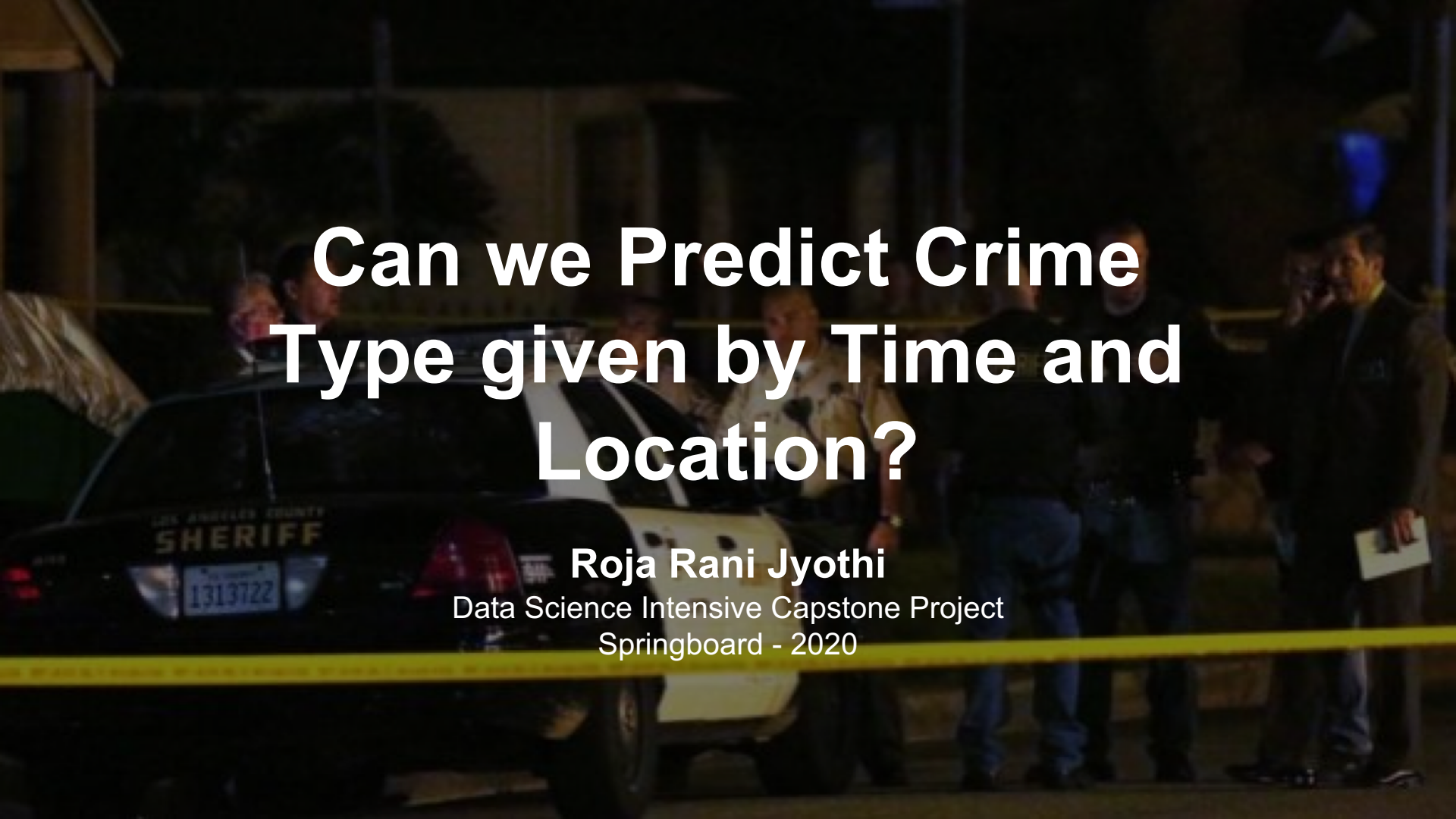


Can we Predict Crime Type given by Time and Location?

