Crime Classification in San Francisco

"When there is a crime in society, there is no justice!" - Plato

Abhishek Jaiswal
A20380004
ajaiswal1@hawk.iit.edu

***Abstract*—The main objective of this project is to implement Big Data technologies in the machine learning realm. As part of this project, we will be working on the San Francisco Crime Classification data set obtained from Kaggle. We are mainly interested in developing a system that could classify crime descriptions into different categories which would help the authorities to assign officers to crimes based on the report. There can be numerous approaches to solving this problem. Out of all those approaches we will be using the crime dataset and working around it. We will be training a model based on the 39 predefined categories, testing the model's accuracy, and deploying it into production. Given a new crime description, the system should assign it to one of the 39 categories. In addition to that, to solve this multi-class text classification problem, we will use various feature extraction techniques along with different supervised machine learning algorithms in Pyspark.**

***Index Terms*—Big Data Technologies (ie- PySpark), Machine Learning, Python Libraries - Keras, SciKit-Learn, Seaborn, MatplotLib, NumPy, Pandas, etc.**

I. Introduction

Everybody has observed crimes, whether they were perpetrators or victims. Therefore, we can say- *"crime is an integral part of our society"*. That is why we are interested in building a crime classification system that could classify crime descriptions into different categories. A system that can help law enforcement assign the right officers to a crime or automatically assign officers to a crime based on the classification.

In our project, we analyzed the crime data that we selected from the "San Francisco Police Department (SFPD) Crime Incident Reporting System" which has the incidents of crimes in San Francisco city from 1/1/2003 to 5/13/015. So, an obvious question that can arise here is- *"Why San Francisco?"*

California's San Francisco serves as the state's administrative, financial, and cultural hub. It is the 17th most populous city in the US and a popular tourist destination known for its cool summers, the Golden Gate Bridge, and some of the best restaurants in the world. San Francisco is a city known for its expansion and liveliness, but because of a rise in criminal and illegal activities, it is still one of the most dangerous places to live in the US.

As per the Hoover Institution- *"San Franciscans face about a 1-in-16 chance each year of being a victim of property or violent crime, which makes the city more dangerous than 98 percent of US cities, both small and large."* On top of that, as per an independent study by Sfgate suggests- *"San Francisco is the nearly the most crime-ridden city in the US."* Therefore, after closely analyzing all the available data and reports we have chosen San Francisco.

II. Problem Definition

We defined a few questions to get a sense of the security conditions in San Francisco, and we responded to them during our project- "Crime Classification in San Francisco (implement Big Data technologies in the machine learning realm)".

1) How has the number of various crimes changed over time (years / months / weeks / hours / minute) in San Francisco?
2) Are there any trends in the crimes being committed over the trend?
3) What is the specific location, time, and year for a specific crime?
4) Which regions are the locations where these crimes are oftenly committed?
5) Which crime (ie- theft) has highest number of occurrences over the years in San Francisco?

III. Data Pre-Processing: Data Extraction

A. Data Exploration

Our data is driven from Kaggle website, and this dataset reflects reported incidents of crime with exceptions of murders where data exists for each victim and these crimes were occurred in San Francisco from 2001 to 2017. Data is derived from CLEAR (Citizen Law Enforcement Analysis Reporting) system. In compliance with privacy act of California state law, personal details like addresses, specific locations are shown as block to not identify them.
This data includes information like date time of reported crime, block where crime happened, type of crime and location descriptions, was there an arrest. Moving further we bring more precise information about our data such as size of our data, sneak peak into our data, etc.

B. Data Extraction

We have various list of available tools for data preprocessing, some of them are Stanford Visualization Group's Data Wrangler, Python pandas, PySpark, Spark Datafrmae and

these tools are great in working and can save hours of work. PySpark is incredibly easy and convenient we used this tool to gain an insight into our data and we found that based on our project for Crime Classification in San Francisco requires data cleaning the most.

