National Crime Victimization Survey (NCVS) Data Analysis Around Police Reporting

CMPE 255 Spring 2022

Sahithi Bommadi
015949219

Uttej Kumar Reddy Gade
016065543

1. Introduction

i. Motivation

The United States of America ranks among the Top 60 nations[1] with the highest crime rate globally. In spite of this, the percentage of crimes reported to the police are not as high as they should be. Through the data collected by the National Crime Victimization Survey, we aim to explore the datasets and understand the key features behind this.

ii. Objective

The objective of this project is to explore the characteristics which influence a person or a household's decision to report a victimization to the police. Then use these characteristics to arrive at the best model among the chosen models to predict whether a person or a household is likely to report a particular victimization.

iii. Approach

1. Conduct Exploratory Data Analysis on Personal and Household Victimization Datasets to understand the features of the dataset and their relevance to the reporting.
2. Perform Data Cleaning and Preprocessing to ready the data to be used by multiple machine learning models.
3. Run the models on the data and compare various evaluation metrics to determine the best model for the prediction task.
4. Explore options to improve model performance.

iv. Literature / Market review

There are very few existing analyses on this dataset and no comparison of models geared towards the aforementioned objective that we could find.

v. Datasets Used

National Crime Victimization Survey collects information on nonfatal crimes from a nationally representative sample of around 150,000 U.S. households and 240,000 people per year. These are crimes against persons age 12 or older, both reported and non-reported to the police. At the time of the project proposal, the Bureau of Justice Statistics website hosted data from the year 1993 to the year 2019. The NCVS produces data on both personal and household crimes.

Personal Victimization
https://www.bjs.gov/developer/ncvs/data/csv/NCVS_PERSONAL_1993-2019.zip
Household Victimization
https://www.bjs.gov/developer/ncvs/data/csv/NCVS_HOUSEHOLD_1993-2019.zip

2. Task Distribution

- Sahithi Bommadi - Exploratory Data Analysis, Model evaluations and improvements, Documentation related to Personal Victimization dataset
- Uttej Kumar Reddy Gade - Exploratory Data Analysis, Model evaluations and improvements, Documentation related to Household Victimization dataset

3. Implementation

The following models were chosen to be evaluated on the task of prediction of reporting.

a. Linear Regression
b. Logistic Regression
c. Decision Trees
d. Random Forest
e. K-Nearest Neighbours
f. K-Means and Support Vector Machine
g. Neural Networks

The following technologies were used for the aforementioned tasks:

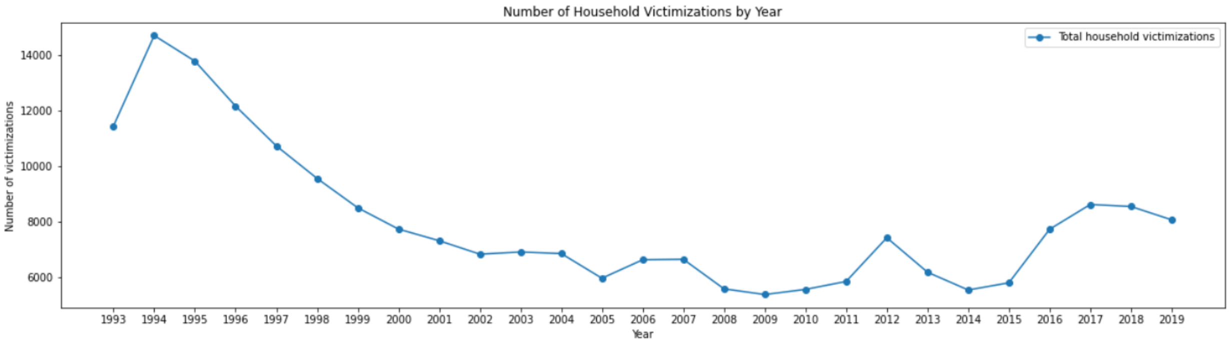
a. pandas
b. matplotlib
c. seaborn
d. numpy
e. sklearn

Implementation - Household Victimization

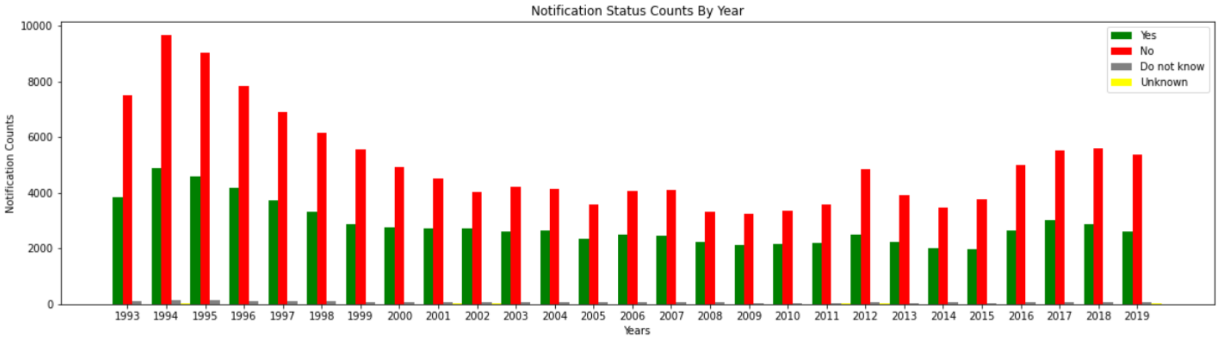
Exploratory Data Analysis

Note: pdfs of reports are stored at https://github.com/uttejkumarreddy/NCVS-Data-Analysis-And-Reporting-Prediction/tree/master/images

1. What is the number of household victimizations per year?

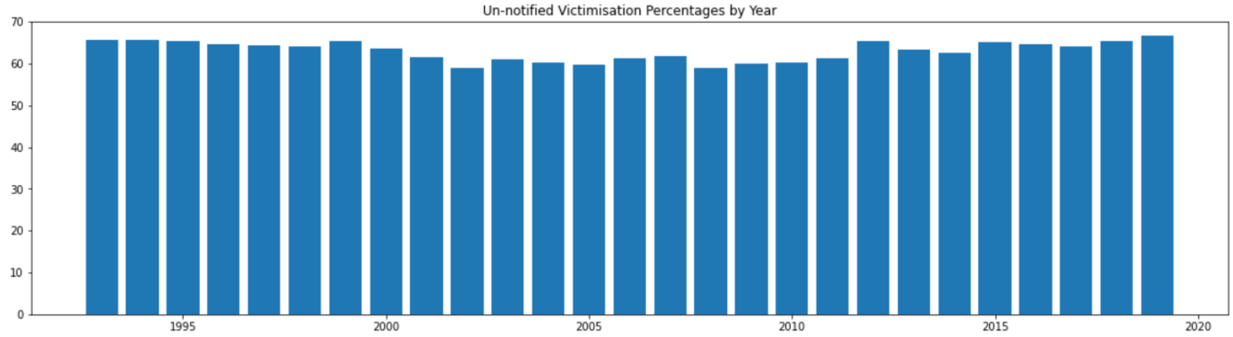


2. What are the reporting status counts through the years



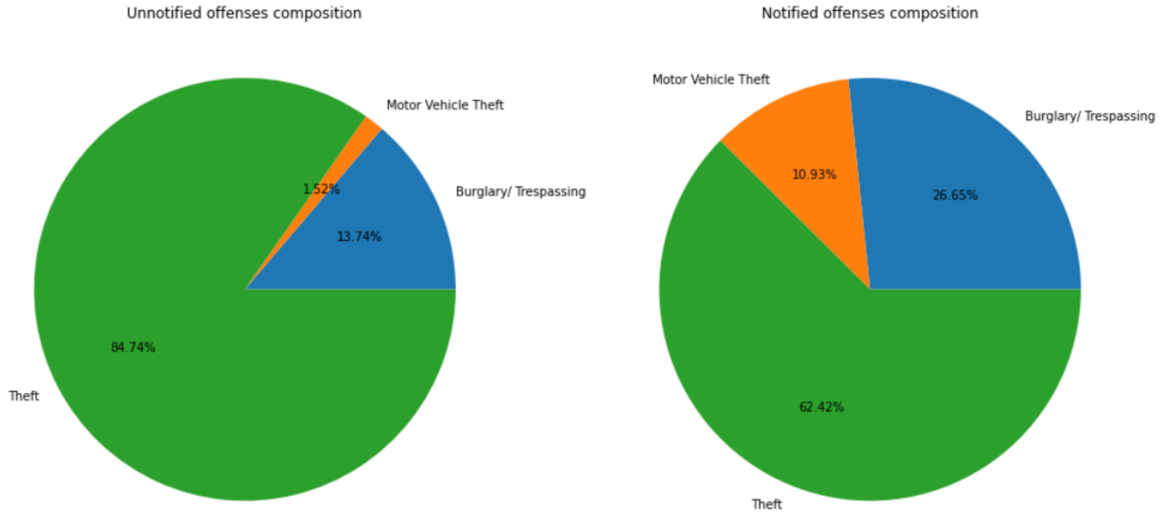
From 1993-2019, the number of unnotified crimes has been consistently and significantly higher than notified crimes

3. By what percentage are the unnotified victimisations higher than the notified through the years



Every year, of the victimisations that occur, the percentage that go unnotified is around 60% higher than those that are notified

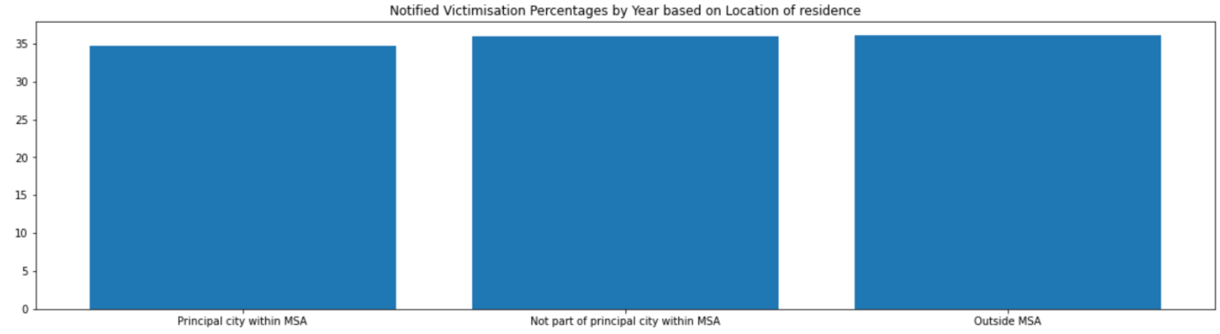
4. What type of crimes are most likely to be reported? Similarly, what crimes are most likely to go unnotified?



