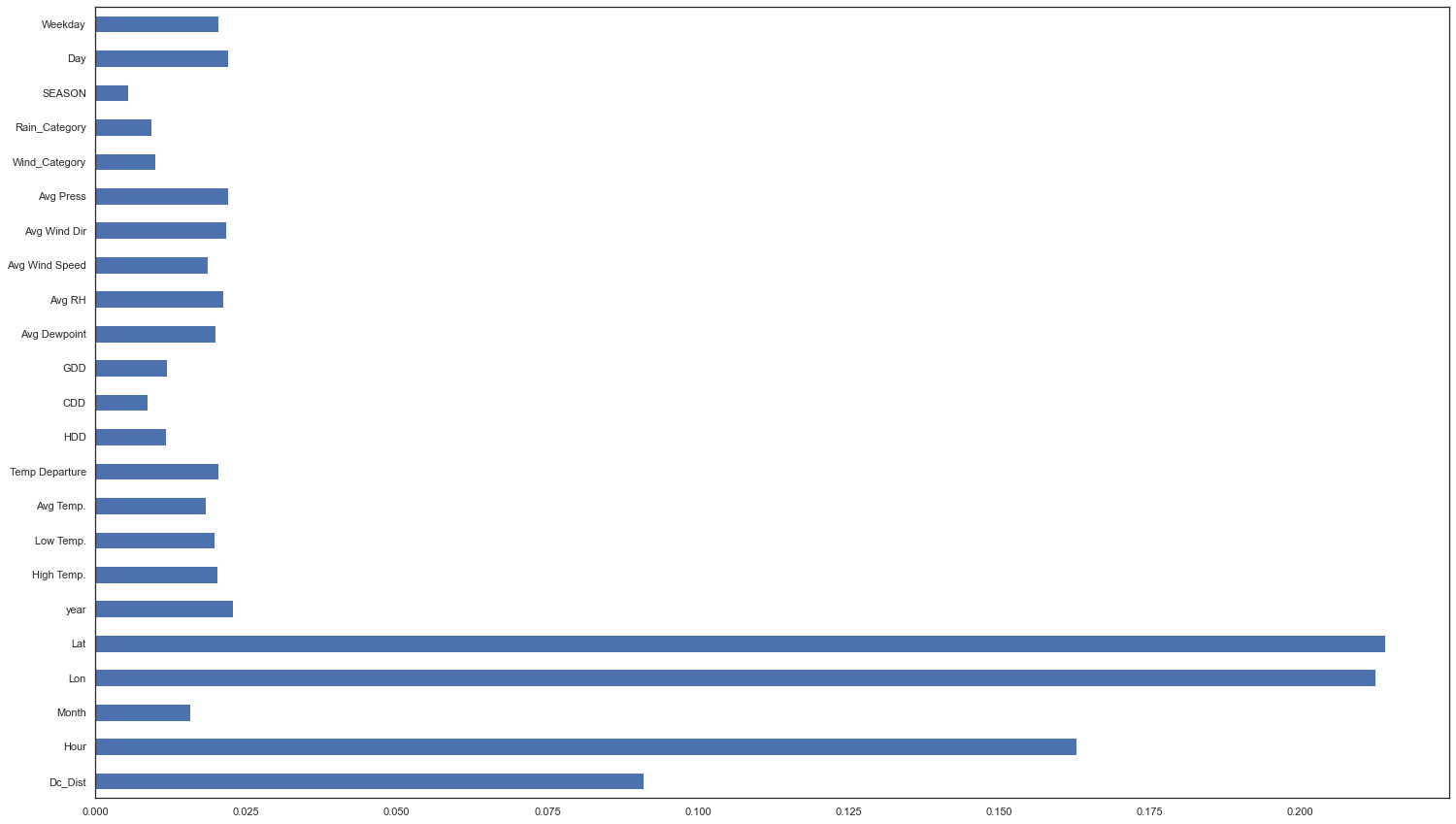
[Addition materials that are not included in the above sections.]



