Exploratory Data Analysis and Crime Prediction for Smart Cities

Isha Pradhan
isha.pradhan@sjsu.edu
Department of Computer Science, San Jose State University
San Jose, California, USA

Katerina Potika
katerina.potika@sjsu.edu
Department of Computer Science, San Jose State University
San Jose, California, USA

Magdalini Eirinaki
magdalini.eirinaki@sjsu.edu
Department of Computer Engineering, San Jose State University
San Jose, California, USA

Petros Potikas
ppotik@cs.ntua.gr
School of Electrical and Computer Engineering, National Technical University of Athens
Zografou, Greece

ABSTRACT

Crime has been prevalent in our society for a very long time and it continues to be so even today. Currently, many cities have released crime-related data as part of an open data initiative. Using this as input, we can apply analytics to be able to predict and hopefully prevent crime in the future. In this work, we applied big data analytics to the San Francisco crime dataset, as collected by the San Francisco Police Department and available through the Open Data initiative. The main focus is to perform an in-depth analysis of the major types of crimes that occurred in the city, observe the trend over the years, and determine how various attributes contribute to specific crimes. Furthermore, we leverage the results of the exploratory data analysis to inform the data preprocessing process, prior to training various machine learning models for crime type prediction. More specifically, the model predicts the type of crime that will occur in each district of the city. We observe that the provided dataset is highly imbalanced, thus metrics used in previous research focus mainly on the majority class, disregarding the performance of the classifiers in minority classes, and propose a methodology to improve this issue. The proposed model finds applications in resource allocation of law enforcement in a Smart City.

CCS CONCEPTS

• **Information systems → Data analytics**; • **Computing methodologies →** *Supervised learning by classification.*

KEYWORDS

Predictive analytics, crime prediction model, multiclass classification, smart city.

ACM Reference Format:
Isha Pradhan, Katerina Potika, Magdalini Eirinaki, and Petros Potikas. 2019. Exploratory Data Analysis and Crime Prediction for Smart Cities . In *23rd International Database Engineering & Applications Symposium (IDEAS'19), June 10–12, 2019, Athens, Greece.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3331076.3331114

1 INTRODUCTION

The concept smart cities encompasses several initiatives that are supported by modern technology and aim at improving the lives of the people living within the city in various domains like urban development, safety, energy and so on [19]. One of the factors that determine the quality of life in a city is the crime rate therein. Although modern cities might offer a lot of technological advancements, the basic requirement of citizens' safety still remains an open problem [11].

Crime continues to be a threat to individuals and to our society and demands serious consideration if we aim at reducing the onset or the repercussions caused by it. Hundreds of crimes are recorded daily by the data officers working alongside the law enforcement authorities throughout the United States. Many cities have signed to the Open Data initiative, thereby making this crime data accessible to the general public. The intention behind this initiative is increasing the citizens' participation in decision-making and utilizing this data to uncover interesting and useful facts [7].

The city of San Francisco is one of many that have joined this Open Data initiative. The data scientists and engineers working alongside the San Francisco Police Department (SFPD) have recorded over 100, 000 crime cases in the form of police complaints they have received [6]. With the help of this historical data, many patterns can be uncovered. This can help us predict crimes that may happen in the future and thereby help the city police better safeguard the population of the city.

Motivated by the ideal scenario, where every citizen lives in a safe environment and neighborhood, we propose some methodologies as well as some initial results that might help the law enforcement of a city predict and tackle crime. We employ the crime data set reported by SFPD over a period of 15 years (2003 to 2018) and analyze them to identify the trends of crimes over the years and predict crimes that might happen in the future. Compared to previous work that has worked with the same data, our proposed data preprocessing methodology improves prediction for the highly imbalanced dataset [1, 4, 15, 23]. We should point out that, even though our proof-of-concept in this work employs the San Francisco crime dataset, a

IDEAS'19, June 10–12, 2019, Athens, Greece
© 2019 Association for Computing Machinery.
ACM ISBN 978-1-4503-6249-8/19/06…$15.00
https://doi.org/10.1145/3331076.3331114

similar approach can be followed to analyze to any city's or region's crime data, so we hope that our approach can help with crime prevention on a larger, national and international level.

1.1 Problem Formulation

The problem being tackled in this paper can be best explained in two distinct parts:

We first perform exploratory data analysis to identify crime patterns by:

- Utilizing the crime data set by the SFPD, to observe existing patterns in the crime throughout the city of San Francisco.
- Determining the classes of crimes within different areas in the city, and analyzing the spread and impact of the crime.
- Studying the crime spread in the city based on the geographical location of each crime, the possible areas of victimization on the streets, seasonal changes in the crime rate and the type, and the hourly variations in crime.

In the second part, we employ machine learning to generate a prediction methodology to identify the type of crime that can take place in the city, at several levels:

- Using the discovered patterns of crime identified during the exploratory analysis part, we inform and improve the data pre-processing process.
- Building a prediction model that treats this problem as a multiclass classification problem, by classifying new raw (unclassified) data into one of the crime categories (classes), thereby predicting the crime that can occur.
- Addressing the problem of an imbalanced dataset, by introducing additional data preprocessing tasks aiming at improving the precision and recall for all classes (including the minority classes) of our data. This improves previous research works that have been proposed on the same dataset.

For the exploratory data analysis, we employ various data analytics tools, along with Spark for initial data preprocessing, to analyze the spread of the crime in the city, and find the crime classes. For the machine learning/prediction part, in order to build a prediction model, we build upon the first part and use different types of algorithms, such as K-Nearest Neighbor, Multi-class Logistic Regression, Decision Tree, Random Forest, and Naïve Bayes.

