A PROJECT ON MUTATED RANDOM FOREST CLASSIFICATION TO PREDICT THE CATEGORY OF CRIMES OCCURED IN SAN FRANCISCO

MAINACK PAUL

DATA DESCRIPTION

In this project, we are using secondary data to analyze, classify and predict the category of crimes that occurred in San Francisco. The variables used are:

- **Dates** - timestamp of the crime incident
- **DayOfWeek** - the day of the week
- **PdDistrict** - name of the Police Department District
- **Address** - the approximate street address of the crime incident
- **X** - Longitude
- **Y** - Latitude
- **Category** - category of the crime incident. This is the target variable you are going to predict.

SCRUTINY

Any data needs to be verified for its consistency and homogeneity before starting the analysis. This verification of the data is known as scrutiny. First, we scrutinize the data to check if there are any missing values. We observe that this data set is free from such values. We will now start with the exploratory data analysis of the data using various statistical tools.

EXPLORATORY DATA ANALYSIS

Descriptive Analysis helps summarize the data points. This converts the raw data into a form that makes it easier to understand and interpret.

Here, we have used heatmaps and different barplots for visualizing the data.

HEATMAPS

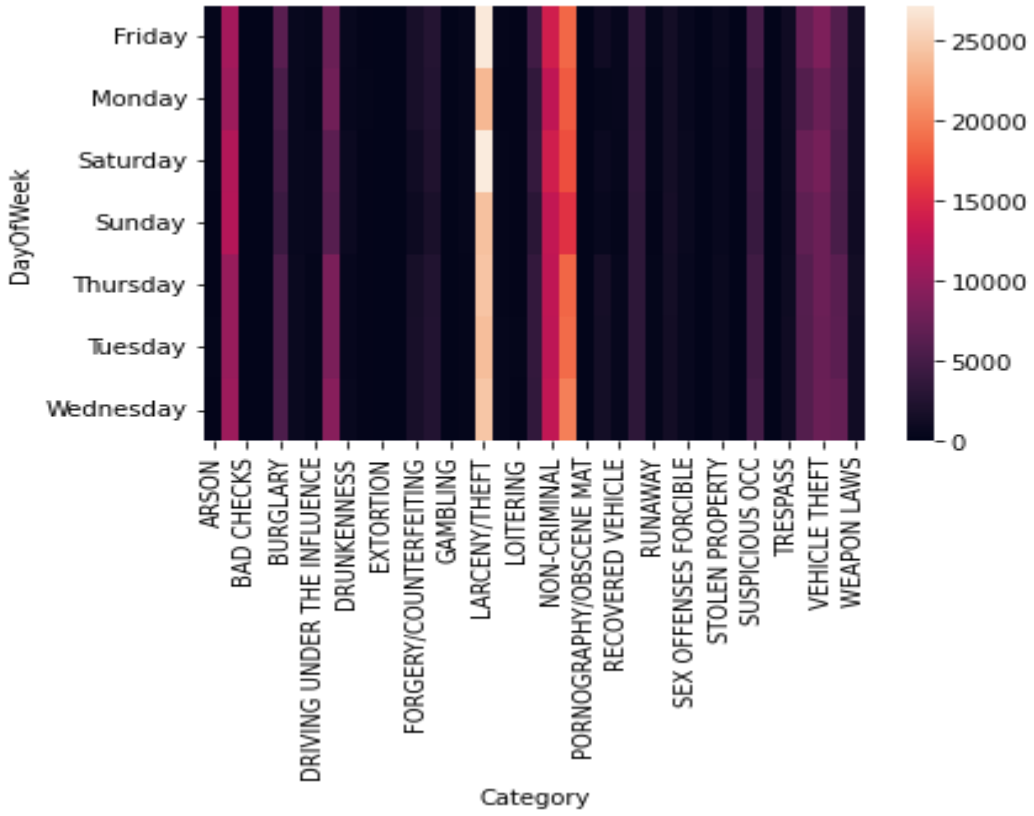

Fig 1: Heatmap between days of week and category

Interpretation: There is no visible change in the distribution of categories within day of week values which indicates lack of importance values of categories.