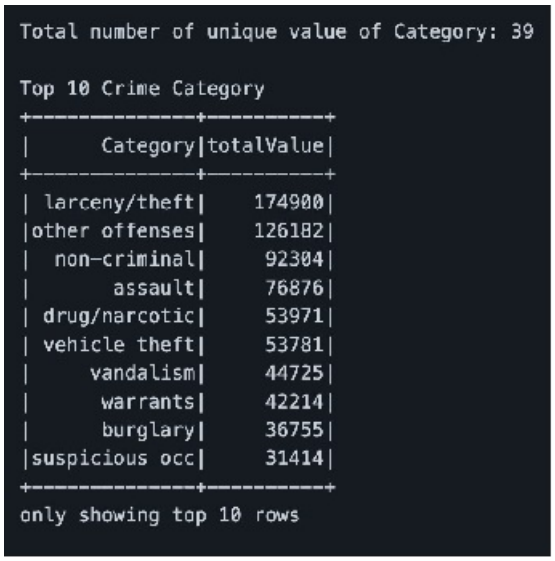
Here is a sample of - total number of unique value of description-

```
Total number of unique value of Description: 879

Top 10 Crime Description
+-----------------+--------+
|        Description|totalValue|
+-----------------+--------+
|grand theft from ...|     60022|
|       lost property|     31729|
|             battery|     27441|
|   stolen automobile|     26897|
|drivers license, ...|     26839|
|      warrant arrest|     23754|
|suspicious occurr...|     21891|
|aided case, menta...|     21497|
|petty theft from ...|     19771|
|malicious mischie...|     17789|
+-----------------+--------+
only showing top 10 rows
```

Fig. 1. Crime Description

Here is a sample of - total number of unique value of category-



```
Total number of unique value of Category: 39

Top 10 Crime Category
+-------------+---------+
|      Category|totalValue|
+-------------+---------+
| larceny/theft|    174900|
|other offenses|    126182|
|  non-criminal|     92304|
|       assault|     76876|
| drug/narcotic|     53971|
| vehicle theft|     53781|
|     vandalism|     44725|
|      warrants|     42214|
|      burglary|     36755|
|suspicious occ|     31414|
+-------------+---------+
only showing top 10 rows
```

Fig. 2. Crime Category

For getting this data we performed the below set up steps-

- Processing of duplicate records
- Handling missing values (Null/NA)
- Keeping only relevant data features for our model and we used heat map chart for the correlation understanding among various features.

IV. Analysis

A. Choosing A Technology

It is important that such machine leaning projects of data analytics get implemented using technologies which work best with respect to data as use of right technologies would be a game changer in the efficiency, speed along the cost of the process. Taking this into account, scope features of this project we concluded to use the Spark would be better.

Additionally, we tried to compare the various parameters like speed, performance, memory optimization type of machine learning algorithm used, type of data format used among others. After close analysis we observed that Spark is better for our project than TensorFlow and H20.

Our detailed analysis is presented below.

We had a challenge of deciding between MapReduce and Spark technologies as these 2 frameworks have their own characteristics benefits and they both have their own nature and proficiency for big data processing. The key difference between them lies in the approach to processing while Spark can do it in-memory, Hadoop MapReduce must read from write to a disk, and the speed of processing differs significantly. Based on the needs of this project Spark in-memory computation is much faster and more efficient so we are using it as our main framework.

Spark vs TensorFlow - Our project is based on classifying and processing San Francisco's crime record data. In addition to that, there was no element of custom model training and deep learning. Whereas TensorFlow is most used for custom learning and Neural Network Design. That is why we have used Spark in place of TensorFlow in our project.

Spark vs H2O - In our project, we are using a few supervised machine learning algorithms, for instance- Regression. While H2O is a great fit for random forest and Gradient Boosting Machines (GBMs). Additionally, H2O works efficiently with some specific data formats like-. hex. On the other hand, Spark uses the most general and common type of data format, i.e.- relational databases.

Conclusion- Therefore, we have used Spark in place of H2O even though we know H2O is faster than Spark.

B. Analysis and Result

As explained in the previous part we use spark SQL to answer each of the questions we defined as our project's goal in section 1. In continuing we explain one by one how we answered each of those questions, in other words, we would be looking into the various insights of data and data samples.

V. Process Workflow

To begin with, we imported and set up Spark and other libraries like- Seaborn, Matplotlib, and NumPy that are required in this project. Thereafter, we extracted the data and selected variables through feature selection. Thereon, the third crucial step that we followed was based on partitioning the dataset into the training and testing datasets.