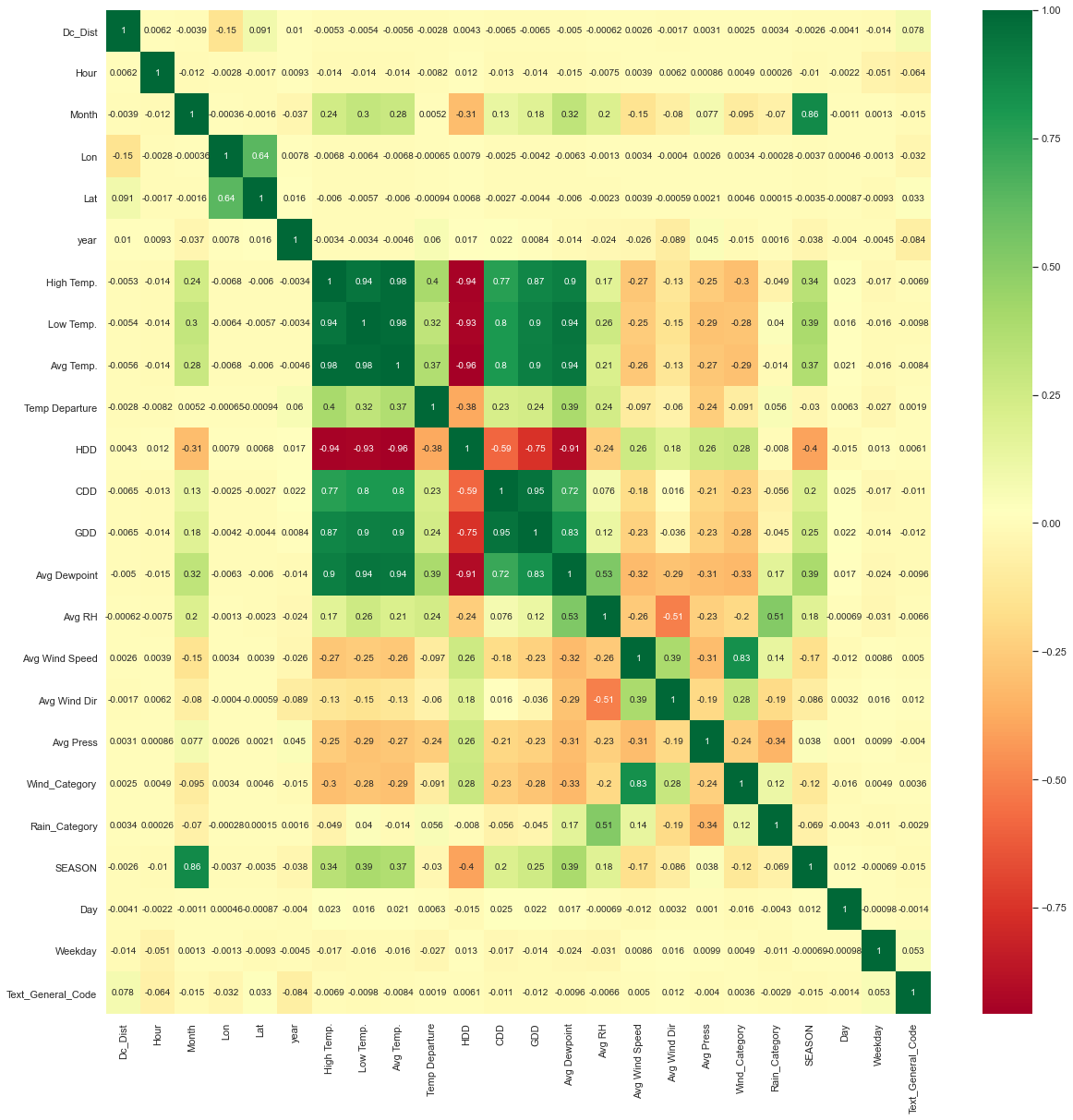
Import the libraries required for machine learning [Extra Trees, KNN, Logistic Regression, Random Forest.



Create functions to apply the classifiers and get the probabilities for each



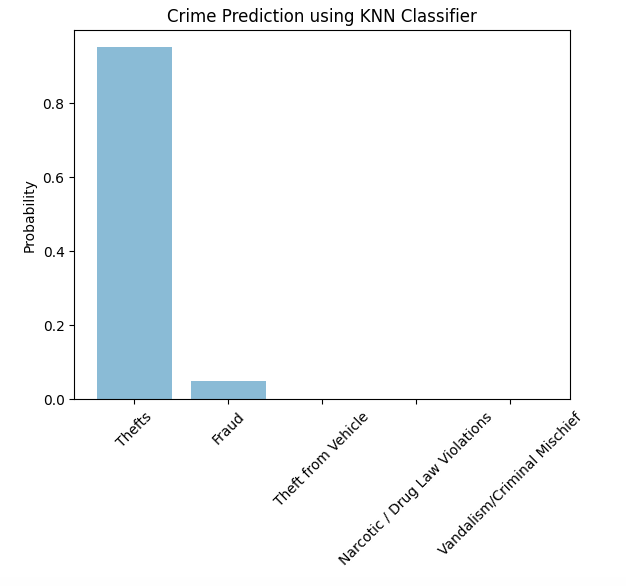
Feature selection using “Extra Trees feature importance” method