Of the victimisations that go unnotified, Theft is the first, followed by Burglary/ Trespassing and then, Motor Vehicle Theft.

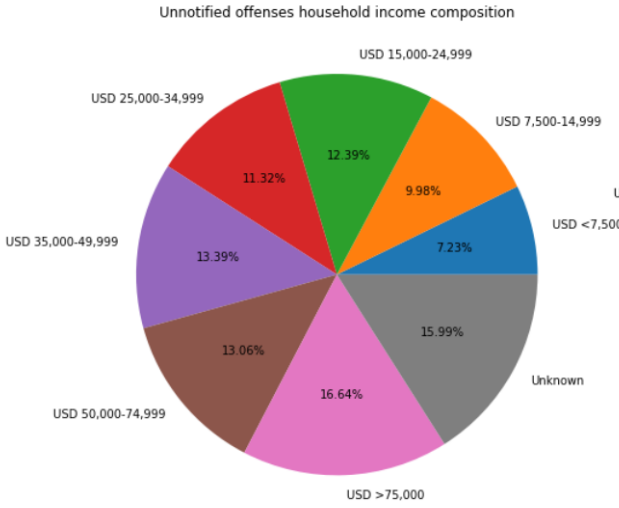
Theft is also the category which is most reported to the police amongst household victimisations which do get reported, followed by Burglary/ Trespassing and finally Motor Vehicle Theft.

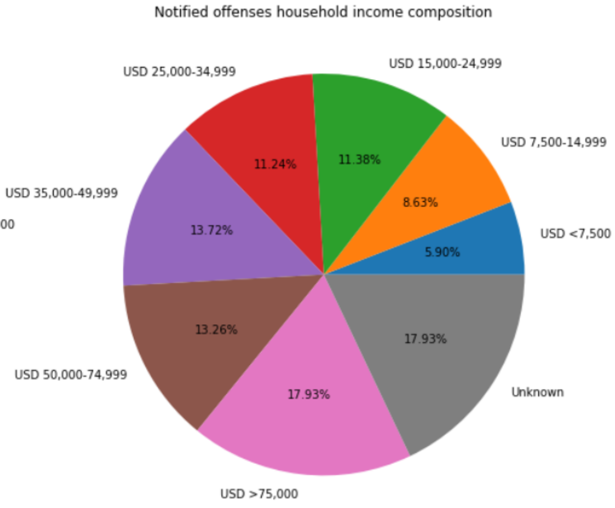
5. How does the location of the residence of the household affect notifications?



There is a very slight difference, but household situated outside MSAs report a higher percentage of victimisations to the police

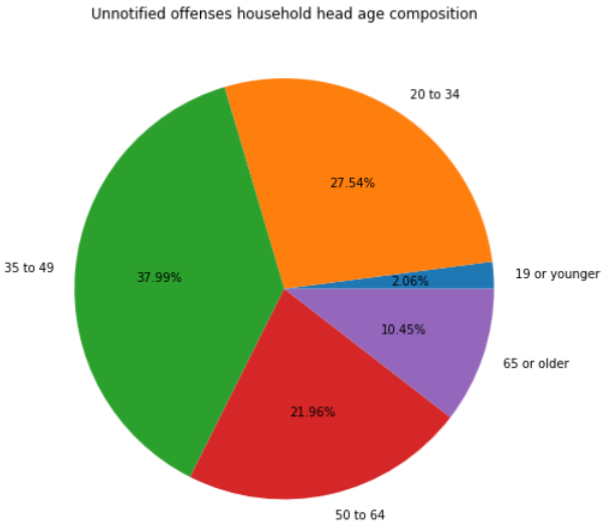
6. How does household income affect notifications?

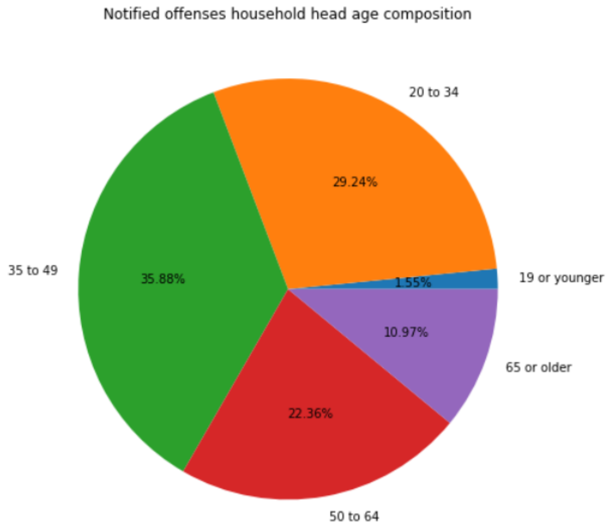




We see that there is not a huge difference in notification status among households of different income levels. The most difference is seen for households with income level less than $7000 and households of income $7500-14,999 wherein the percentage difference between notified and unnotified is ~1.3% favoring towards unnotified.

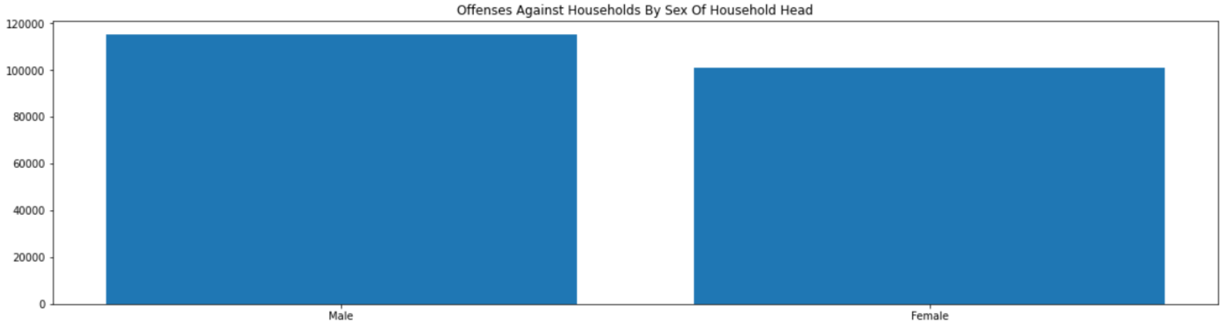
7. How does age affect the decision to report a victimization?


Unnotified offenses household head age composition


Notified offenses household head age composition

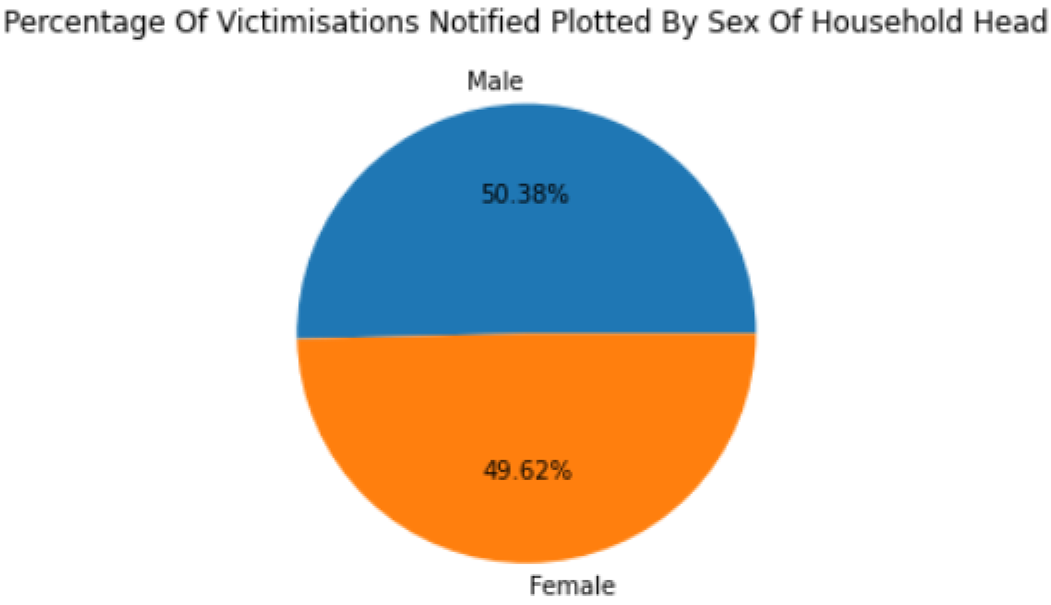
From the above charts, it can be concluded that age of the household head does not hugely influence the decision to report victimisations.