Roja Rani Jyothi

Data Science Intensive Capstone Project

Springboard - 2020

Context

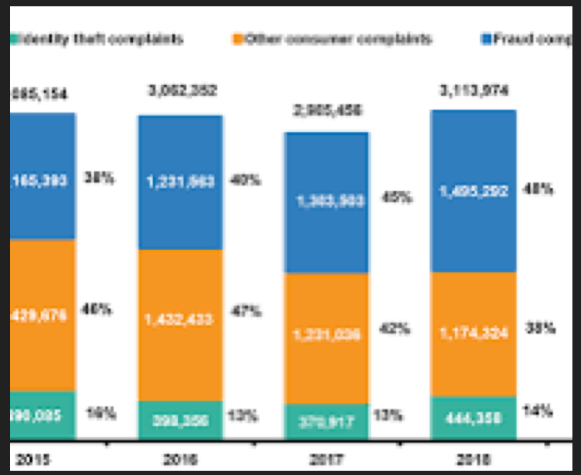
1. Los Angeles, California is one of the largest metropolitan areas in the country and has a colorful history of crime,including organized criminal activity, gang wars, riots and more.
2. The purpose of this project is to give insights to the authorities to concentrate their forces in order to contrast crimes and to assess the type of crime ahead.

Problem

- What type of crimes that frequently occur in LA neighborhoods?
- What factors contribute to the type of crime?
- Can we predict type of crime given by time and location?

Who might care?

- Police Department
- US Newspapers that cover crime news
- US online backed by data analytics to create awareness in the public



Data Information

- Data Acquired for the period : 2010 - 2019
  - Crime data: Los Angeles - Open Data Portal
    - Number of records: 2110000
    - Number of fields: 28

Data Wrangling

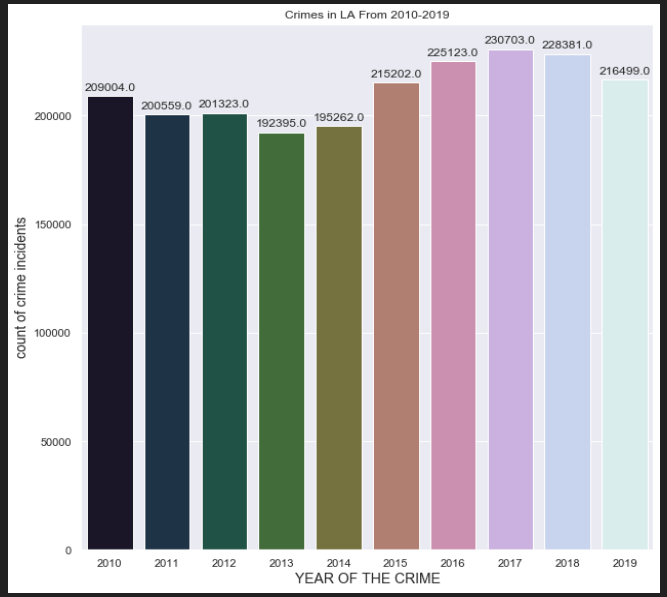
Data Cleaning:

1. Changing the original column names into more accessible format.
2. Dropping rows that contains NaN's.
3. Typecasting columns
4. Removing outliers
5. Extracting additional information from Time and Date Occurred such as day of week,,month,year,date,hour,minute.

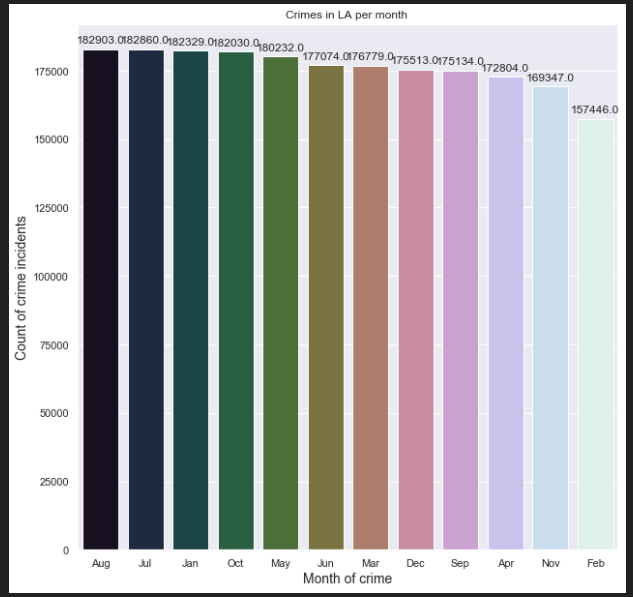
DR_NO	Date Rptd	DATE OCC	TIME OCC	AREA	AREA NAME	Rpt Dist No
1307355	02/20/2010 12:00:00 AM	02/20/2010 12:00:00 AM	1350	13	Newton	1385
11401303	09/13/2010 12:00:00 AM	09/12/2010 12:00:00 AM	45	14	Pacific	1485
70309629	08/09/2010 12:00:00 AM	08/09/2010 12:00:00 AM	1515	13	Newton	1324
90631215	01/05/2010 12:00:00 AM	01/05/2010 12:00:00 AM	150	6	Hollywood	646
100100501	01/03/2010 12:00:00 AM	01/02/2010 12:00:00 AM	2100	1	Central	176
100100506	01/05/2010 12:00:00 AM	01/04/2010 12:00:00 AM	1650	1	Central	162