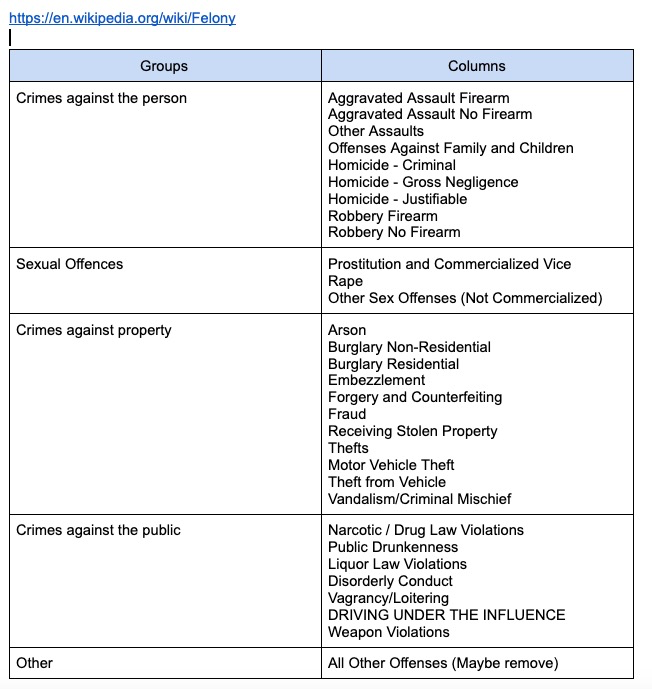
This image above is a correlation matrix that shows how interrelated the attributes of the final data set are. The dark green color denotes high correlation while the dark red represents little to no correlation.



Use the functions above to predict and calculate accuracy

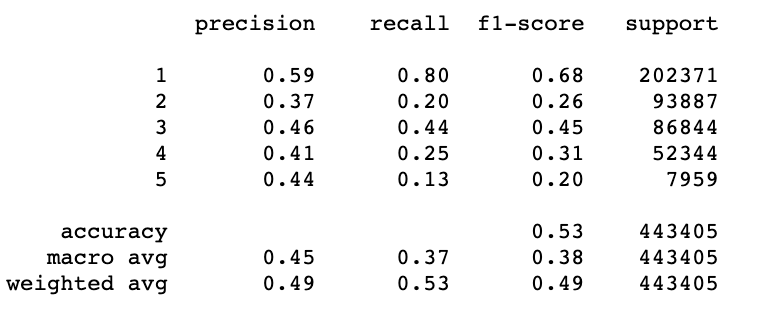


This image is an output displaying the predictions made. Theft has the highest probability of occurring with fraud coming in second. The remaining three crimes do not have any substantial probability of occurring.