8. Does victimization count change depending on the sex of the household head?


Offenses Against Households By Sex Of Household Head

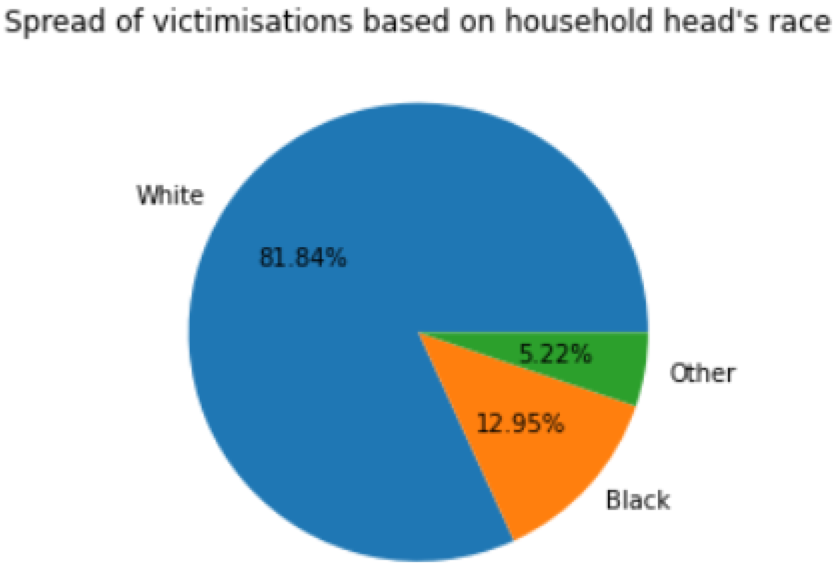
There is only a slight difference

9. Is there a change in the notification percentage based on the sex of the household head?



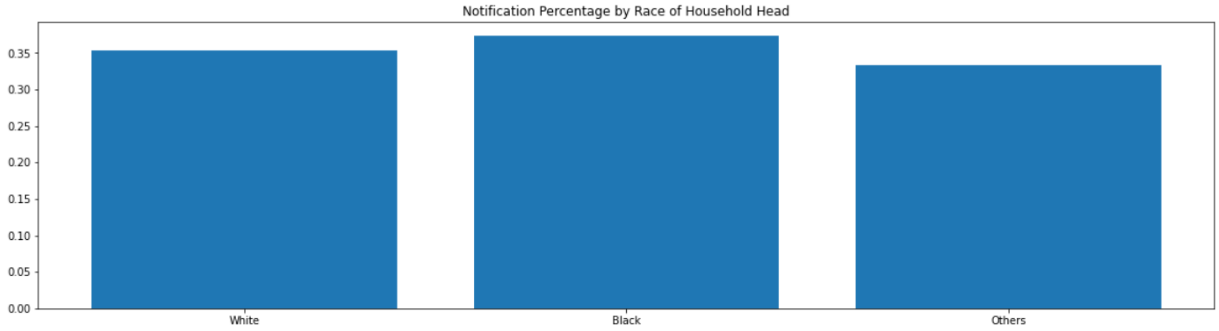
The percentage of reportings based on the sex of the household head is nearly the same for both sexes

10. Which race (of household head) is most affected by household victimisations?



Predominantly, households with head's of race White are more prone to victimisations

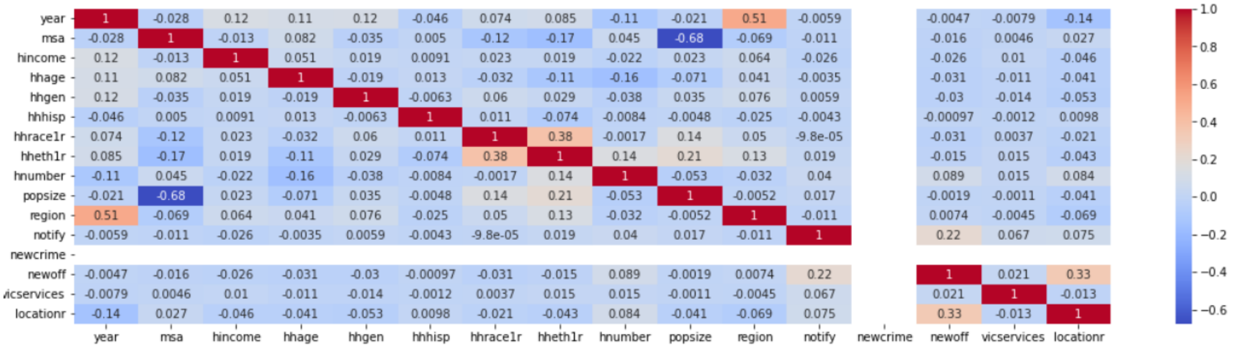
11. Does the race of the household head affect reporting?



Household head's of Black origin have a slightly higher reporting percentage

Data Preprocessing

1. Handle empty values - Out of the 216,269 rows, 38,832 had empty popsize values. Assuming that msa (location of residence) and popsize (population size) have a direct correlation, the missing popsize values were filled with the median of popsize values for their corresponding MSAs.
2. Handle empty values - There were 1275 rows with empty vicservices. As that is a very small percentage and because there is no accurate way to ascertain whether a service was provided after victimization or not, these rows were dropped.
3. Remove outliers - We are interested in predicting given certain characteristics of a victimization, if it is likely to be reported to the police or not. The 'notify' column however contains two more values, 'Do not know' (3) and Unknown (8). As these values (3 and 8) make up for a very small percentage of the total data, these rows can be considered as outliers and dropped.
4. Drop columns - weight column was determined to be unnecessary and dropped.
5. Drop columns - After plotting a heatmap with correlations, ‘newcrime’ column was determined to be unnecessary and dropped.

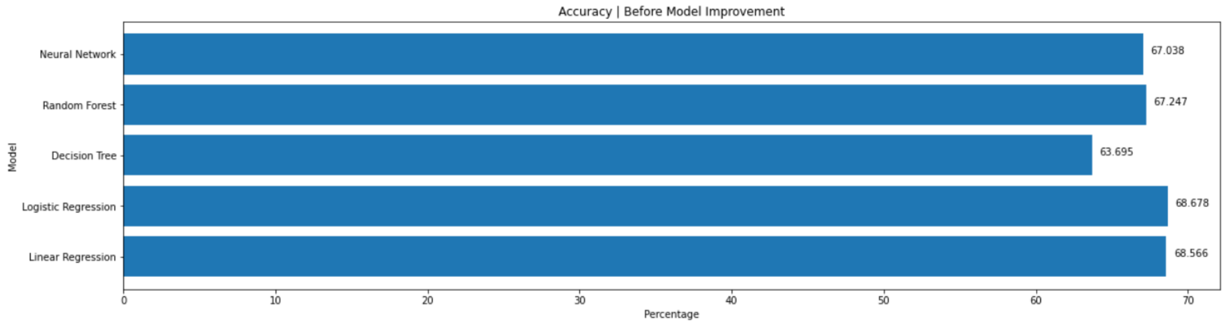


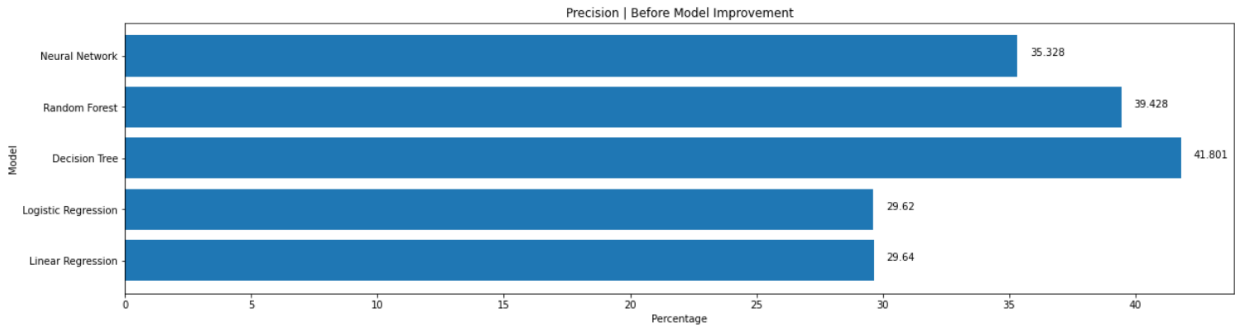
6. One hot encoding - After the above step, the dataset was entirely made of categorical values. Hence, everything was decided to be one-hot encoded so as to make the models give equal importance to them.
7. Replace notify values - To make ‘notify’, our target column consistent with other columns (binary), 2 (No) was replaced with 0, to make this column binary too.