Exploratory Data Analysis

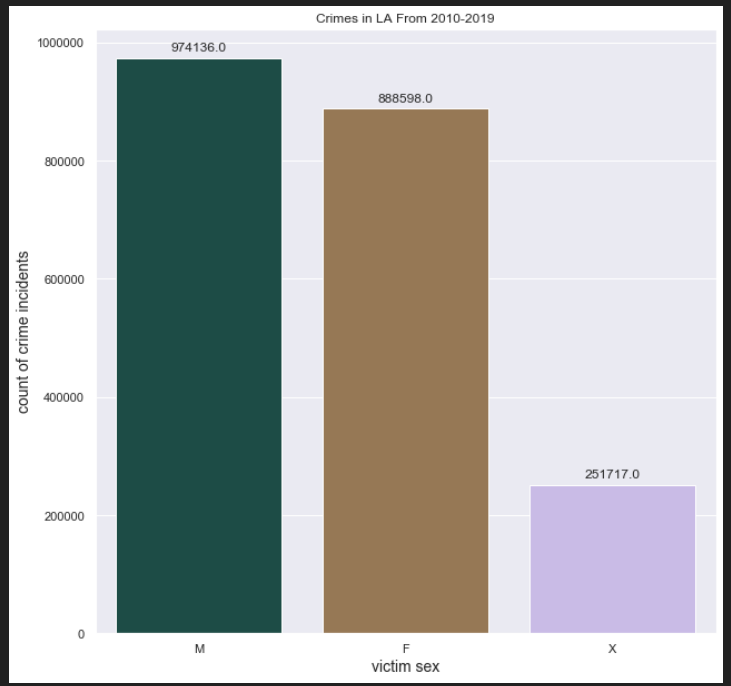
- Crime count on yearly basis



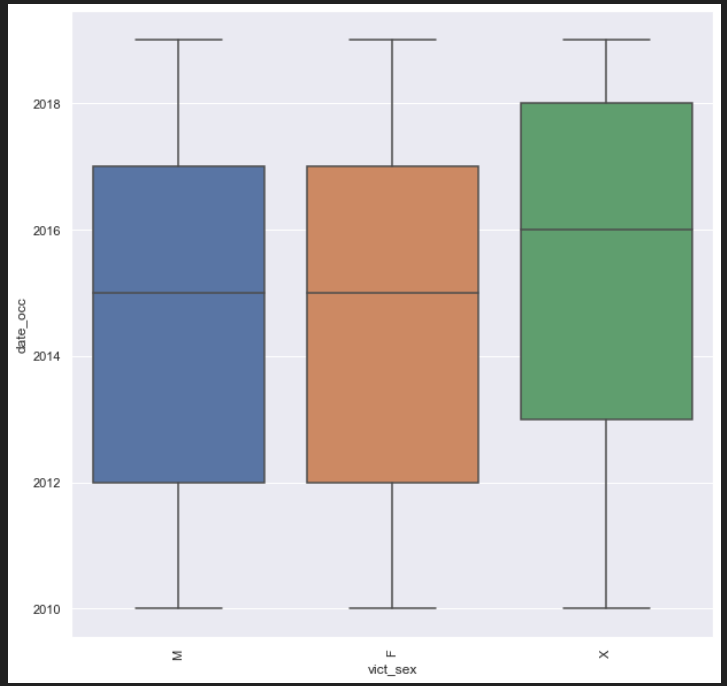
- Crime count on monthly basis



- Crime counts with respect to victim’s sex

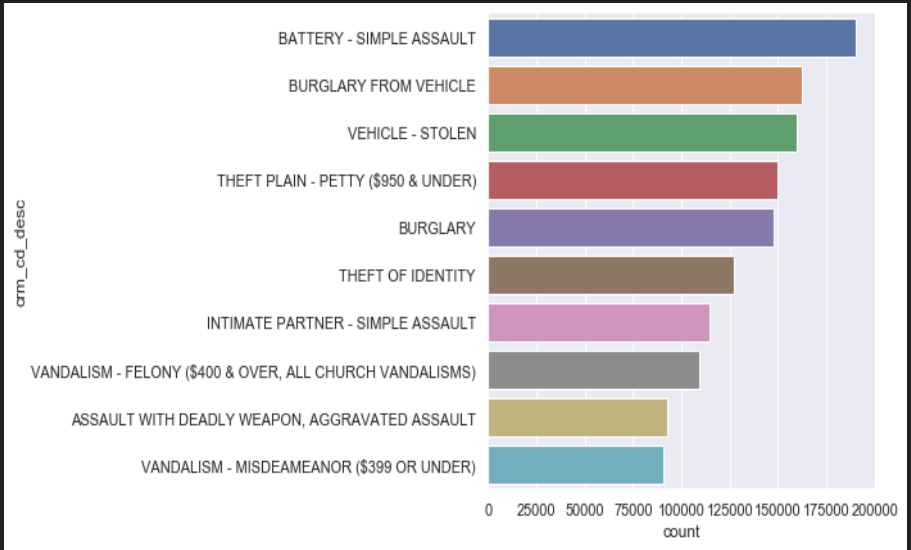


- Yearly crimes on victim’s sex

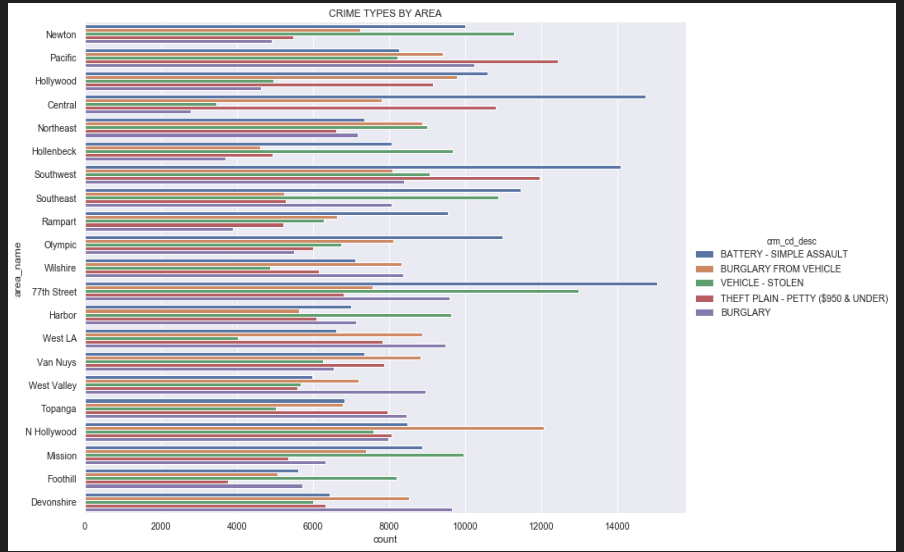


Crime types with respect to Area

- Top 10 crimes in LA



- Crime types by area

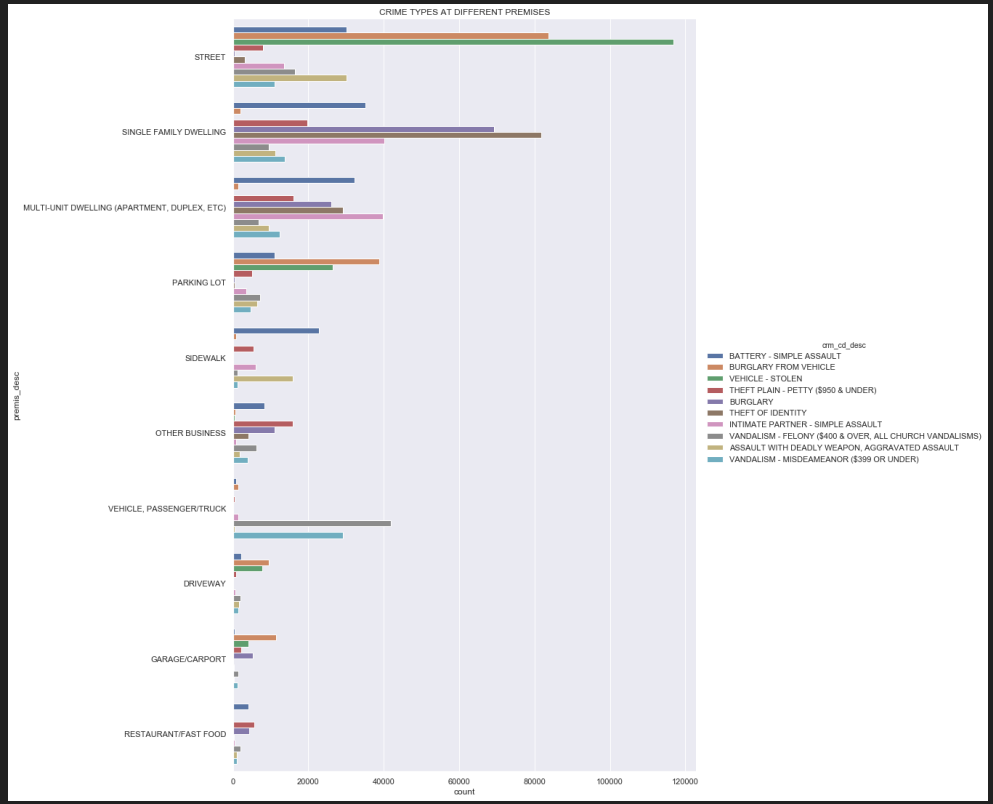


Hypothesis Testing: Is there any relationship between crime type and location

- After performing chi-squared test, we got p-value as 0.
- Results:

We concluded that there is a relation ship between crime type and location.

- Crimes at different locations



Machine Learning Modeling

We tried to build two models

1. Multiclass classification for all the crime types including low frequency crimes
2. Multiclass classification for High frequency (top 10) crimes.

Multiclass classification for all crime types including low frequency crimes

Modeling overview

Type: Supervised and unsupervised learning

Multi class classification

Highly imbalanced data

Tools: Python’s scikit learn and imblearn

Modeling steps

- Data pre-processing steps:

  1. Binary encoding

  2. Stratified Sampling (sample of 20000 out of the original dataset)

  3. Data splitting into training and test sets (80%-20%)

  4. Scaling

- Cross validation (CV) for hyperparameter tuning:
  - StratifiedKFold cv
  - Using scikit-learn random search method
  - Evaluation metric: F1-score(weighted)

Classifier training using optimal parameters and 80% of the whole data

Performance evaluation using holdout dataset (50% of the whole data)

Classification Algorithms Used

1. Logistic Regression
2. Support Vector Machine
3. Random Forest
4. K-Nearest Neighbours
5. XGBoost

Resampling/Weighting Techniques Used

Resampling techniques (imblearn):
SMOTETomek
Weighting technique (sklearn):
Using class_weight (='balanced')
parameter in several scikit-learn classifier implementations