Some of the important steps involved in this process were defining -

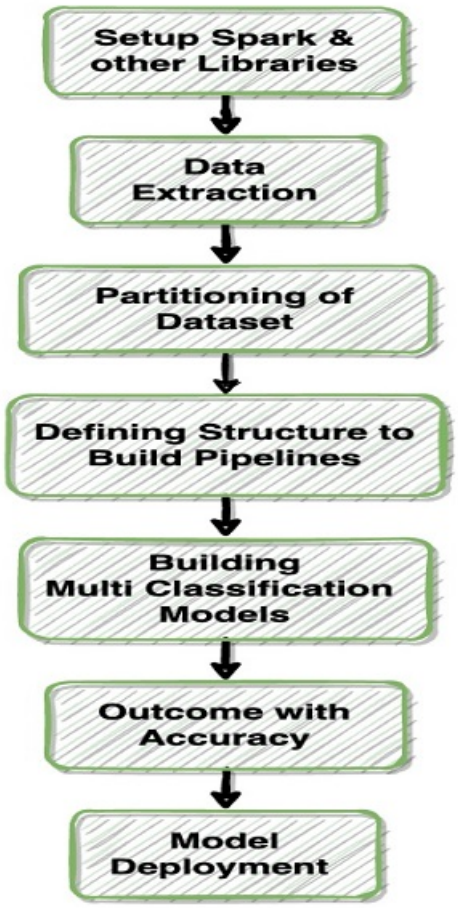


Fig. 3. WORKFLOW

The count that we got for both datasets was - 6,15,417 and 2,62,632 respectively. These datasets would now help in training the models for making the right classifications which are the backbone of the whole project.

Thereafter, in the next part, we defined various structures for building pipelines for our project.

1) Tokenization function using RegexTokenization
2) Stop remover function using StopWordsRemover
3) Bag of words functions for DeScript variable using CountVectorizer
4) Function to encode the values of category variable using StringIndexer.
5) Pipeline to call these functions.

Furthermore, we moved on to the Building multi-classification models step, which includes the building baseline model and secondary model based on the supervised machine learning approaches. Some of the processes that might get involved would be -

Model training and evaluation -

1 Building a Baseline Model -
   - Logistic Regression using CountVectorizer features

2 Build a Secondary Model -
   - Naive Bayes
   - Logistic Regression Naive Bayes using TF-IDF features
   - Logistic Regression Naive Bayes using word2Vec

Therefore, we will try to classify the crime rate based on two supervised machine learning algorithms and it's accuracy will decide which would be the better model use.

VI. Data visualization

Data Visualization is one of the crucial aspect of Data Science and analysis. Therefore, we tried to use some of the crucial graphs like- Bar chart, column chart, among others. As we get to the analysis and result by using Spark, we get to see lot of information easily using the different visualizations and will try to solve some of the problem statements that we mentioned above-

A. How the features are correlated?

The below Heat map chart of San Francisco tells us how the various features from our dataset are correlated to each other and we can make decisions based on the heatmap as well the strongly related features.

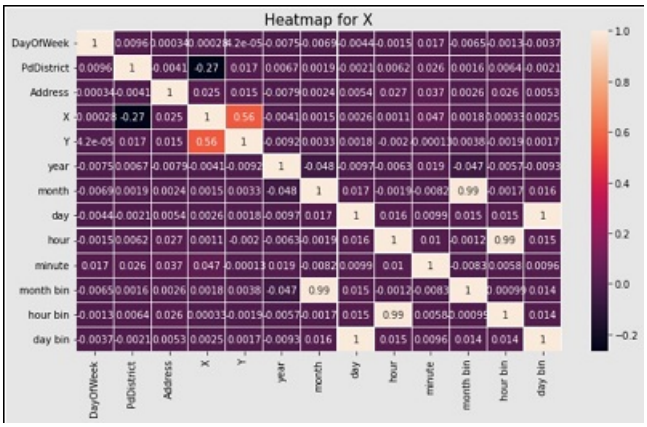


Fig. 4. Heat Map

For example- for which particular region highest number of crimes are recorded? As we can see from the heat map, that region(district) at 0.27 location has the highest number of recorded crime rate.