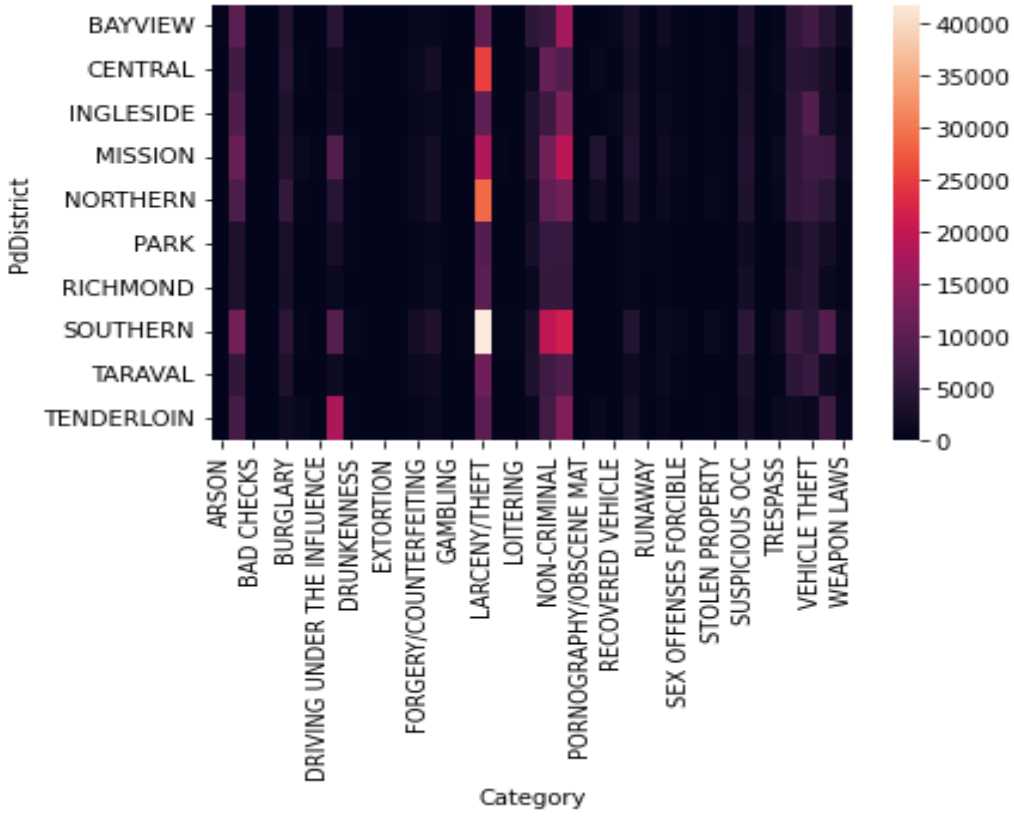

Fig 2: Heatmap between PdDistrict and category

Interpretation: There is some visible change in the distribution of categories within PdDistrict values which indicates importance of the values of categories.