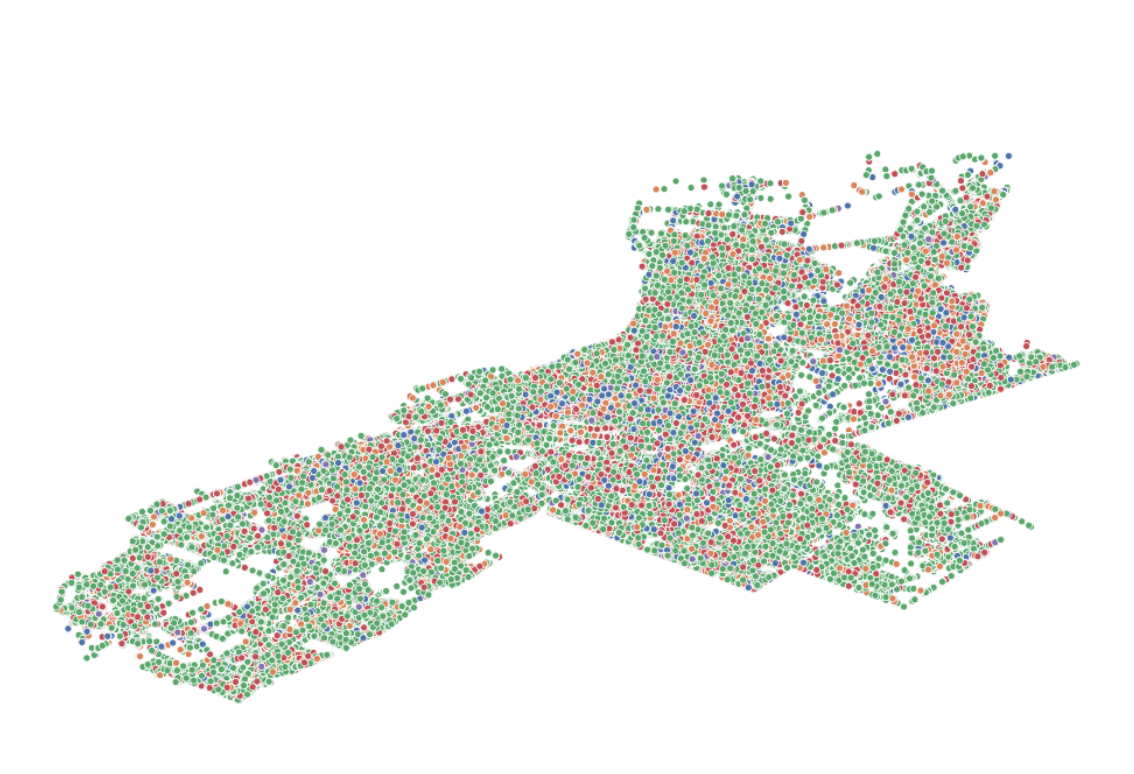


In order to enhance the performance of the model, crimes were grouped to reduce noise. There were 30+ types of crimes before we grouped them into broader categories. The new categorization reduced the crime type count to only 5 variations.

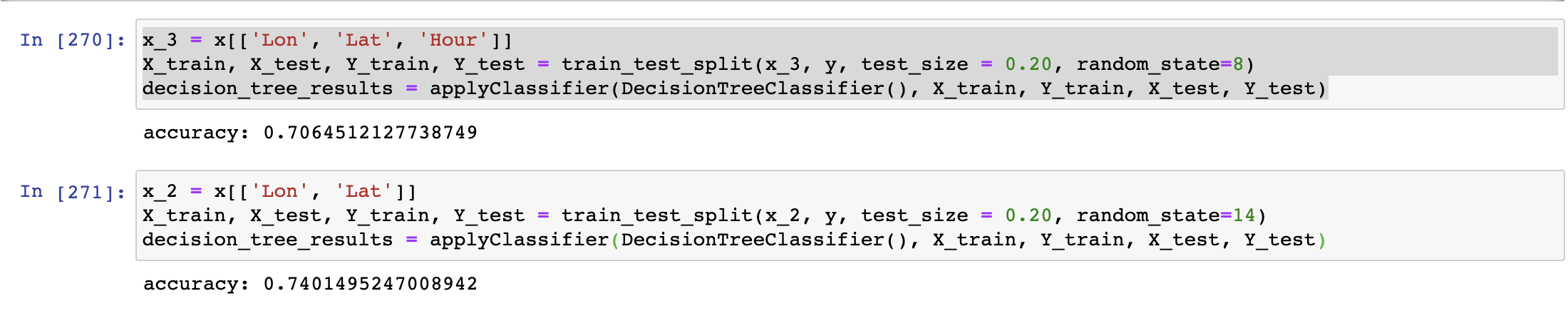


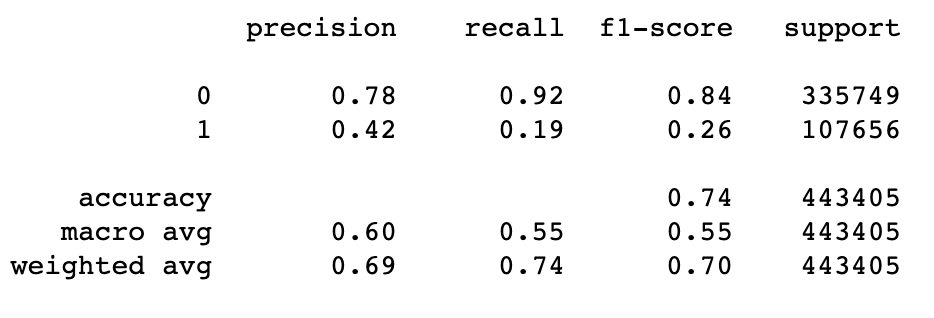


The models ranged from 45-53% accuracy. We see from the classification report that we get very low precision, recall, and f1-score for classes outside of class 1.



A majority of the nearest neighbors or latitude points will be classified as green or class 1.





For binary classification (‘Non Violent’ and ‘Violent’), the accuracy increased significantly as we grouped the crimes in violent and non violent but the model performs very badly for classifying class 1.

In the first predictive modeling report due in Week 5, you only need to use one method in the predictive modeling. Through the experience of implementing and testing the predictive modeling method, you may learn how it works and fails. In the final predictive modeling report due in Week 10, you will need to try more methods and compare with the one you use in the first report.

Table of Contributions

The table below identifies contributors to various sections of this document.

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Section** | **Writing** | **Editing** |
| **1** | **Predictive Modeling Problem Definition** | **Hong & Raj** | **Kunal** |
| **2** | **Predictive Models** | **Hong & Raj** | **Kunal** |
| **3** | **Evaluations** | **Hong & Raj** | **Kunal** |
| **4** | **Appendix** | **Hong & Raj** | **Kunal** |

**Grading**

The grade is given on the basis of quality, clarity, presentation, completeness, and writing of each section in the report. This is the grade of the group. Individual grades will be assigned at the end of the term when peer reviews are collected.