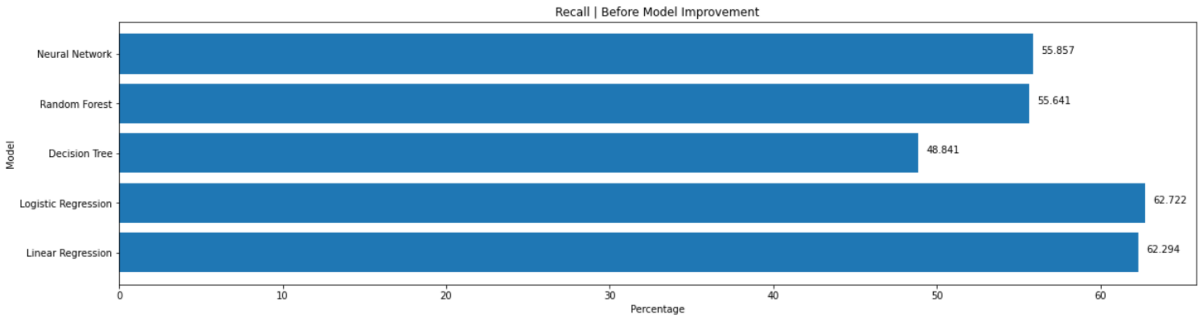
Model Training

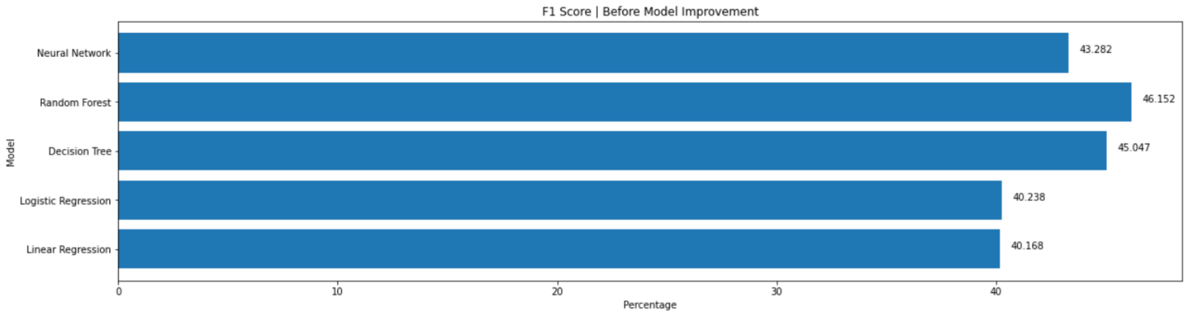
1. Linear Regression, Logistic Regression, Decision Trees, Random Forest and Neural Networks were trained with default configurations.
2. KNN is trained with default configuration too. However, for K Means, the classification is combined with SVM to see if it can produce a better performance.

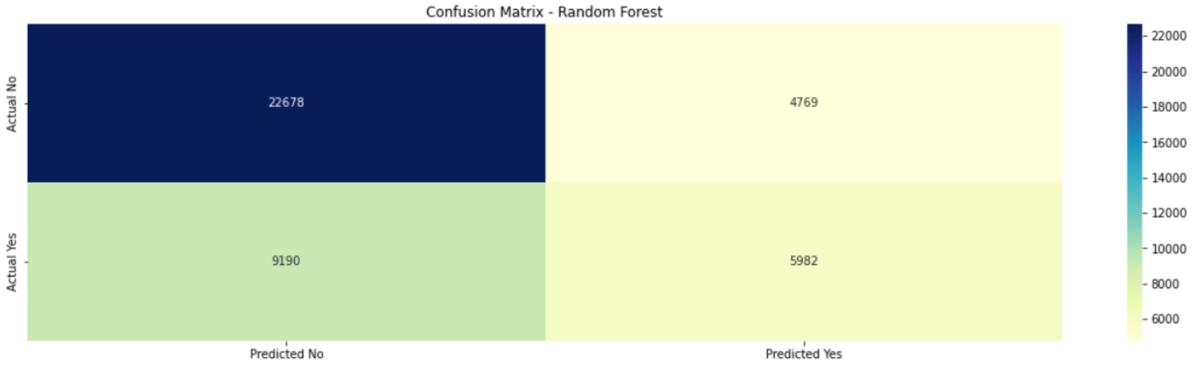
Model Evaluation - I




Precision | Before Model Improvement
Model
Neural Network
Random Forest
Decision Tree
Logistic Regression
Linear Regression
35.328
39.428
41.801
29.62
29.64
0
5
10
15
20
25
30
35
40
Percentage


Recall | Before Model Improvement
Model
Neural Network
Random Forest
Decision Tree
Logistic Regression
Linear Regression
55.857
55.641
48.841
62.722
62.294
0
10
20
30
40
50
60
Percentage


F1 Score | Before Model Improvement
Model
Neural Network
Random Forest
Decision Tree
Logistic Regression
Linear Regression
43.282
46.152
45.047
40.238
40.168
0
10
20
30
40
Percentage


Confusion Matrix - Random Forest
Actual No
Actual Yes
Predicted No
Predicted Yes
22678
4769
9190
5982
22000
20000
18000
16000
14000
12000
10000
8000
6000

Logistic Regression has the highest accuracy score amongst the other 5 models - 68.678 % . It also has the least precision and highest recall which implies that even though its accuracy is high it is not the best model for the task.

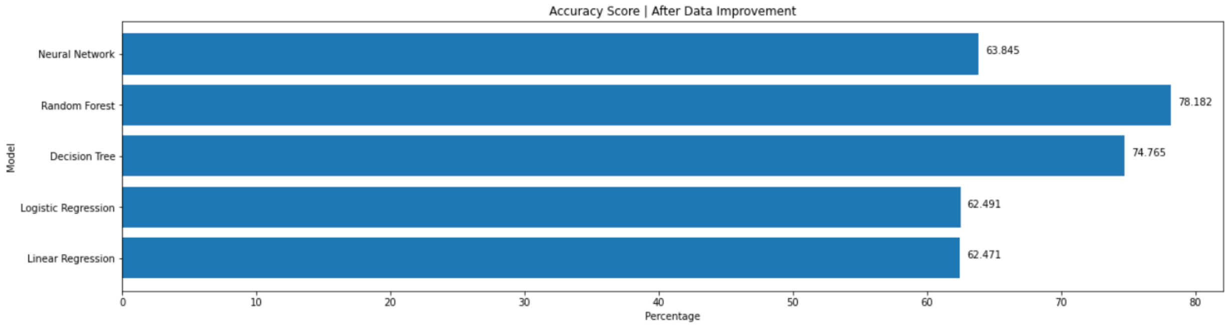
On the other hand, even though Random Forest and Decision Tree have slightly lesser accuracy than Logistic Regression, they have a higher precision and lower recall which imply that they are better models for the task.

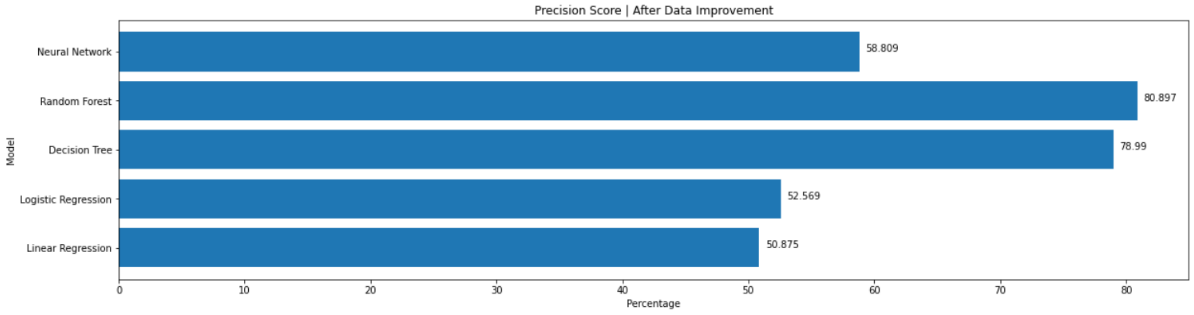
Between Random Forest and Decision Tree, Random Forest has better accuracy and a better F1 score, which builds up to the conclusion that Random Forest is the best suited model.