The rest of the paper is organized as follows: in Section 2 we present an overview of the related work; our design and implementation details are presented in Section 3, while the results of our analysis and experimental evaluation are included in Section 4. We conclude with our plans for future work in Section 5.

2 RELATED WORK

Over the years, there have been a lot of studies involving the use of predictive analytics to observe patterns in crime. Some of these techniques are more complex than others and involve the use of more than one data sets. Most of the data sets used in these researches are taken from the Open Data initiative [7] supported by the government. In this section, we will study the various techniques used by different authors which will help answer questions such as: what is the role of analytics in crime prediction, what techniques are used for data preprocessing and what are the classification techniques which have proved to be most efficient.

2.1 Temporal and Spectral Analysis

A lot of research in the area of crime analysis and prediction revolves around the analysis of spatial and temporal data. The reason for this is fairly obvious as we are dealing with geographical data spread over the span of many years.

The authors of [17] have studied the fluctuation of crime throughout the year to see if there exists a pattern with seasons. In their research, they have used the crime data from three different Canadian cities, focusing on property related crimes. According to their first hypothesis, the peaks in crime during certain time intervals can be distinctly observed in the case of cities where the seasons are more distinct. Their second hypothesis is that certain types of crimes will be more frequent in certain seasons because of their nature. They were able to validate their hypothesis using Ordinary Least Squares (OLS) Regression for Vancouver and Negative Binomial Regression for Ottawa. Since their research focused on crime seasonality, a quadratic relationship in the data was predicted. Crime peaks were observed in the Summer months as compared to Winter.

In a similar study, the authors of [2] have analyzed the crime data of two US cities - Denver, CO and Los Angeles, CA and provide a comparison of the statistical analysis of the crimes in these cities. Their approach aims at finding relationships between various criminal entities as this would help in identifying crime hotspots. To increase the efficiency of prediction, various preprocessing techniques like dimensionality reduction and missing value handling were implemented. In the analysis, they compared the percentage of crime occurrence in both cities as opposed to the count of crimes. Certain common patterns were observed in both the cities such as the fact that Sunday had the lowest rate of crime in both the cities. Also, important derivations like the safest and the most notorious district were noted. Decision Tree classifier and Naive Bayes classifier were used.

L. Venturini *et al.* [22] have discovered spatio-temporal patterns in crime using spectral analysis. The goal is to observe seasonal patterns in crime and verifying if these patterns exist for all the categories of crime or if the patterns change with the type of crime. The temporal analysis thus performed highlights that the patterns not only change with the month but also with the type of crime. Hence, the authors of [22] rightly stress the fact that models built upon this data would need to account for this variation. They have used the Lomb-Scargle periodogram [18] to highlight the seasonality of the crime as it deals better with uneven or missing data. The AstroML Python package was used to achieve this. In their paper they have described in detail how every category of crime performs when the algorithm is applied to the data. Further, the authors suggest that researchers should focus on the monthly and weekly crime patterns.

2.2 Prediction using Clustering and Classification techniques

The authors of [20] have described a method to predict the type of crime which can occur based on the given location and time.

Apart from using the data from the Portland Police Bureau (PPB), they have also included data such as ethnicity of the population, census data and so on, from other public sources to increase the accuracy of their results. Further, they have made sure that the data is balanced to avoid getting skewed results. The machine learning techniques that are applied are Support Vector Machine (SVM), Random Forest, Gradient Boosting Machines, and Neural Networks [20]. Before applying the machine learning techniques to predict the category of the crime, they have applied various preprocessing techniques such as data transformation, discretization, cleaning and reduction. Due to the large volume of data, the authors have sampled the data to less than $20,000$ rows. They used two data sets to perform their experiments - one was with the demographic information used without alterations and in the second case, they used this data to predict the missing values in the original data set. In the first case, ensemble techniques like as Random Forest or Gradient Boosting worked best, while in the second case, SVM and Neural Networks showed promising results.

Since a smart city should give importance to the safety of their citizens, the authors of [11] have designed a strategy to construct a network of clusters which can assign police patrol duties, based on the informational entropy. The idea is to find patrol locations within the city, such that the entropy is maximized. The reason for the need to maximize the entropy is that the entropy, in this case, is mapped to the variation in the clusters, i.e. more entropy means more cluster coverage [11]. The dataset used for the research is the Los Angeles County *GIS* Data. The data has around 42 different crime categories. Taking the help of a domain expert, the authors have assigned weights to these crimes based on the importance of the crime. Also, the geocode for each record is taken into consideration and the records that do not have a geocode are skipped. Because the authors in [11] are trying to maximize the entropy in this case, by considering the equation $H_{c1} = -p(c_1)lnp(c_1)$. The probability $p(c1)$ is defined as the ratio of the weight of the centroid of the crime to the weight of the system, plus the ratio of the quickest path between two centroids, to the quickest path in the whole system.

The authors of [9] have taken a unique approach towards crime classification where unstructured crime reports are classified into one of the many categories of crime using textual analysis and classification. For achieving this, the data from various sources, including but not limited to the databases which store information about traffic, criminal warrants of New Jersey (NJ) and criminal records from NJ Criminal History, was combined and preprocessed. As a part of the preprocessing activity, all the stop words, punctuations, case IDs, phone numbers and so on were removed from the data. Following this, document indexing is performed on the data to convert the text into its concise representation. In order to identify the topics or specific incident types from the concise representation, the authors used Latent Semantic Analysis (LSA). Next, the similarity between these topics was identified using the Topic Modeling technique where the closer the score is to 1, the more similar it is to the topic which was followed by Text Categorization. The classification methods used in this research were Support Vector Machines (SVM), Random Forests, Neural Networks, MAXENT (Maximum Entropy Classifier), and *SLDA* (Scaled Linear Discriminant Analysis). However, the authors observed that SVM performed consistently better of them all.