Clustering Algorithms Used

DBSCAN

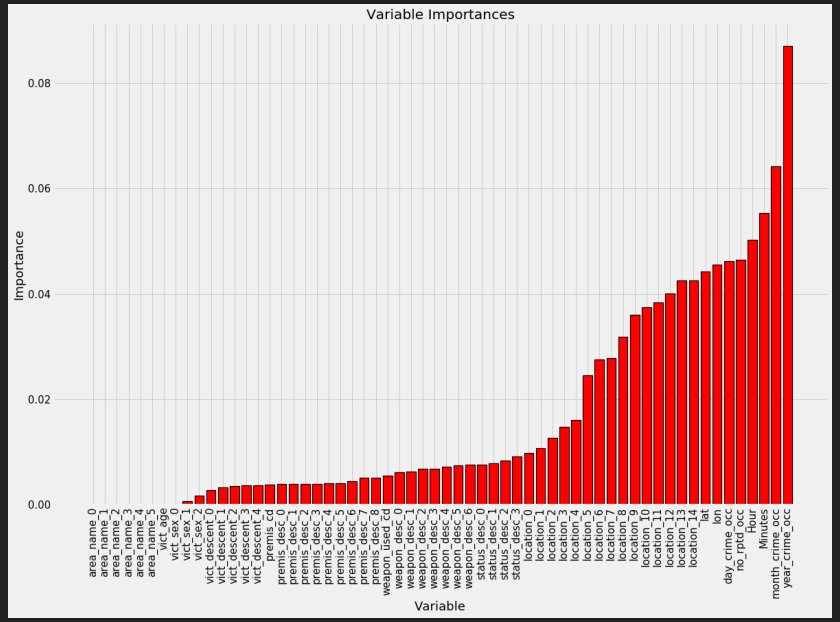
Choosing the best model: We choose the best model as Cost sensitivity(without clusters) even though SMOTETomek outperforms Cost sensitivity, because SMOTETomek is computationally so expensive.

Model	Algorithm	f1_score(weighted)	Precision	Recall	Training	Testing
Without Clusters	Random Forest	0.665	0.687	0.692	1.000	0.692
Without Clusters	Cost Sensitivity	0.684	0.712	0.705	1.000	0.705
Without Clusters	SMOTETomek	0.708	0.719	0.720	0.992	0.720
With Clusters	Random Forest	0.662	0.681	0.689	1.000	0.689
With Clusters	Cost Sensitivity	0.678	0.705	0.701	1.000	0.701
With Clusters	SMOTETomek	0.703	0.713	0.716	0.992	0.716

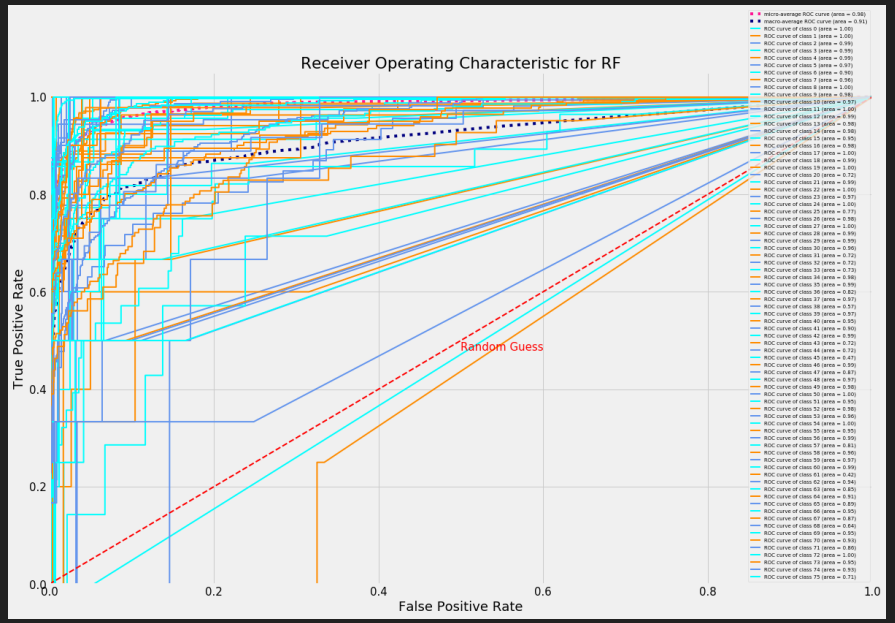
Reviewing the model performance :

After tuning the parameters, model achieved accuracy F1-score(weighted) as 0.71

- Feature Importance



- AUC-ROC for multiclass



MULTICLASS CLASSIFICATION FOR HIGH FREQUENCY(TOP 10) CRIMES

- Classification Algorithms Used

Support Vector Machine (SVM)
k-Nearest Neighbours (KNN)
Logistic Regression (LogReg)
Random Forest (RF)
XGBoost(XGB)