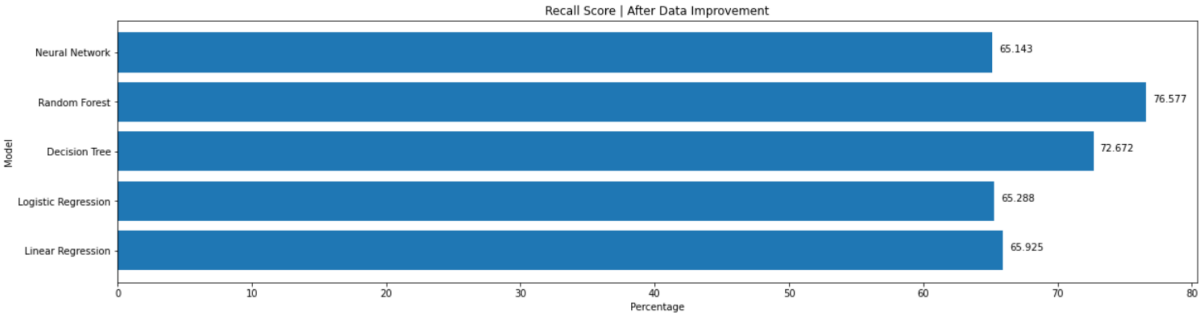
Model Improvement

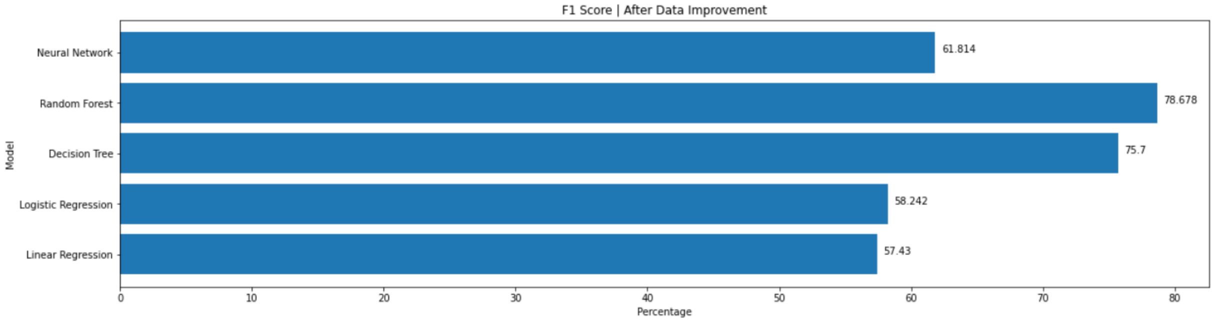
The number of unnotified crimes (notify = 2) is almost twice the number of notified crimes (notify = 1). This might be leaning the model predictions more towards unnotified crimes. To resolve this, we will 'oversample' rows with notify = 1 to make them equal in count with those with notify = 2. This enhanced dataset, with equal number of rows of classes notify is again sent to the above models.

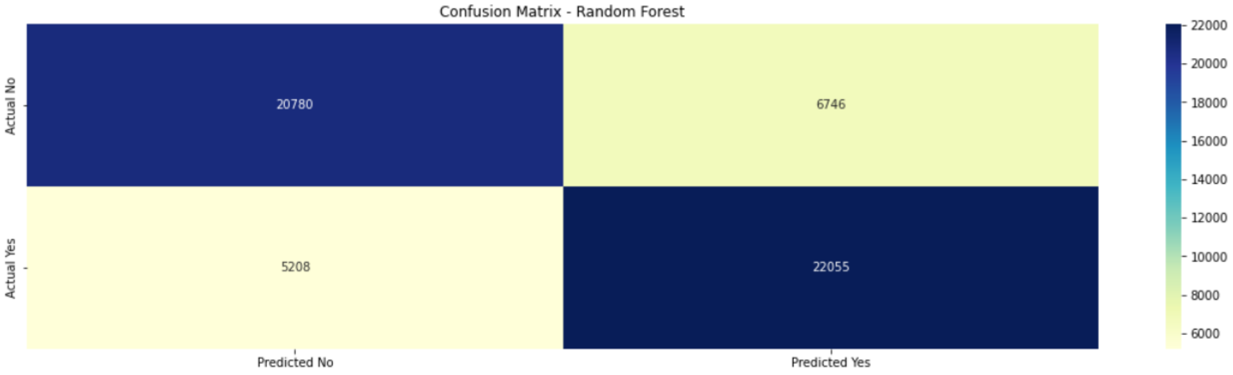
Model Evaluation - II







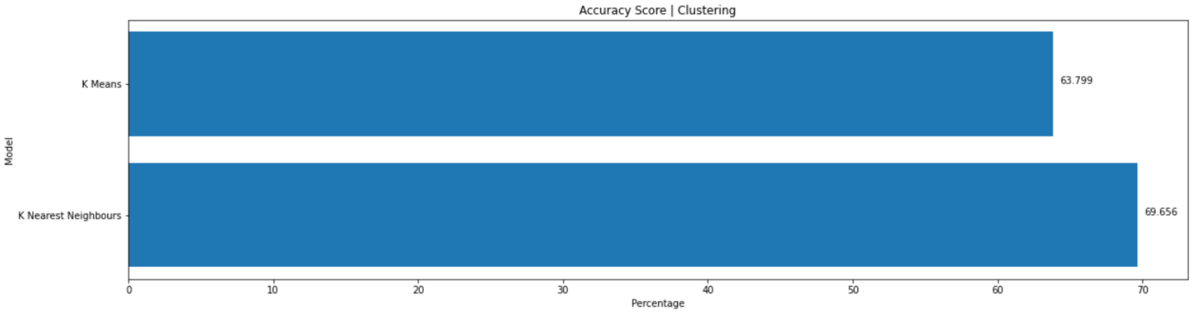




With oversampling of the data, Random Forest model performed the best with an accuracy of 78.182% and a F1 score of 78.678%

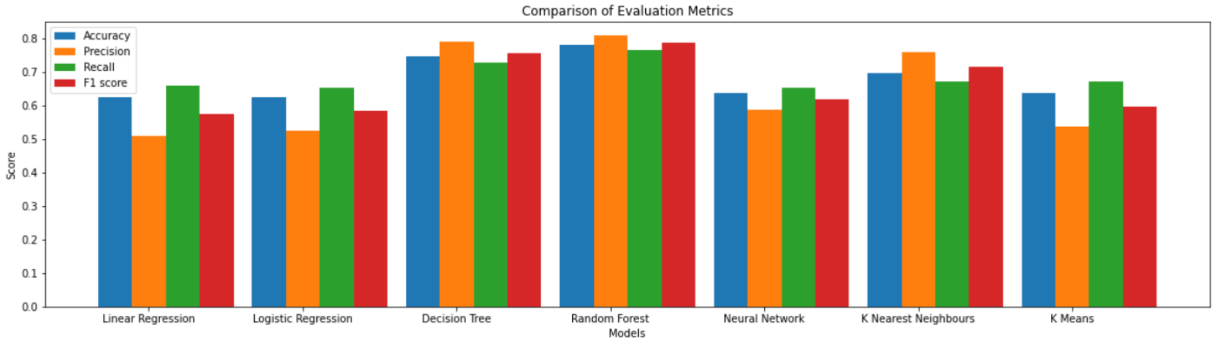
Model Evaluation - III

Due to the high dimensionality (81 columns), clustering will not be very efficient. To prepare data for clustering models, we will perform PCA and reduce dimensions to 4 (chosen arbitrarily). Then clustering algorithms are run.



Clustering algorithms with the current configuration have a lesser accuracy percentage than Random Forest

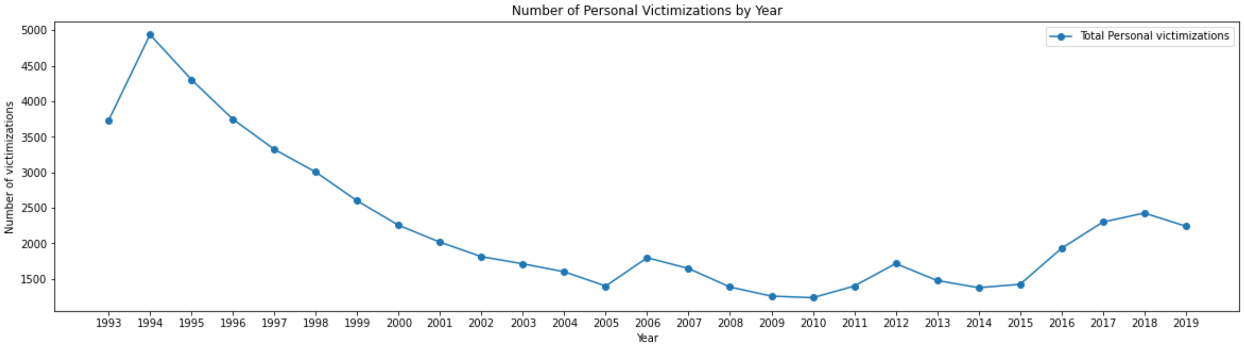
Conclusion



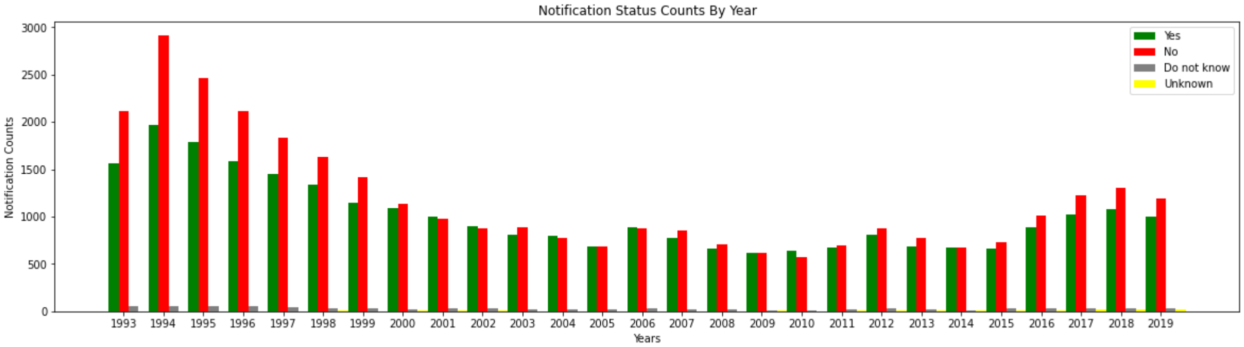
Random Forest is the model (among chosen models) best suited to predict whether a household is going to report a victimisation or not based on the given characteristics.