2.3 Hotspot Detection

A crime hotspot is an area where the occurrence of crime is high as compared to other locations [8]. Many researchers have taken an interest in determining crime hotspots from the given dataset. The authors of [8] mainly discuss two approaches for detecting hotspots - circular and linear. The authors also discuss the fundamentals of Spatial Scan Statistics, as a useful tool for hotspot detection. The results on the Chicago crime data set are also discussed in detail using both the approaches.

3 DESIGN AND IMPLEMENTATION

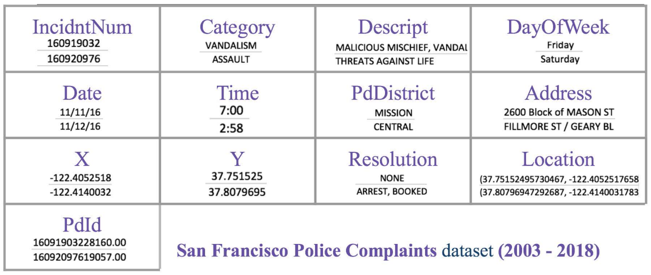
The fundamental goal of this work is to build a model, that can predict the crime category that is more likely to happen given a certain set of characteristics like the time, location, month and so on. Also, we take the help of statistical and graphical analysis to help determine which attributes contribute to the overall improvement in the Log Loss score. Our proof-of-concept application focuses on the San Francisco crime dataset. We used parallel processing using Apache Spark. Apache Spark is a big data tool which distributes the data over a cluster and achieves parallel processing. It has become popular in the recent few years [12].

3.1 Overview of the data set

We used the San Francisco crime data set [7]. The data set consists of the following attributes:

- IncidntNum: the incident number of the crime as recorded in the police logs, it is analogous to the row number,
- Descript: brief description of the crime and provides slightly more information than the *Category* field but is still quite limited,
- DayOfWeek (Date): day of the week when the crime occurred: *Monday, Tuesday, Wednesday, Thursday, Friday, Saturday, Sunday* (exact date of the crime),
- PdDistrict: police district the crime occurred in, San Francisco has been divided into 10 police districts: *Southern, Tenderloin, Mission, Central, Northern, Bayview, Richmond, Taraval, Ingleside, Park*,
- Resolution: resolution for the crime, one of these values: *Arrested, Booked, None*,
- Address: street address of the crime,
- X (Y): longitudinal (latitudinal) coordinates of the crime,
- Location: a pair of coordinates, i.e. (X, Y),
- PdId: a unique identifier for each complaint registered for database update or search operations,
- Category: category of the crime, originally, there are 39 distinct values (such as *Assault, Larceny/Theft, Prostitution, etc.*) and it is also the dependent variable we will try to predict for the test set.

There are about 1.4 million rows and the size of it is approximately 450 MB. It contains data from the year 2003 to (February) 2018. A snapshot of the actual data set is shown in Figure 1.



IncidntNum 160919032 160920976	Category VANDALISM ASSAULT	Descript MALICIOUS MISCHIEF, VANDAL THREATS AGAINST LIFE	DayOfWeek Friday Saturday
Date 11/11/16 11/12/16	Time 7:00 2:58	PdDistrict MISSION CENTRAL	Address 2600 Block of MASON ST FILLMORE ST / GEARY BL
X -122.4052518 -122.4140032	Y 37.751525 37.8079695	Resolution NONE ARREST, BOOKED	Location (37.75152495730467, -122.4052517658 (37.80796947292687, -122.4140031783
PdId 16091903228160.00 16092097619057.00	San Francisco Police Complaints dataset (2003 - 2018)		

Figure 1: Snapshot of the actual Data set

3.2 Data Preprocessing

For our preprocessing, we employ Apache Spark. This provides several advantages, especially in terms of distributed and parallel processing. It can also significantly decrease the processing time of such a huge volume of data.

The implementation of the rest of the project has been done using Python and hence we have used the PySpark distribution of Spark for preprocessing.

The data set is mostly complete with no null values. However, there are a few outliers which must be handled (see 3.2.1). The dataset provided a lot of potential to extract more meaningful information from the existing columns. Hence, a few columns have been added or transformed to improve the score of the resulting prediction. The decision to add or transform columns has been taken by studying the graphical analysis which has been performed on the data prior to building a model.

3.2.1 Data Cleaning. One of the primary steps of data cleaning is outlier detection. Using the longitude/latitude coordinates, we identify 196 outliers that fall outside the minimum boundary of San Francisco and filter them out.

The next step in data cleaning is taking care of incorrect or missing data. Although the data set does not contain Null or missing values, the Category column does contain a few columns which have been incorrectly labeled, like the *TREA* category which should actually be *TRESPASSING*.

There are 39 distinct categories in the data set. However, some of the categories are very similar to each other. For example, when the Category column contains values or keywords like *INDECENT EXPOSURE* or *OBSCENE* or *DISORDERLY CONDUCT*, we can group those together in one category *PORNOGRAPHY/OBSCENE MAT*. The decision on which categories should be clubbed together is taken by looking at the Description column of the data set which provides more information on what the corresponding Category column represents. The complete list is presented for reference in Table 1.