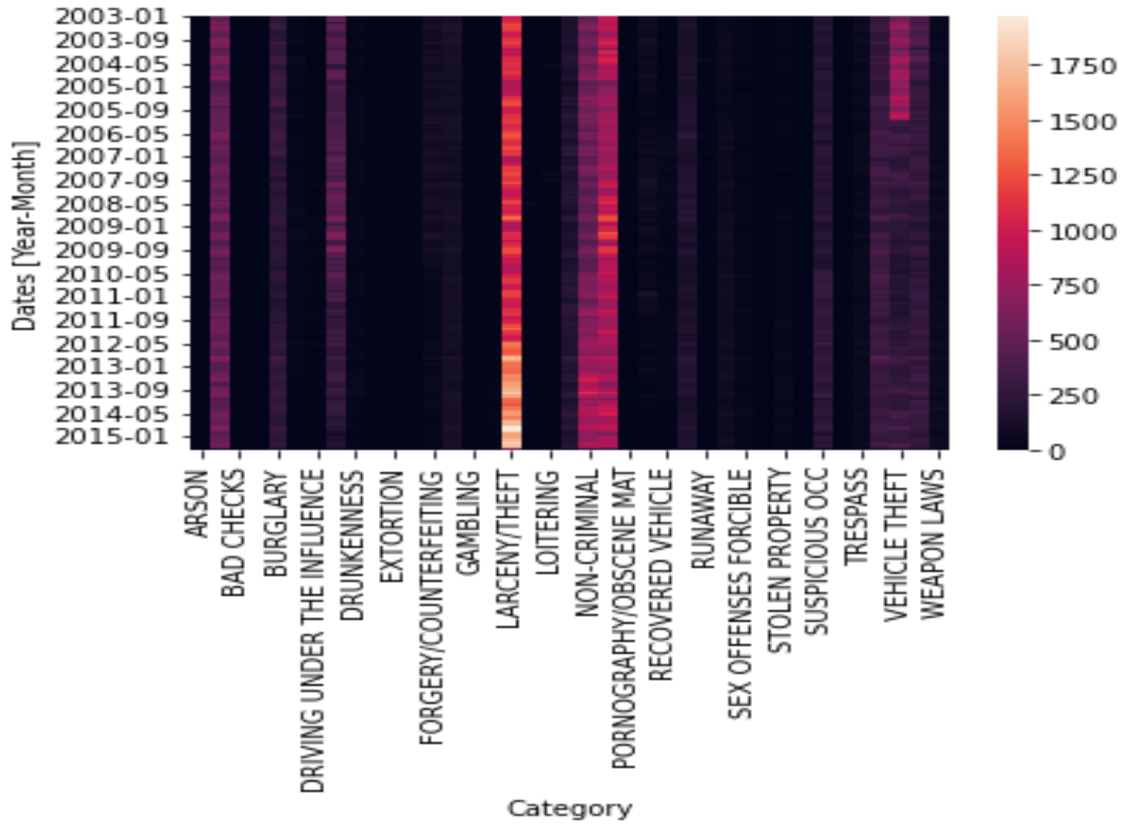

Fig 3: Heatmap between Year-Month and category

Interpretation: There is some visible change in the distribution of categories within Year-Month values which indicates importance of the values of categories.