Implementation - Personal victimization:

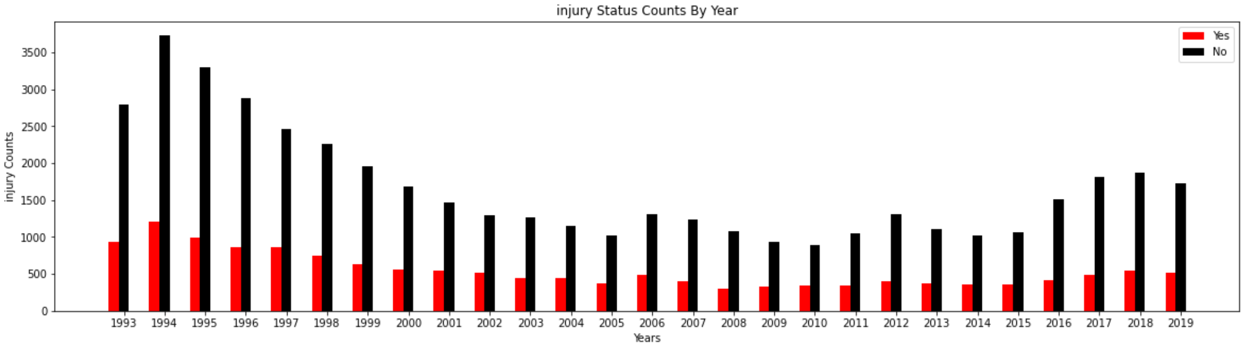
1. What is the number of victimization per year?



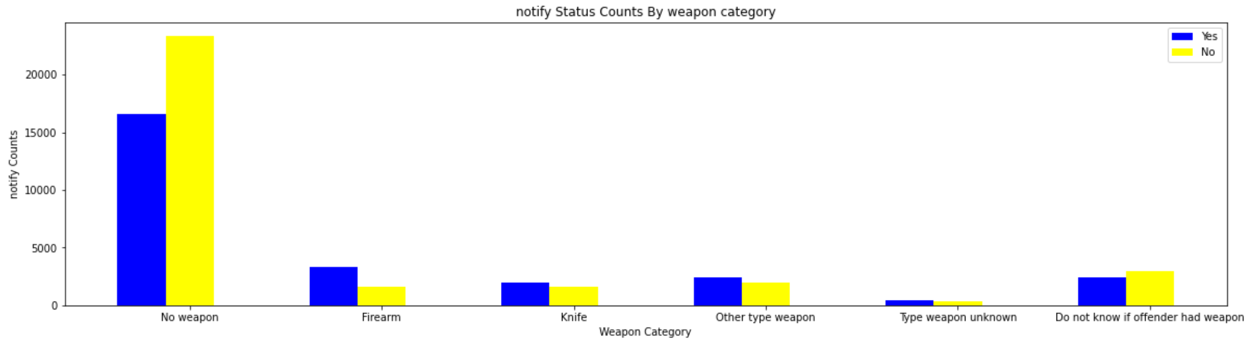
2. What are the reporting status (Notified, Not notified, Do not know, Unknown) counts through the Years



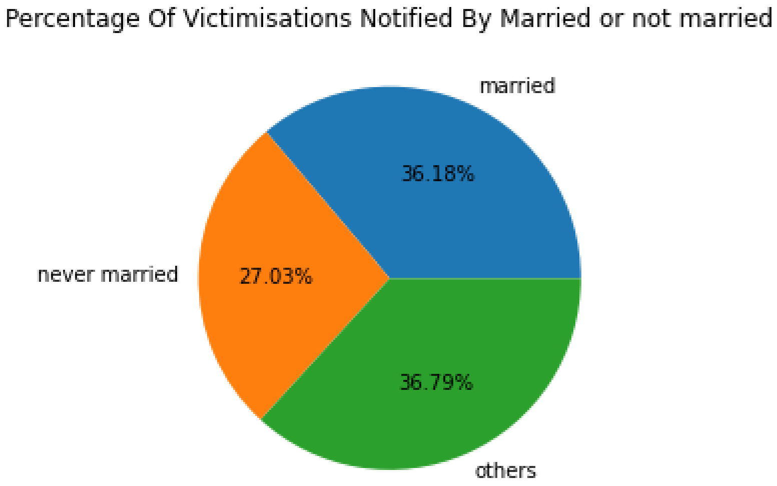
3. By what percentage are injury statuses higher than the notified through the years?



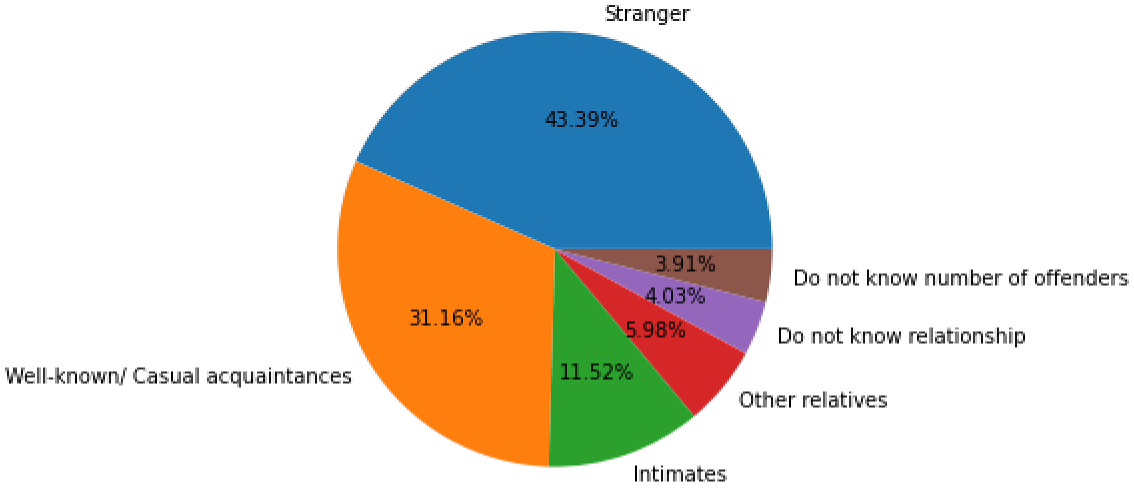
4. By what percentage are the weapon category report a victimization



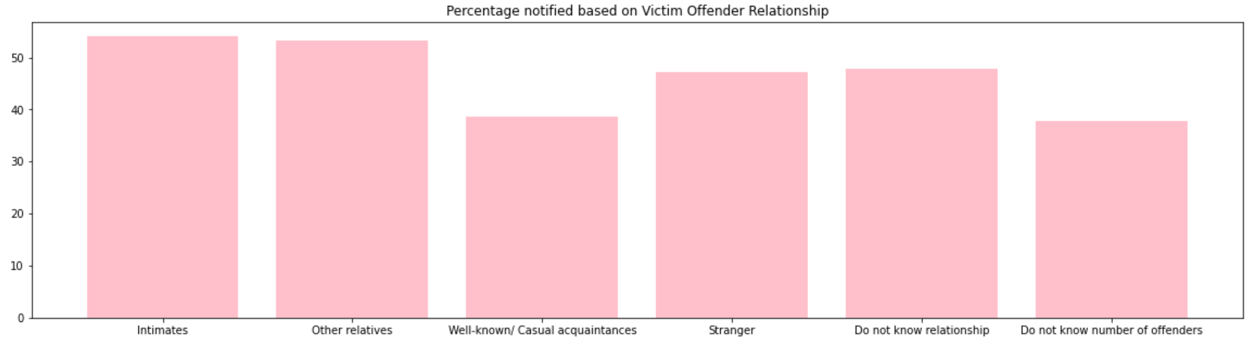
5. How does the marital status affect the decision to report a victimization



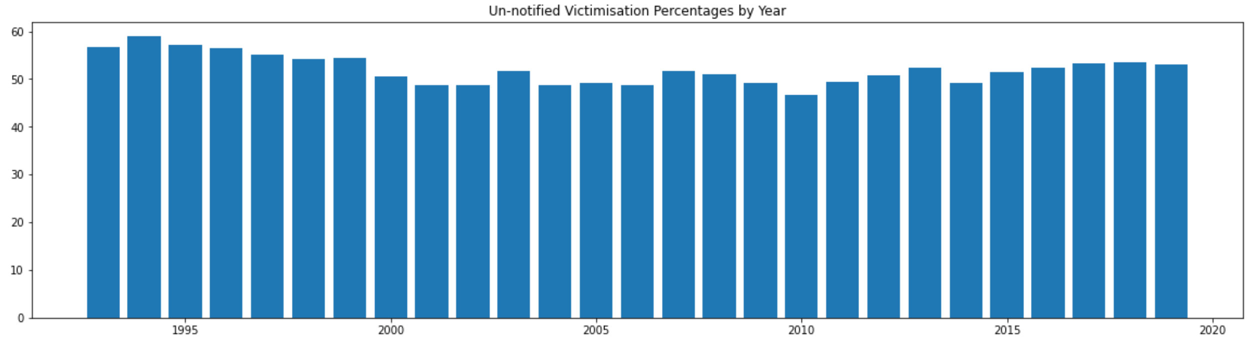
6. How does the victim-offender relationship affect whether the victimization gets reported or not?