3.2.2 Data Transformation. Data transformation is one of the most important data preprocessing techniques. Usually, the data is originally present in the form that makes more sense if it is transformed. In this case, the main transformations performed are as follows:

Description Containing	New Category
License, Traffic, Speeding, Driving	Traffic Violation
Burglary Tools, Air Gun, Tear Gas, Weapon	Deadly Tool Possession
Sex	Sexual Offenses
Forgery, Fraud	Fraud/Counterfeiting
Tobacco, Drug	Drug/narcotic
Indecent Exposure, Obscene, Disorderly Conduct	Pornography/obscene Mat
Harassing	Assault
Influence Of Alcohol	Drunkenness

Table 1: Extracting Information from Description Column

Extracting Information from Other Attributes: On taking a closer look at the Description column, it is observed that it contains a lot of useful information which has not been captured in the Category column. For example, although the Description column explains that the crime has something to do with *WEAPON LAWS*, the Category column has classified it under *OTHER OFFENSES*. This might cause us to miss out on significant information. Hence, we extract such information from the Description column and rename the categories in the Category column. The complete list is shown in Table 2 for reference.

Original Category containing	New Category
Weapon Laws	Deadly Tool Poss.
BadCheck, Counterfeit., Embezzl.	Fraud/Counterfeiting
Suspicious Occ	Suspicious Person/act
Warrants	Warrant Issued
Vandalism	Arson

Table 2: Combining Similar Categories

Feature Extraction: There exist several features like Address, Time, Date, X and Y which can be transformed into new features that hold more meaning as compared to the existing ones. Hence, all of these features have been used to generate new features and some of these old features have been eliminated.

Address to *BlockOrJunc*: In its original form, the Address feature has a lot of distinct values. Thus, if given a logical consideration, it is not hard to realize that the exact address of the crime might not be repeated or be useful in predicting the type of crime in the future. However, this column can be used to see if the crime occurred on a street corner/junction or on a block. We can also check if there exists a pattern among certain types of crime to occur more frequently on a street corner rather than a block. To achieve this, a simple check of whether '/' occurs in the address or not, is performed. If it does contain it, it means that the crime occurred on a corner and we return 1, otherwise it is a block and we return 0.

Time to *Hour*: The Time feature is in the Timestamp format. It would be interesting to observe patterns in crime by the hour. Hence the Hour field is extracted from the Time field. It is worth noting that if the minute part is greater than 40, i.e. if the time is for example, 12 : 42, then the hour is rounded off to 13, otherwise it would be 12.

Date to *Season, Day, Year and Month*: The Date field is a very important one for prediction. Using this single field, we are able to extract four features. Spark provides inbuilt methods to extract the Day, Month and Year from the Date and hence our script makes use of the same. After extracting the Month from the Date, we make use of this feature to extract the Season.

X and Y to *Grid*: The X and the Y coordinates provide the exact location of the crime. However, we can see some interesting patterns on dividing the entire San Francisco area into $20X20$ grids. This is inspired by the work of [15], who give specific details on the formula used for the generation of these 400 cells.

3.2.3 Data Reduction. As previously mentioned, there are 39 categories of crime in the original data set. Some of them include labels like *NON-CRIMINAL, RECOVERED VEHICLE* and *SECONDARY CODES*. Since we are trying to predict the future occurrences of crimes, it is essential to have categories pertaining to actual criminal activities. However, the above labels do not provide any additional information to help us achieve our goal. Thus, these categories are completely filtered out from our data set. This reduces the number of rows from about 2.1 million to about 1.9 million after all the preprocessing.

3.3 Classification Techniques

Classification techniques are used to automatically put the data into one or more categories also known as classes.

We focus on Pigeonhole Multiclass Classification algorithms. Multiclass Classification involves classifying the data into more than two classes. One of the most common types of Multiclass Classifiers[14] is the Pigeonhole Classifier, where every item is classified into only one of the many classes. Hence, for a given item, there can be only one output class assigned to it. Below, we briefly describe the classification techniques that we used in our analysis.

(1) Naïve Bayes classifier is a supervised learning algorithm which is based on the Bayes' theorem. The Bayes' theorem can be stated as shown in $P(A|B) = P(A)\frac{P(B|A)}{P(B)}$, where $P(A|B)$ is the conditional probability of A happening given that B is true, similar for $P(B|A)$, $P(A)$ and $P(B)$ are the individual probabilities of A and B happening independently. The Naïve Bayes classifier relaxes the conditional dependence assumption of the Bayes Theorem, introducing the "naïve" assumption that there exists independence between all pairs of features. Although these classifiers are fairly simple, they tend to work very well in a large number of real world problems.

(2) Decision Tree classifiers use decision trees to make a prediction about the value of a target variable. The decision trees are basically functions that successively determine the class that the input needs to be assigned. Using decision trees for prediction has many advantages. An input is tested against only specific subsets of the data, determined by the splitting criteria or decision functions. Another advantage is that we can use a feature selection algorithm in order to decide which features are worth considering for the decision tree classifier. The fewer the number of features, the better will the efficiency of the algorithm be [21].
To construct a decision tree, generally a top down approach is applied until some predefined stopping criterion is met.

(3) Random Forest classifiers generate multiple decision trees on different sub-samples of the data while training, and then predict the accuracy or loss score by taking a mean of these values. This helps to control over-fitting that might happen when a single decision tree is used, as this algorithm is biased towards always selecting the same root of the tree (the one that gives the less entropy after the split.
To alleviate this problem, in Random Forests the split for each node is determined from a subset of the predictor variables which are randomly chosen at the given node [16].