SUBDIVIDED BARPLOTS

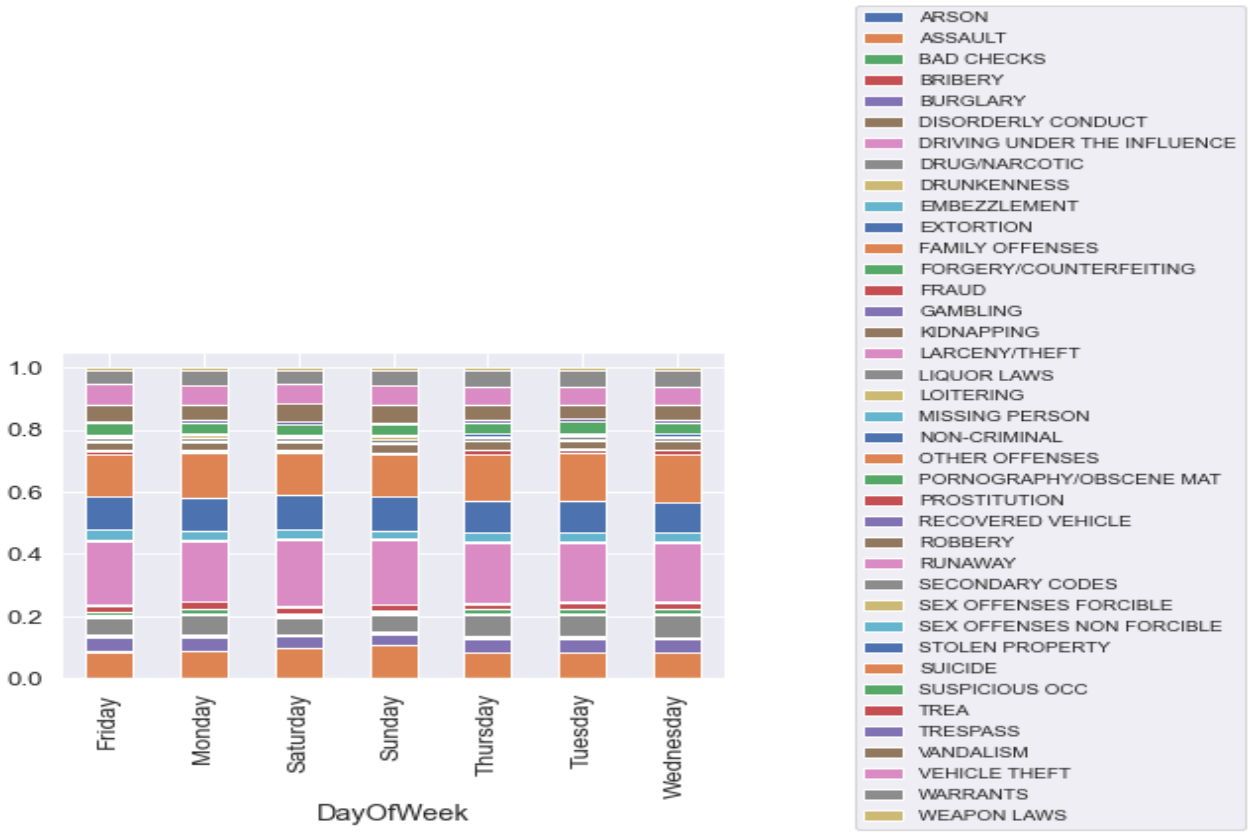

Fig 4: Subdivided bar plot between Category and DayofWeek

Interpretation: The above graph gives a very good visualization. If we closely observe, we can see that a green line (indicating Forgery) at the lower part of the bars is absent for Saturday and Sunday. Again, we can observe that purple line (indicating gambling) is absent on Friday and Monday; and also, a yellow line (indicating drunkenness) is present only on Saturday and Monday.