B. How has the number of crimes changed over time (years / months / weeks / hours) in San Francisco?

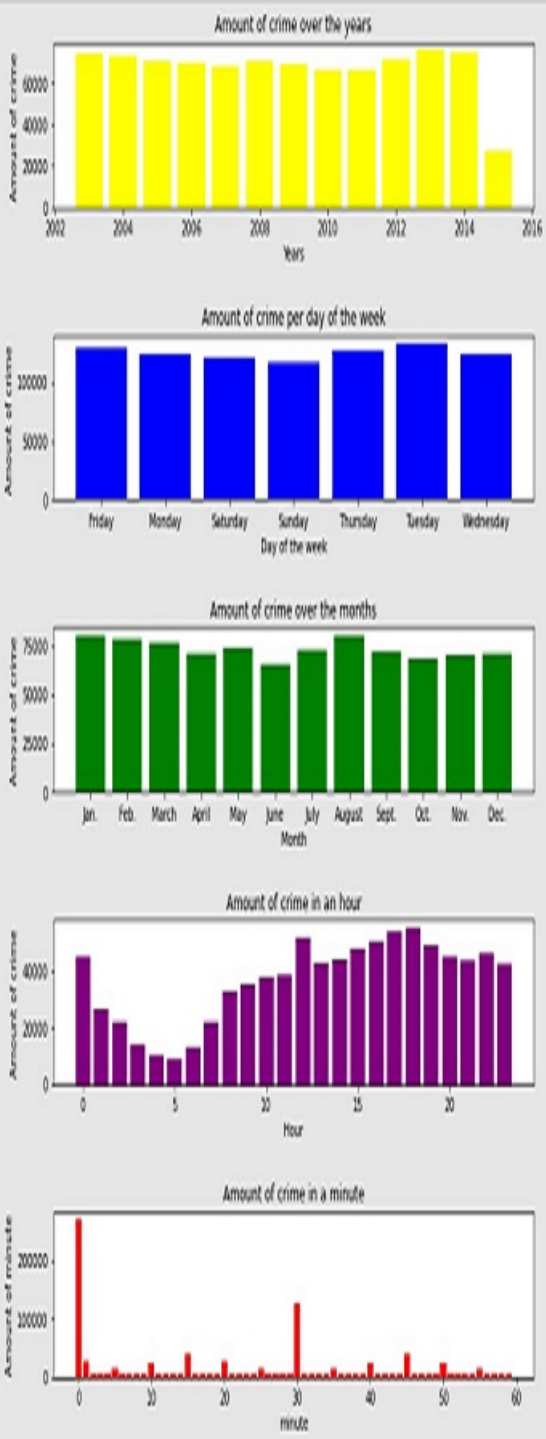


Fig. 5. Crime over Time

From the problem statement- we can see the crime in San Francisco is sort of constant over period of time. There are hardly any changes in it. For instance- if we take the third scenario of months, then we can see the crime rate has some vertical shift but most often it's same (approx 75000). Thus, this chart tells us the sad story of crime in the San Francisco.

Likewise some other insights can also be drawn from that single visualization. For example- Which particular crime is showing a significant growth over the years? For this problem statement we need to club some other graphs as well. We will see some of them later in this report.

C. What are the types of crime? and which particular crime has highest number of repetition or registration rate?

Using bar chart, we are easily able to understand that what are types of crime that are being most reported in San Francisco. For example- theft is one of the crime which has the highest number of occurrences over every region of San Francisco.

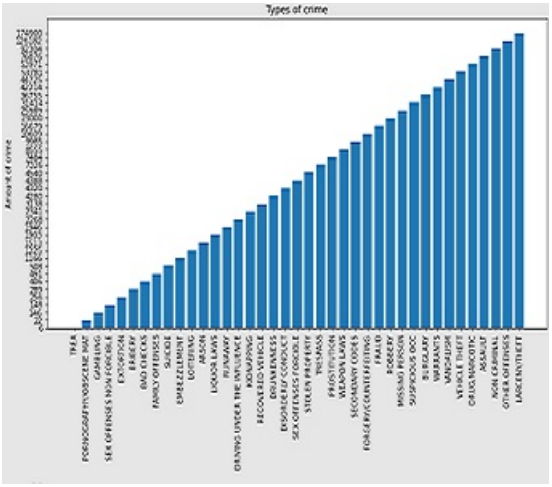


Fig. 6. Column Chart

D. What is the most reported crime across San Francisco?

When we closely look at this scatter plot, we observe that there is a varied amount of density (occurrences) as well as repetition of several crimes. For instance- Vehicle theft is mostly occurring in the outskirts of the city, while bribery is more towards the inner part of the city. Similarly, the crimes like - assault, burglary, kidnapping, prostitution, drug/narcotics, vandalism, etc. are reported mostly across the core area of the city. On the other hand crimes like- Bad checks, extortion, bribery, bad checks, arson, suicide, embezzlement, among others are those crimes which are happening at the least. However, the analysis from the other graphs suggests that there can be some sort of variation based on the location, time of the year, etc.