(4) K- Nearest Neighbor (KNN) classifiers classify data into one of the many categories by taking a majority vote of its neighbors. The label is assigned depending on the most common of the categories among its neighbors. The number of neighbors to consider is a user-defined parameter K that is set after experimentation.

(5) Multinomial Logistic Regression classifiers are a generalized version of Logistic Regression for multiclass problems like ours. The log odds of the output are modeled as a combination of the various predictor variables [5]. There are two variants of Multinomial Logistic Regression based on the nature of the distinct categories in the dependent variable-nominal and ordinal [10]. Multinomial regression uses the Maximum Likelihood Estimation (MLE) method. Logistic Regression is a discriminative classifier [13](Ch 7). This means that it tries to learn the model based on the observed data directly and makes fewer assumptions about the underlying distribution.

4 EXPERIMENTAL EVALUATION

4.1 Exploratory Data Analysis

We begin by exploring our data. Exploratory data analysis is the first step of any big data analytics process. Using graphs we can get useful and interesting insights into our data. This step will also help us make data preprocessing decisions, such as which features to include for predictions. Some of these graphs show interesting patterns in crime, which might not be apparent otherwise.

Figure 2 shows the trend of the crime over the years in various districts (neighborhoods) of San Francisco. These are the Police Districts and each of those include many other city districts. Looking at this graph, we can observe that the crime in *SOUTHERN, CENTRAL* and *NORTHERN* districts is on the rise. On the other hand, crimes in *TENDERLOIN* and *INGLESIDE* have declined over the years.

Figure 3 shows how crimes happen on different hours of the day. We can observe that there is a clear pattern in crime and the hour of the day. Generally, the crime rate is low in the early morning hours from around 3:00 AM to 6:30 AM and it rises to its peak in the evening rush hours, i.e., from 4:30 PM to 7:00 PM and is generally high at night. However, it would be really interesting to see if this pattern is followed by all the different types of crime. For this, we plot graphs for the top four crimes that we found interesting.

Figure 4, focuses on Theft/Larceny crimes per hour. It pretty much follows the trend of the previous graph.

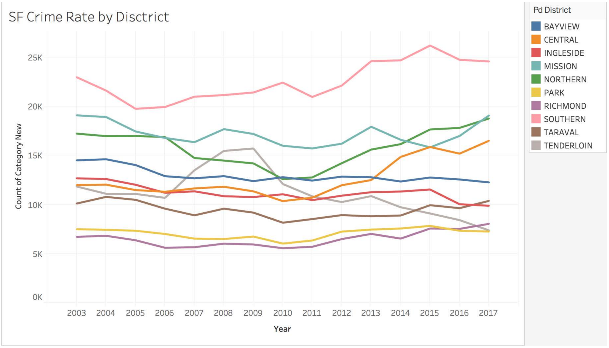


Figure 2: Rate of Crime per District by Year

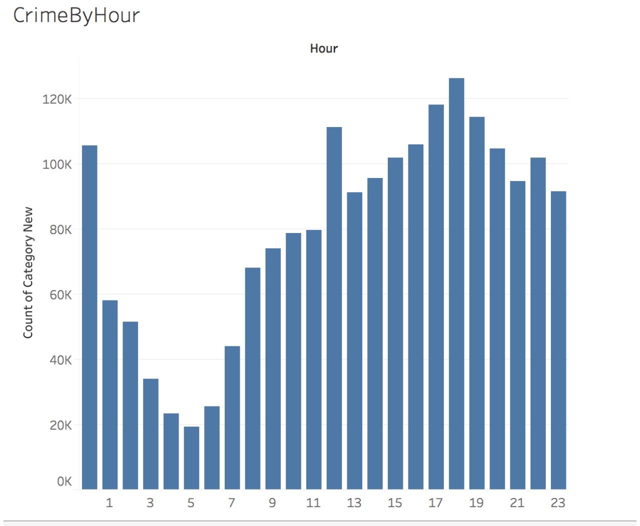


Figure 3: Rate of overall crime every Hour

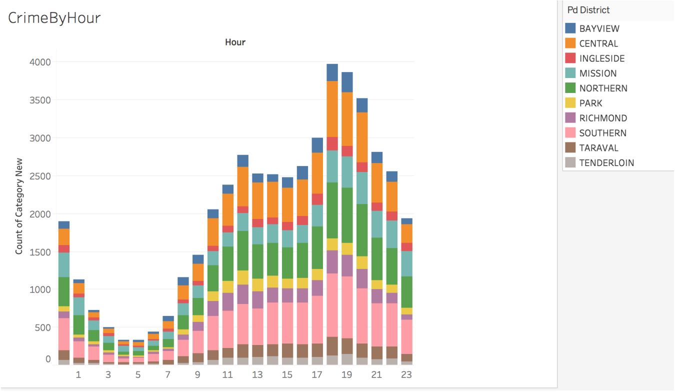


Figure 4: Rate of Theft/Larceny by the Hour

However, the same pattern is not followed by other crime types. For example, as shown in Figure 5 that plots Prostitution crimes, there are clear areas where Prostitution is high as compared to others and we can also see that Prostitution is higher during midnight and late hours (something that was expected). However, it is also very high around 11 : 00 AM in the Central district, which is unusual and can be further looked into by the police department and law enforcement agents.