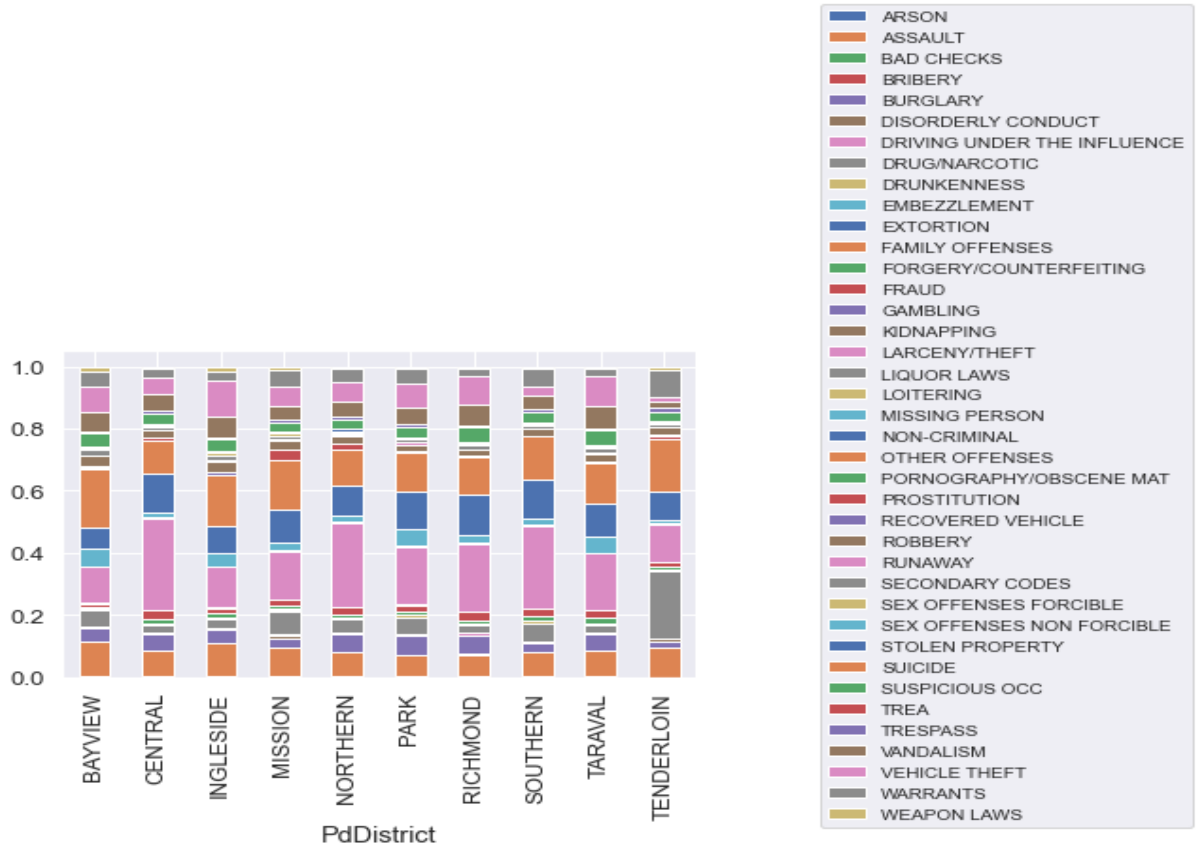

Fig 5: Subdivided bar plot between Category and PdDistrict

Interpretation: The above graph gives a very good visualization. If we closely observe the bottom of the bars, we can see that a very high part of the bar of Tenderlion is warrants (grey portion) relative to other bars. Theft (pink portion at the center) varies rapidly within the districts. Other categories also vary rapidly. Hence, PdDistrict is a very important predictor.