E. Are there any specific high crime locations across the San Francisco?

As we can observe the northern part of the city as well the central business district (CBD) has the highest number of registered crime rate. Whereas when we move more towards the western part of the city the crime rate decreases significantly.

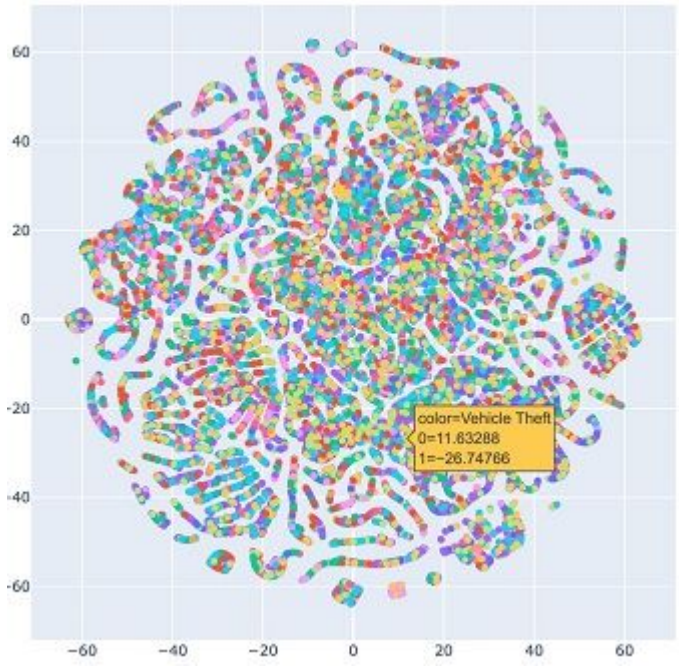


Fig. 7. Scatter Plot

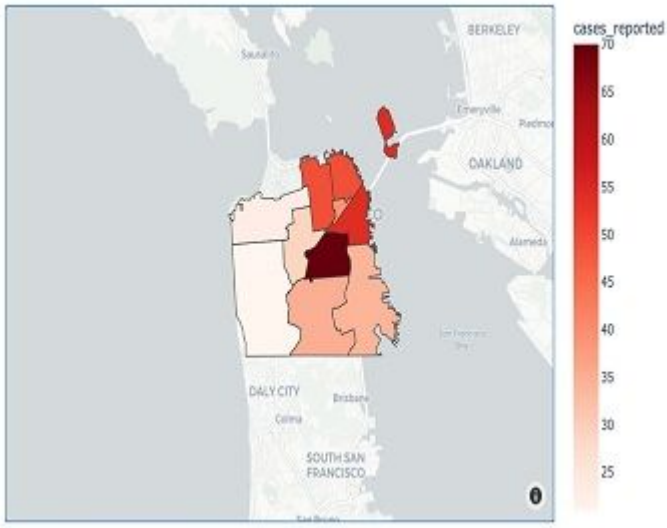


Fig. 8. Geo Map

functioning of the model.