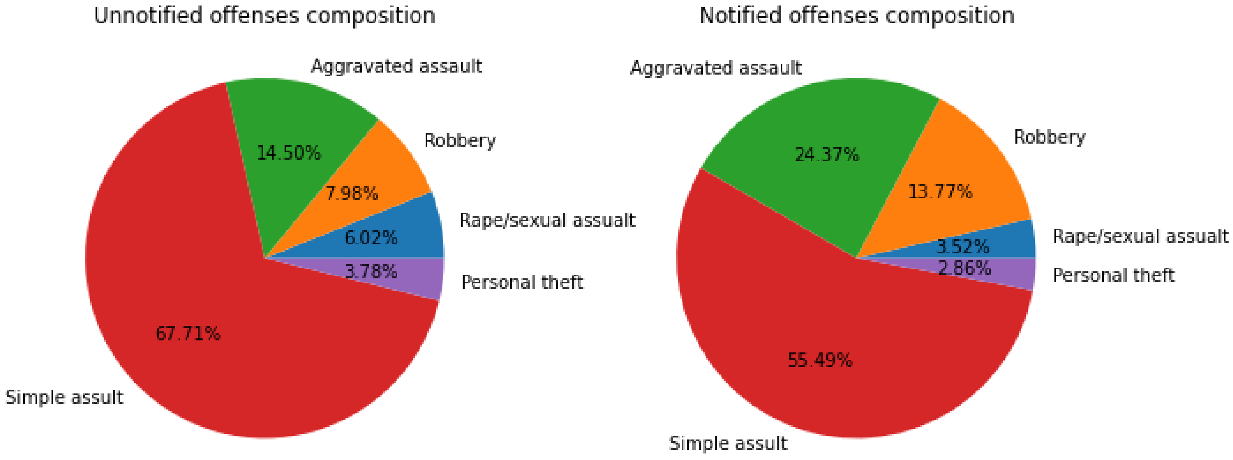
7. By what Percentage of victimizations reported grouped by offender relationship to victim



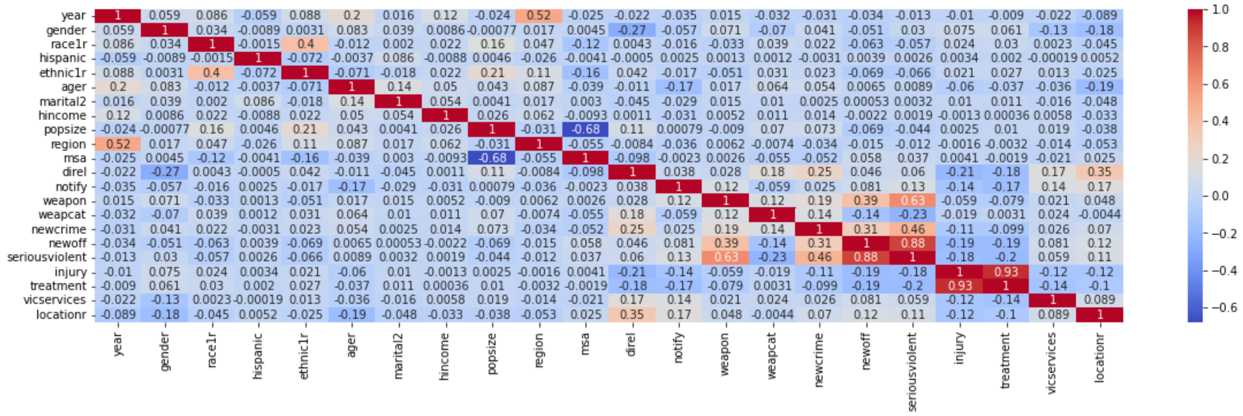
8. By what percentage are the un-notified victimisations higher than the notified through the years?



9. What type of crimes are most likely to be reported? Similarly, what type of crimes are most likely to go unnotified?



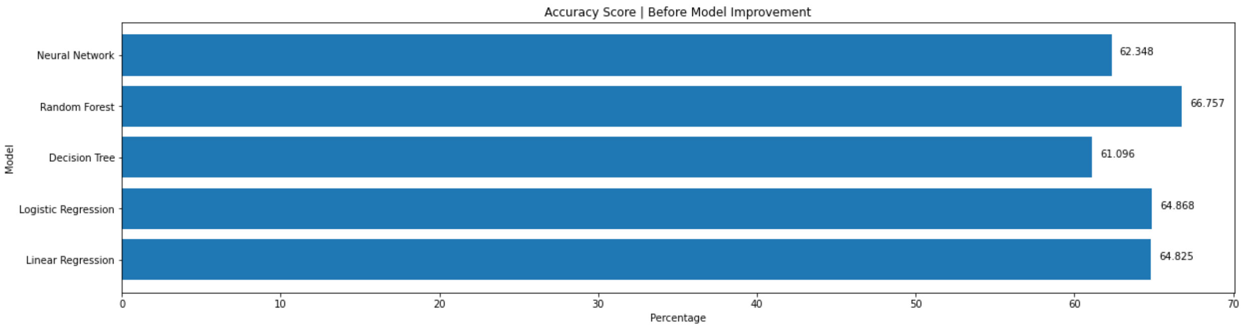
10. Check correlations of other columns with notify



Training Models:

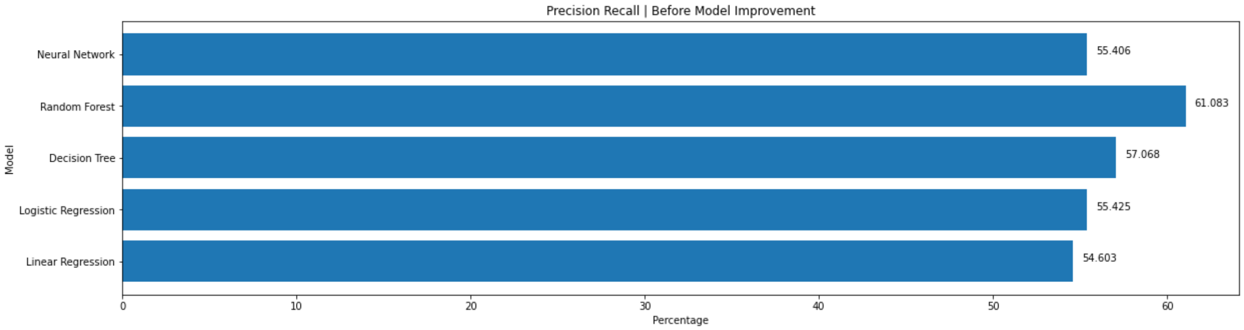
Model Evaluation - 1

1. Accuracy Score - Before data Improvement

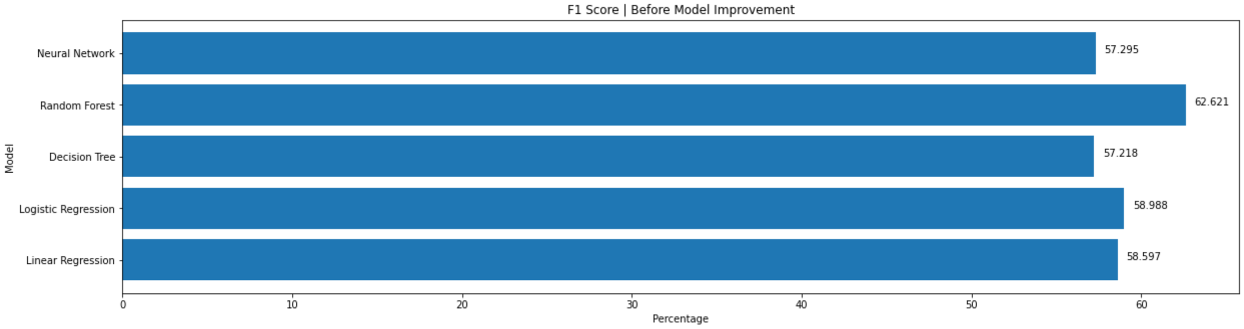


Random Forest has the highest accuracy score amongst the other 5 models - 66.757%

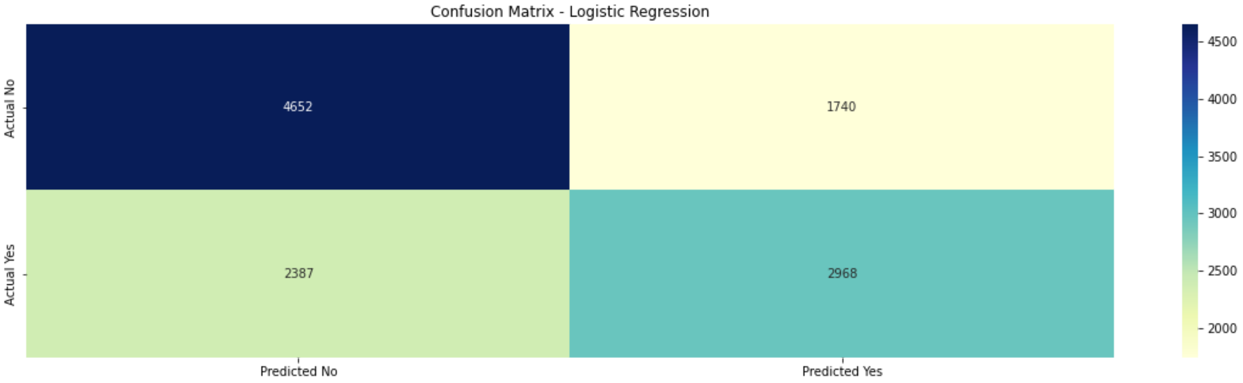
2. Precision recall Score - Before data Improvement