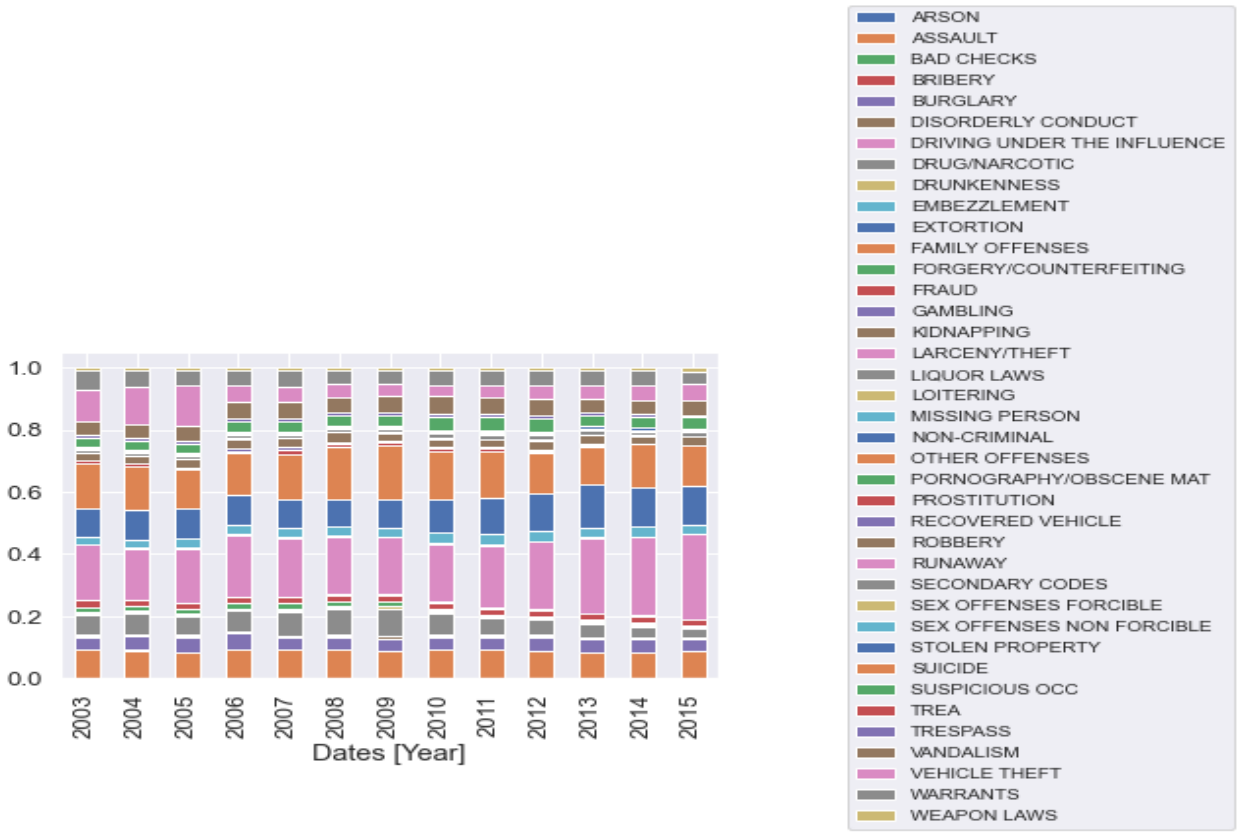

Fig 6: Subdivided bar plot between Category and Year

Interpretation: The above graph gives a very good visualization. Red line (Bribery) is maximum in the year 2007. Theft (pink portion at the upper and center of the bars) varies rapidly within the years. Other categories also vary rapidly. Hence, Year is a very important predictor.

Now, as we proceed, since it is a time series data, we have considered lagged versions of each category (based on PdDistrict) of our response for better understanding. We have considered the weekly average of the past 3 months of each category of our response as our lagged predictors. Then, we standardized each of our lagged predictors and added our new predictors to our dataset. There were many ‘NaN’ values and we replaced all those with 0’s.

We have used target encoding to our Response variable for converting the categorical values into numerical numbers for performing mathematical operations.

Decision Tree Classifier

Decision Tree is a Supervised learning technique that can be used for both classification and Regression problems, but mostly it is preferred for solving Classification problems. It is a tree-structured classifier, where internal nodes represent the features of a dataset, branches represent the decision rules and each leaf node represents the outcome.

We first used a decision tree classifier for fitting of the model and then we performed Cross Validation using Time Series Split (It provides train/test indices to split time series data samples that are observed at fixed time intervals, in train/test sets. In each split, test indices must be higher than before, and thus shuffling in cross validator is inappropriate.) to judge how our model is behaving. We used classification report for better understanding of the model.

A **Classification report** is used to measure the quality of predictions from a classification algorithm. How many predictions are True and how many are False. Classification Report is based on mainly 3 criteria: Precision, Recall and F_1 score.

Precision is the ability of a classifier not to label an instance positive that is actually negative. For each class it is defined as the ratio of true positives to the sum of true and false positives.

Recall is the ability of a classifier to find all positive instances. For each class it is defined as the ratio of true positives to the sum of true positives and false negatives.

The **F_1 score** is a weighted harmonic mean of precision and recall such that the best score is 1.0 and the worst is 0.0.

	precision	recall	f1-score
ASSAULT	0.64	0.63	0.64
MISSING PERSON	0.41	0.38	0.39
ROBBERY	0.26	0.28	0.27

The above is one of the classification reports of our fitted decision tree model.

Random Forest Classifier

Random forest consists of a large number of individual decision trees that operate as an ensemble. Each individual tree in the random forest spits out a class prediction and the class with the most votes become our model's prediction.

	precision	recall	f1-score
ASSAULT	0.65	0.91	0.76
MISSING PERSON	0.74	0.37	0.49
ROBBERY	0.33	0.09	0.14

The above is one of the classification reports of our fitted random forest model.

Mutated Random Forest Classifier

A mutated random forest classifier is a classifier consisting of decision trees each with different parameter values instead of decision trees with all of them having same parameter values.

	precision	recall	f1-score
ASSAULT	0.65	0.89	0.75
MISSING PERSON	0.71	0.36	0.48
ROBBERY	0.32	0.13	0.18
accuracy			0.63
macro avg	0.56	0.46	0.47
weighted avg	0.60	0.63	0.59

The above is one of the classification reports of our fitted mutated random forest model.

COMPARISION

For comparision, we will use the F1 Score and we note down our observations:

- F1 score of Robbery of Random Forest Classifier is very low, Mutated Random Forest Classifier giving mediocre result and decision tree classifier is working fine.
- F1 score of Assault and Missing Person is very high for both Random Forest Classifier and Mutated Random Forest Classifier whereas it is not so good for decision tree classifier.

Hence, using the above observations, one thing we can say is that, Mutated Random Classifier is working better overall for all of the Categories relative to the other two.