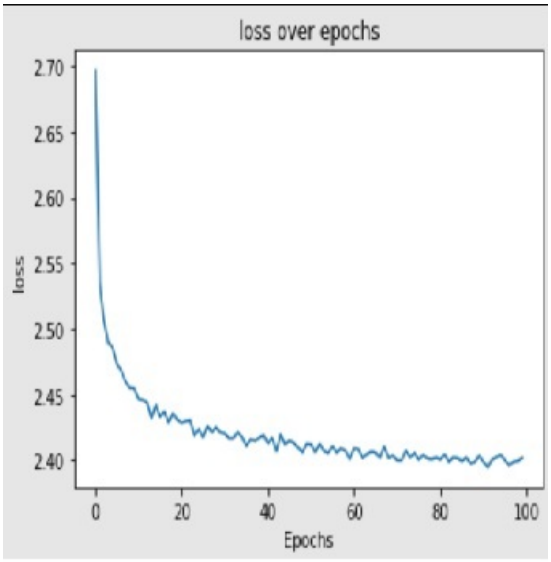


Fig. 9. Prediction Accuracy

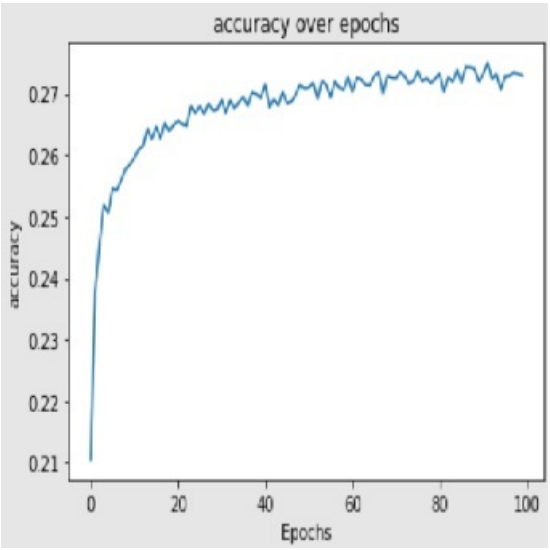


Fig. 10. Prediction Accuracy

VII. Model Analysis and Accuracy

A. Model Performance Analysis

Below we are showing a set of epochs purposely. Since the loss function is calculated across all data items during an epoch and is guaranteed to give the quantitative loss measure at the given epoch. However, plotting the curve across iterations only provides the loss on a subset of the total data set.

Therefore, plotting validation loss alongside training loss provides more context not just to the model but also to the

Furthermore, we are providing the model accuracy to give more insights about out machine learning model.

B. Model Accuracy

After a critical analysis of crime record data we got to know that- Naive Bayes is more accurate than the logistic regression for our crime record data set. This is because logistic regression makes the prediction on the basis connected functional form. Whereas, the Naive Bayes uses the throughput of the provided data.

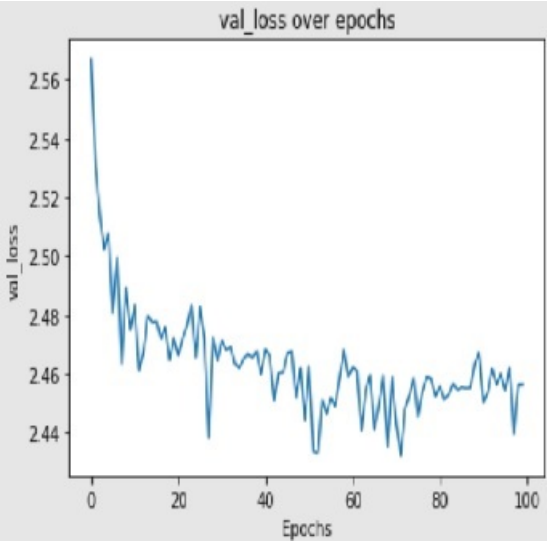


Fig. 11. Prediction Accuracy

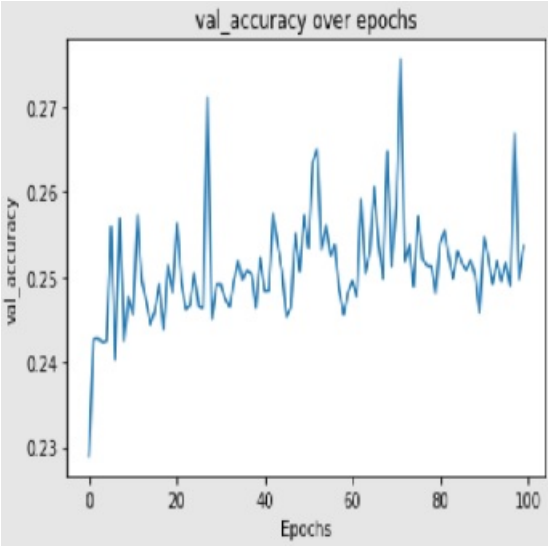


Fig. 12. Prediction Accuracy

The summary table about the accuracy of the model states the same-

	Logistic Regression	Naive Bayes
Count Vectoriser	77.2%	87.5%
TF-IDF	87.2%	90.7%
Word2Vec	89.7%	

Fig. 13. Model Accuracy

VIII. Conclusion

This machine learning project gave us a scientific view about the crime status and its rate of change for the San Francisco. As per the big data analysis report, we can observe the frequently occurring crimes along with the locations where they occurred. For example- Northern and the Central Business District (CBD) are the worse affected regions. In addition to these, the reports suggests that- the most reported crime in San Francisco was theft, even in the theft the vehicle theft crime was reported the most number of times.