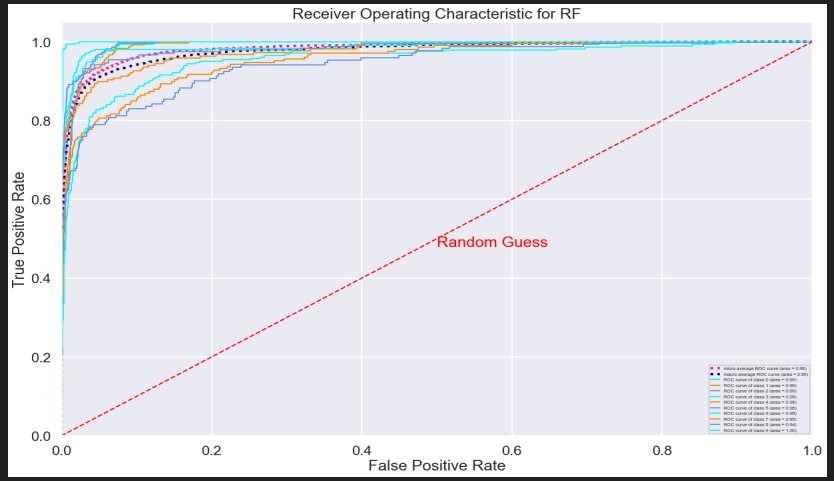
- Model Comparison

Model	Accuracy(mean)	Accuracy(std)
SVM	0.656	0.0041
KNN	0.562	0.0072
RF	0.692	0.0053
LogReg	0.604	0.0084
XGB	0.708	0.008

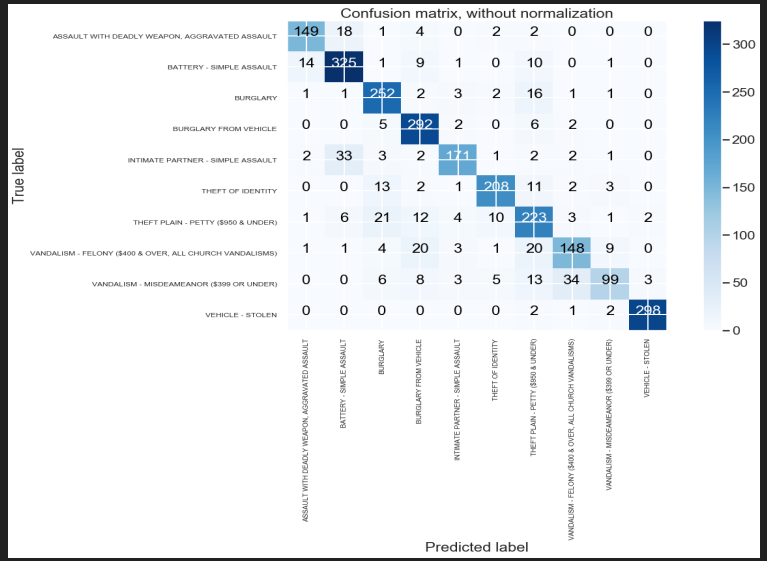
Reviewing the model performance :

After tuning the parameters, model achieved accuracy F1-score as 0.84

- AUC-ROC for multiclass



- Confusion matrix



Conclusion

- We built two models to predict
    1. All the crimes including low frequency and
    2. High Frequency (top 10) crimes.
- Out of all models we tried, we choose the Random Forest(cost sensitivity) classifier as the final model.
- With 80%-20% splitting, the test data set gave F1-score(weighted) = 0.71 for all the crimes and F1-score = 0.84 for high frequency crimes.
- With more ideas, the model can be improved in the future.

Future Improvements

- Could incorporate other information like Modus Operandi(MO) , events information and weather information to see whether these features can help the classification.
- We could use a Clustering algorithm, namely K-Modes Clustering, to figure out some *cluster centroids* which can potentially be interesting for authorities to contrast crimes indicated by centroids and similar ones.
- Due to memory constraints on Jupyter notebook, I had to train a sample of size 20000 from the original dataset. Without resource limitations, I would love to train on the full dataset.

THANK YOU!