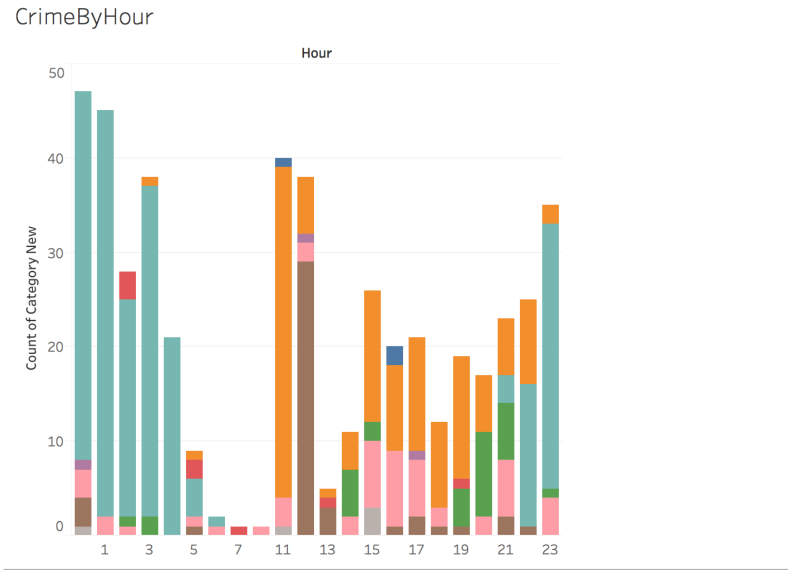


Figure 5: Rate of Prostitution by the Hour

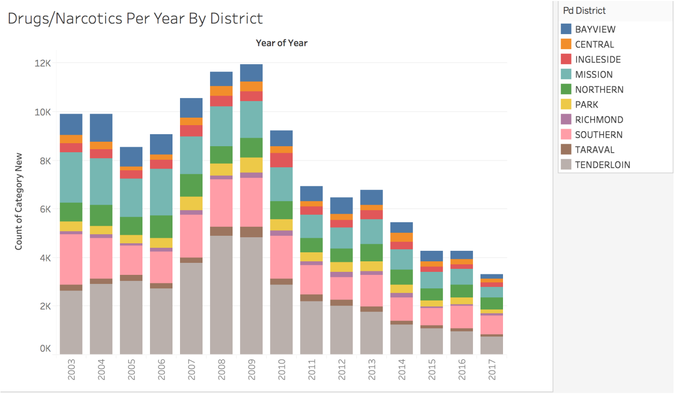


Figure 6: Sum of Drugs/Narcotics cases per Year

In Figure 6, we can see that the Drug/narcotics related crimes were highest in the year 2009 followed by 2008. Anyone even slightly familiar with San Francisco might mention that the Tenderloin is one of the most notorious districts in San Francisco with a high crime rate, especially, with high rate Drugs and Narcotics related crimes, and this impression is supported by the numbers shown here. From Figure 6 we can see that Tenderloin district has the highest number of Drug related crimes till 2009. However, in recent years, these crimes have seen a huge dip, going down by more than 50% since 2009. This might be due to the fact that

SFPD has focused their efforts on fighting crime in this notoriously crime-prone neighborhood.

A great way to study the growth or decrease in the rate of crime is by using area charts. An area chart is another way to look at the growth (or fall) rate in the data In Figure 7 we study the rise in the number of thefts over the years in most of the districts in San Francisco, except Tenderloin and Taraval. On the other hand, by plotting the area chart of Drugs and Narcotics as shown in Figure 8 we can see a clear decrease in these crimes in San Francisco.

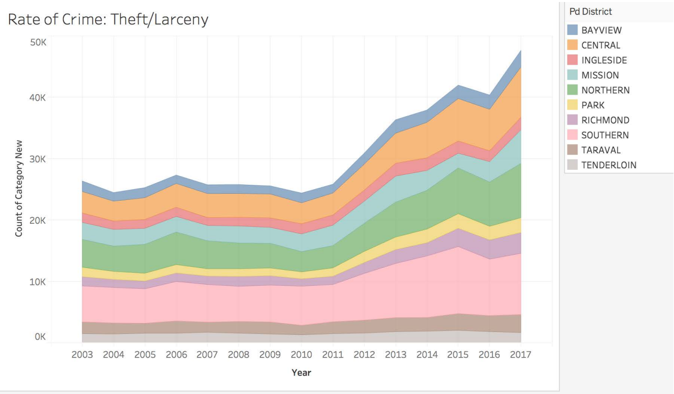


Figure 7: Area of Theft/Larceny by the Year

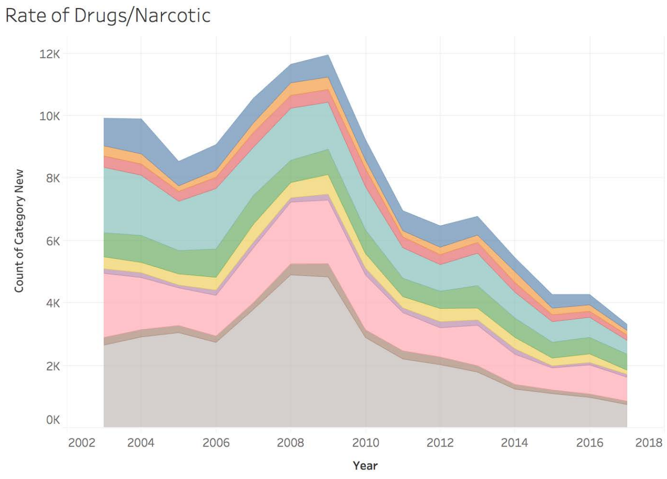


Figure 8: Area of Drugs/Narcotics by the Year

4.2 Comparison with existing results

As discussed previously, several researchers have also worked with the SF crime dataset. In this section, we provide a comparative analysis.