Moreover, if we go by the detailed descriptive analytics then we noticed that the most occurred crimes except theft were - drug/narcotics, assault, vandalism, and most crime affected regions except the northern region were - eastern, and central part of the main city. Whereas western parts and south-western parts reported lesser no of crimes. The most commons places for these crimes were - streets, sidewalks, residence apartments. These are the locations where people are mostly moving and residing. One another interesting outcome suggests that - even though the rate of vehicle theft is around heavy 90 percentage, the vehicle recovery rate is still far below 30 percentage. This clearly suggests that San Francisco police department is taking ineffective measures for their investigation. But this problem can be solved if our analysis can be used. For example- if the police department needs to stop the vehicle then they need to keep an active patrolling around mid-eastern region.

At last, though we can not be very sure that our model training and testing and it's predictions are very accurate. This is because of limitation of resources for this project were in limited scope and we could not go up to certain miles to develop this project that can give us more accurate information and results.

Data Science Humour of The Semester

"Data Science Humour (off Semester) - We are not slow, we are just a gradient descent function with a very (s)low learning rate."

References

[1] C. -H. Yu, M. W. Ward, M. Morabito and W. Ding, "Crime Forecasting Using Data Mining Techniques," 2011 IEEE 11th International Conference on Data Mining Workshops, 2011, pp. 779-786, doi: 10.1109/ICDMW.2011.56.

[2] Feng, M., Zheng, J., Han, Y., Ren, J., Liu, Q. (2018). "Big Data Analytics and Mining for Crime Data Analysis, Visualization and Prediction". In: , et al. Advances in Brain Inspired Cognitive Systems. BICS 2018. Lecture Notes in Computer Science(), vol 10989.

[3] U. Thongsatapornwatana, "A survey of data mining techniques for analyzing crime patterns," 2016 Second Asian Conference on Defence Technology (ACDT), 2016, pp. 123-128, doi: 10.1109/ACDT.2016.7437655.

[4] GK Bottomley and C Coleman, "Understanding Crime Rates - Police and Public Roles in the Production of Official Statistics," NCJ Number 79264, United Kingdom, 1981.

[5] Jia Wang, Jun Hu, Shifei Shen, Jun Zhuang, Shunjiang Ni, "Crime risk analysis through big data algorithm with urban metrics", Physica A: Statistical Mechanics and its Applications, Volume 545, 2020.

[6] J. Hu, "Big Data Analysis of Criminal Investigations," 2018 5th International Conference on Systems and Informatics (ICSAI), 2018, pp. 649-654, doi: 10.1109/ICSAI.2018.8599305.

[7] M. I. Pramanik, W. Zhang, R. Y. K. Lau and C. Li, "A Framework for Criminal Network Analysis Using Big Data," 2016 IEEE 13th International Conference on e-Business Engineering (ICEBE), 2016, pp. 17-23, doi: 10.1109/ICEBE.2016.015.

[8] Farsi, M., Daneshkhah, A., Far, A.H., Chatrabgoun, O., Montasari, R. (2018). Crime Data Mining, Threat Analysis and Prediction. In: Jahankhani, H. (eds) Cyber Criminology. Advanced Sciences and Technologies for Security Applications. Springer, Cham.

[9] Matthias Hoffmann, Felipe G. Santos, Christina Neumayer, Dan Mercea. (2022) "Lifting the Veil on the Use of Big Data News Repositories: A Documentation and Critical Discussion of A Protest Event Analysis". Communication Methods and Measures 16:4, pages 283-302.

[10] Wu, S., Wang, C., Cao, H., Jia, X. (2020). "Crime Prediction Using Data Mining and Machine Learning". In: Liu, Q., Mısır, M., Wang, X., Liu, W. (eds) The 8th International Conference on Computer Engineering and Networks (CENet2018). CENet2018 2018. Advances in Intelligent Systems and Computing, vol 905. Springer, Cham.