3. F1 Score - Before data Improvement

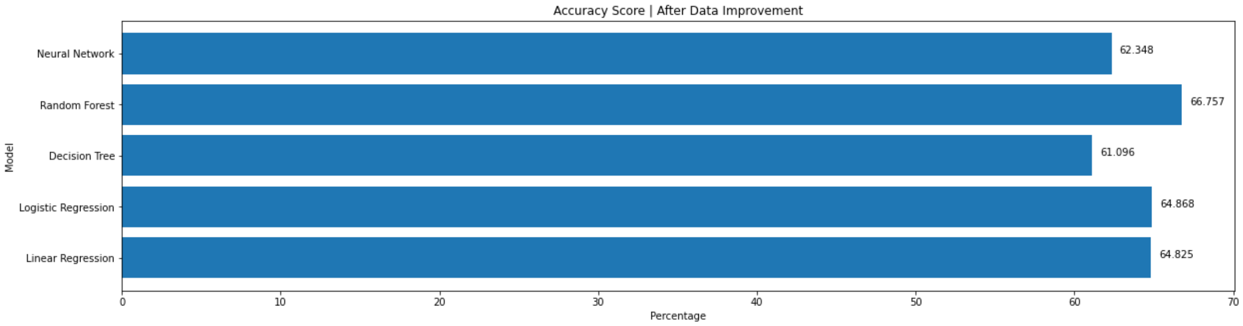


4. Confusion Matrix - Logistic Regression

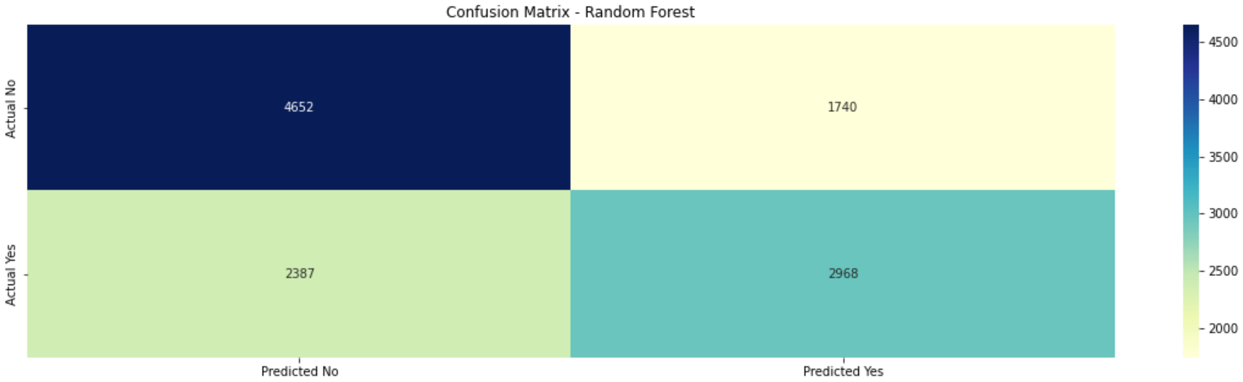


Model Evaluation - II

1. Accuracy Score - After Data Improvement



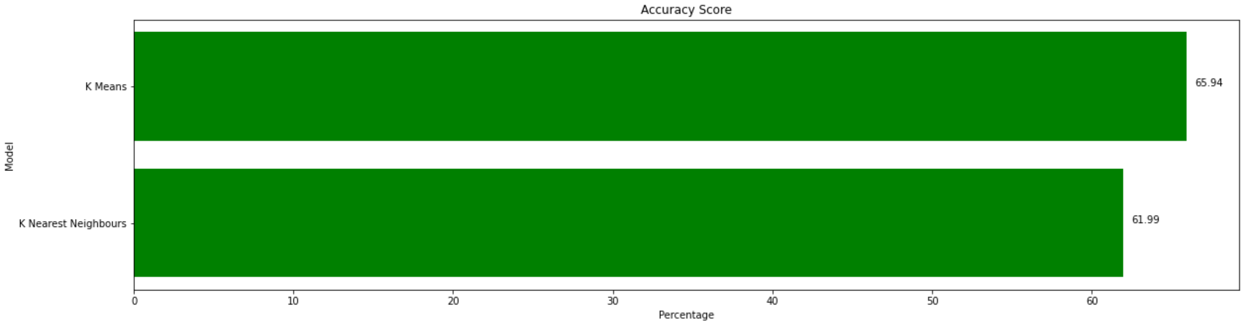
2. Confusion Matrix - Random Forest



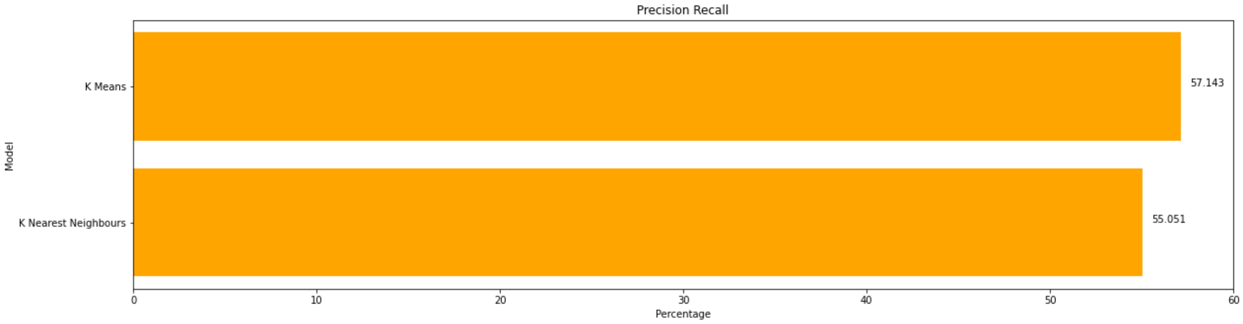
Model Evaluation - III

Here, we evaluate clustering algorithms

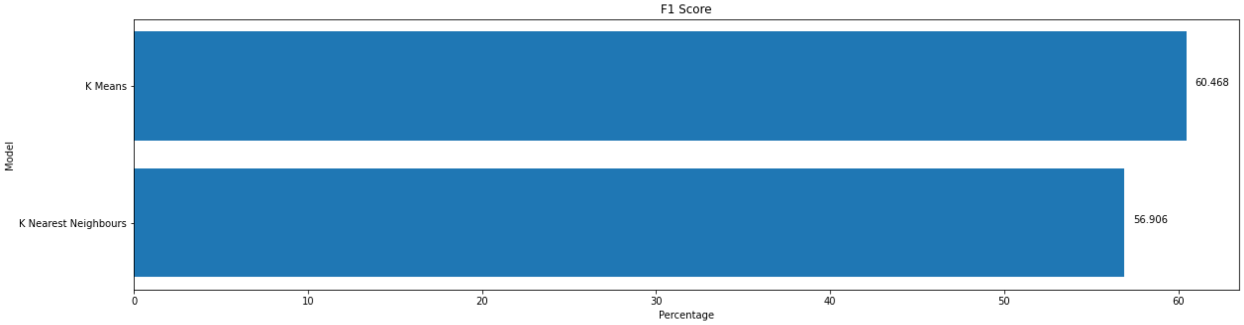
1. Accuracy Score



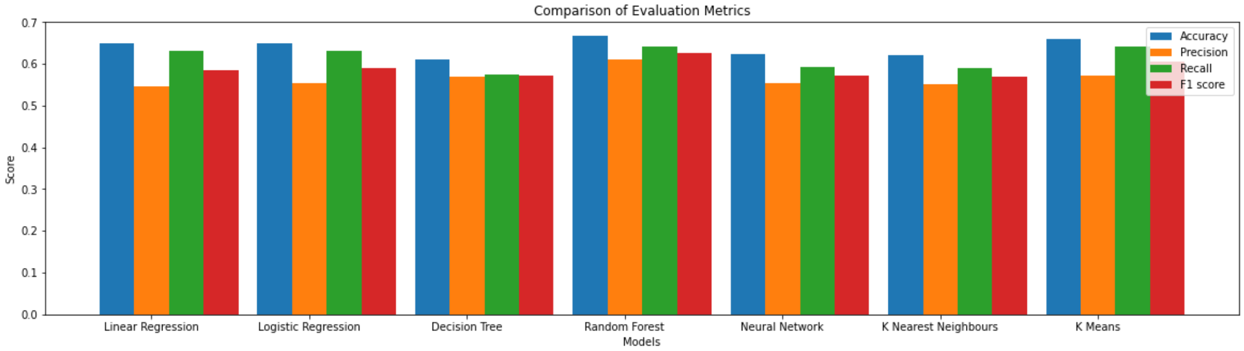
2. Precision-Recall



3. F1 Score



4. Comparison of Evaluation Metrics



Random forest is the model (among the chosen models) that is best suited for predicting whether a person is going to report victimization or not, based on the characteristics given.

4. Conclusion

Random Forest is the best model for prediction of reporting of victimization for both household and personal victimization datasets.

References:

https://worldpopulationreview.com/country-rankings/crime-rate-by-country
https://www.bjs.gov/developer/ncvs/index.cfm
https://www.bjs.gov/developer/ncvs/personalFields.cfm
https://www.bjs.gov/developer/ncvs/householdFields.cfm