Our data preprocessing results in a reduction in the number of rows in the dataset from 2.19 to 1.92 million. We split the dataset to training and test chronologically as follows: as training data we use data from year 2003 to year 2015, consisting of 1, 636, 217 rows; as test data we use data from year 2016 to year 2018, consisting of 284, 165 rows.

Following the practice of researchers in related work, we use the Log Loss score for our models. In this scoring metric, false classifications are penalized. The less the Log Loss score, the better is the model. For a perfect classifier, the Log Loss score would be zero [3].

Mathematically, the Log Loss function is defined as follows:

$$-\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{M} y_{ij}\, log\, p_{ij}$$

where N is the total number of samples, M is the number of distinct categories present in the output variable, y_{ij} takes the value of 0 or 1 indicating if the label j is the expected label for sample i and p_{ij} if the probability that label j will be assigned to the sample i [3].

While we cannot directly compare the results, as we cannot know how the other researchers split their datasets, we can gauge how well our approach is performing compared to this of previous work, and also provide some insights on which classifier seems to work best for this particular dataset. We first review the features used for building the various models in other existing approaches and their reported results.

Source 1 [4]: The authors have used the features *DayOfWeek, PdDistrict, X, Y, Month, Year, Hour and Grid(of 8 X 8)* for the prediction. Their best model is Random Forest with *LogLoss* = 2.496, with second best the Decision Tree with *LogLoss* = 2.508. The authors also evaluated Naive Bayes and Logistic Regression.

Source 2 [1]: The authors have used *Hour, Month, District, DayOfWeek, X, Y, Street No., Block and 3 components of PCA*. Their best model is Random Forest with *LogLoss* = 2.366, with second best the KNN classifier with *LogLoss* = 2.621. The authors also evaluated Naive Bayes.

Source 3 [23]: The attributes/features used for prediction in this work are *Year, Month, Hour, DayOfWeek, PdDistrict, X, Y and Block/Junction*. Their best model is Logistic Regression with *LogLoss* = 2.45. The authors also evaluated KNN, which yielded very high log loss.

Source 4 [15]: The features used for prediction are *Hour, DayOfWeek, Month, Year, PdDistrict, Season, BlockOrJunction, CrimeRepeatOrNot, Cell and 39-d Vector*. The authors only evaluated Logistic Regression, with *LogLoss* = 2.365

As shown in Table 3, in our approach, Random Forest is also the best model. In terms of Log Loss, our model yields the best results among the reported related work ones, with *LogLoss* = 2.276 while the second best model is the Decision tree (*LogLoss* = 2.3928).

One important aspect, left out by the previous papers focusing on crime classification in San Francisco, is the issue of data imbalance. The data set is highly skewed, as shown in the sum of the distinct categories of Figure 9. We discuss how we address this problem in what follows.

4.3 Improving Classification of Imbalanced Datasets

Most of the existing work uses accuracy or Log Loss score to evaluate the efficiency of the model. However, these metrics provide an overall assessment of the classifier, without focusing on how well

Algorithm	Log Loss
Random Forest	**2.2760**
Naive Bayes	2.5008
Logistic Regres.	2.4042
KNN	2.4634
Decision Tree	2.3928

Table 3: Results of Experiments (Log Loss**)**

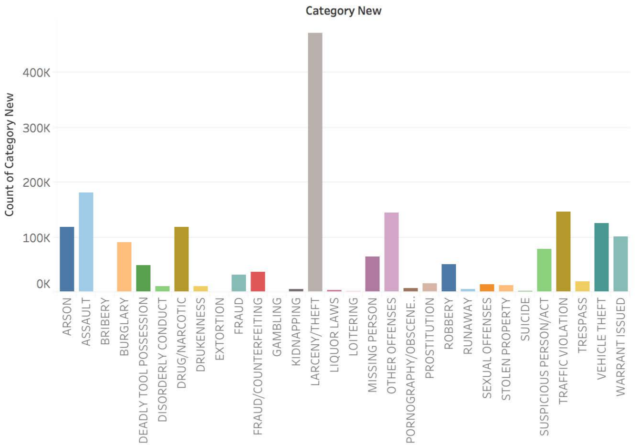


Figure 9: Count of Distinct Categories in the Dataset

the classifier does for each class. Accuracy measures the percentage of correct predictions overall predictions, so even if the classifiers don't work well with minority classes, accuracy can still be very high. The Log Loss metric does discriminate among different classes, however, it weighs each type of misclassification equally. Again, a similar misleading result might be calculated, if the classifier works well for the majority classes (which is most often the case, as it is trained using such an imbalanced dataset). Instead, we need the model to correctly identify maximum samples but at the same time, we want those correctly identified samples to include the minority classes as well, in other words increasing precision and recall for each and every class in the model.

Looking at the SF crime data, we observe that even after preprocessing, the dataset is imbalanced with the *LARCENY/THEFT* category acting as the majority class. We tried three techniques to handle the imbalance: oversampling the minority classes, oversampling the majority class, and adjusting weights on the classifiers. However, none of them showed a significant improvement in the Recall or Precision scores. Hence, the following preprocessing was performed in addition to the approaches described previously:

1. The *LARCENY/THEFT* category was split taking into consideration the Description column. It was observed that separating out the samples with *Grand Theft From Auto* in their description proved to be a good split. The resulting classes were *LARCENY/THEFT* and *THEFT FROM AUTO.*
2. Combined classes with less than 2000 samples into *OTHER OFFENSES* category.
3. Created a new category called *VIOLENT/PHYSICAL CRIME* which includes former categories of *ARSON, WEAPON LAWS, VANDALISM* and instances of *ROBBERY*, where physical harm or guns were involved.

