Crime Classification in San Francisco

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?”
So, 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- “*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.

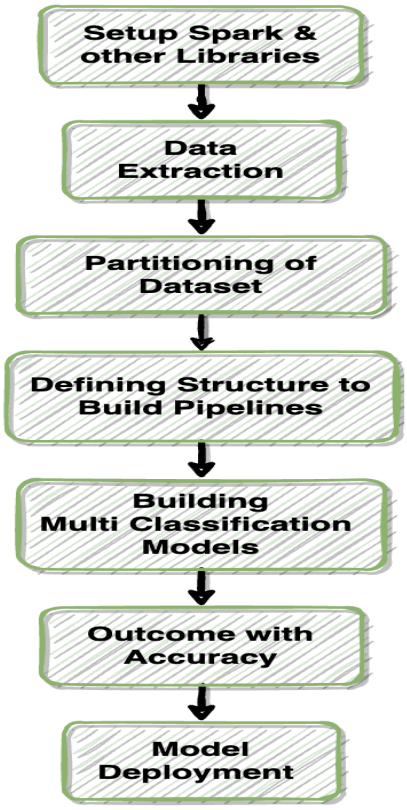
Now, coming back to our dataset, it has over 8 hundred thousand rows and we extracted around 9 variables through feature selection. These variables are as follows-

Dates- timestamp of the crime incident.
Category- category of the crime incident.
Descript- detailed description of the crime incident.
DayOfWeek- the day of the week.
PdDistrict- the name of the Police Department District.
Resolution- how the crime incident was resolved.
Address- approximate street address of the crime incident.
X- Longitude
Y- Latitude

Since, our problem statement is based on solving a multi-class text classification. So, we are using supervised machine learning techniques like - ‘**Logistic Regression**’ and ‘**Naive Bayes**. Furthermore, the end goal would be to make better classification by harnessing Big Data technologies in the machine learning realm by using the below approach/procedure (Fig-1).

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. 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 is the backbone of the whole project.



(Fig-1)

Thereafter, in the next part, we defined various structures for building pipelines for our project.
Some of the important steps involved in this process were defining -

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 -
    a. **Logistic Regression** using **CountVectorizer** features
2. Build a Secondary Model -
    a. Naive Bayes
    b. Logistic Regression & Naive Bayes using **TF-IDF features**
    c. Logistic Regression & Naive Bayes using **word2Vec**

We have completed some of the above parts. Henceforth, we are looking forward to working on it to make some analysis and prediction of the data. This will help us to get insights about classification parameters and compare the accuracy of Logistic Regression and Naive Bayes approaches.

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

Project Division and Milestones -

We believe- "*Teamwork makes the dream work!*"
Our team comprises 4 students – Gyanesh Tiwari, Abhishek Jaiswal, Ramgopal Vaka, and Raghukarn Sharma and we have split the work in the below fashion-

1. Gyanesh Tiwari - Responsible for design and development of the project by closely collaborating with the other team members.
2. Abhishek Jaiswal - Preprocessed the data and worked on training & evaluating the model with two other teammates. Additionally, I collaborated on the documentation of the project from scratch.
3. Ramgopal Vaka - Would be responsible for working closely with Abhishek doing proper detailed documentation for each step - such as an explanation of the models, reasons for feature selection, results and comparisons, improving the aesthetics of the created visualizations for the end user, summary writing, etc.
4. Raghukarn Sharma - Worked on feature selection and data extraction for the dataset. Thereafter, I helped in training & evaluating the model for classification.

Since there are roughly 4 weeks left in this semester, we are dividing our project into 3 major chunks in the following manner-

Milestone I (Week 1) - We plan on getting started with - setting up spark & required libraries and then extracting the data followed by getting training and testing datasets along with its documentation. So far we are through with this part.

Milestone II (Week 2 & 3) - In this part, we are thinking of coding the base and building the required pipeline so that we can move on to building the classification models. Additionally, we would be trying to analyze and register the outcomes and compare them for better insights.

Milestone III (Week 4) - In this final week, we hope to work on drawing conclusions from the analysis, summarization, model deployment, and improving the overall aesthetics of the project.

Overall, our aim is to make our project super easy to understand, crisp, and visually captivating. Therefore, making it not only helpful for us but for the other people who are going to use it as a reference. Thus, we hope that we would be able to overcome all the challenges to make this project a great learning launchpad for us.

References -

1. https://www.neighborhoodscout.com/ca/san-francisco/crime#description
2. https://www.sanfranciscopolice.org/stay-safe/crime-data/crime-dashboard
3. https://www.neighborhoodscout.com/ca/san-francisco/crime
4. https://www.sfgate.com/bayarea/article/san-francisco-crime-not-improving-17478702.php
5. https://www.hoover.org/research/why-san-francisco-nearly-most-crime-ridden-city-us