This made the dataset more balanced (see Figure 10) than the original set. We can observe this by comparing the recall of the model for the original (Figure 11) and the balanced (Figure 12) dataset.

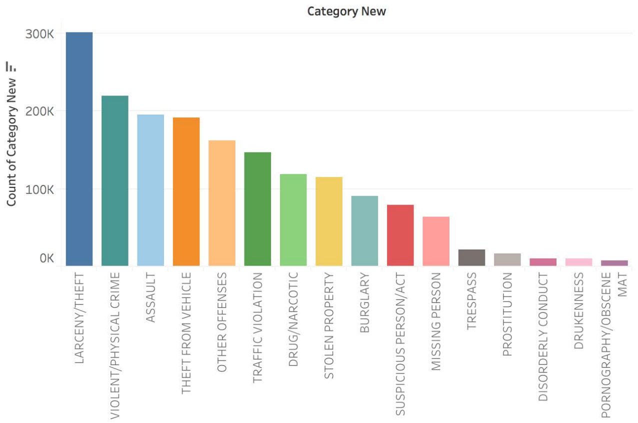


Figure 10: More Balanced data set

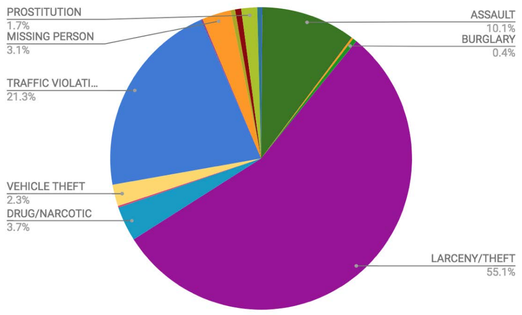


Figure 11: Recall for imbalanced dataset

5 CONCLUSION AND FUTURE WORK

In this work, we conducted a detailed analysis of the Open Data set of crime activity over 15 years for the city of San Francisco. We performed exploratory data analysis and extensive data preprocessing. Compared to previous work, we tried to alleviate the problem of an imbalanced dataset in order to improve the results of multi-class

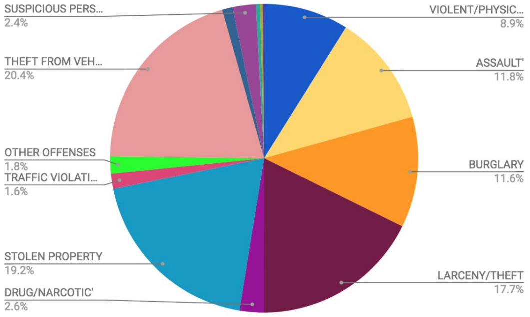


Figure 12: Recall for the more balanced dataset

classification. As a part of the future work, we plan to evaluate how other classifiers, such as neural networks, can be employed to further improve the results of the classification process. We also plan to enhance this dataset with additional metadata, such as population, housing and transportation data to gain more insights on the crime prediction process. Finally, we should stress that the proposed approach can be applied to other cities' crime datasets and see if there are any similarities and differences depending on the region.

REFERENCES

[1] Yehya Abouelnaga. San Francisco crime classification. *arXiv preprint arXiv:1607.03626*, 2016.

[2] Tahani Almanie, Rsha Mirza, and Elizabeth Lor. Crime prediction based on crime types and using spatial and temporal criminal hotspots. *arXiv preprint arXiv:1508.02050*, 2015.

[3] Exegetic Andrew B. Collier. Making Sense of Logarithmic Loss. http://www.exegetic.biz/blog/2015/12/making-sense-logarithmic-loss/, 2015.

[4] Shen Ting Ang, Weichen Wang, and Silvia Chyou. San Francisco crime classification. *University of California San Diego*, 2015.

[5] J. Bruin. Ucla: Multinomial logistic regression @ONLINE, February 2011.

[6] City and County of San Francisco. Police Department Incidents. https://data.sfgov.org/Public-Safety/Police-Department-Incidents/tmnf-yvry/, 2017.

[7] DataSF. Open government. https://www.data.gov/open-gov/. Accessed 2018-04-12.

[8] Emre Eftelioglu, Shashi Shekhar, and Xun Tang. Crime hotspot detection: A computational perspective. In *Data Mining Trends and Applications in Criminal Science and Investigations*, pages 82–111. IGI Global, 2016.

[9] Debopriya Ghosh, Soon Chun, Basit Shafiq, and Nabil R Adam. Big data-based smart city platform: Real-time crime analysis. In *Proceedings of the 17th International Digital Government Research Conference on Digital Government Research*, pages 58–66. ACM, 2016.

[10] Jelle J Goeman and Saskia le Cessie. A goodness-of-fit test for multinomial logistic regression. *Biometrics*, 62(4):980–985, 2006.

[11] Jacob Hochstetler, Lauren Hochstetler, and Song Fu. An optimal police patrol planning strategy for smart city safety. In *2016 IEEE 18th International Conference on HPCC/SmartCity/DSS*, pages 1256–1263. IEEE, 2016.

[12] Dennis Hsu, Melody Moh, and Teng-Sheng Moh. Mining frequency of drug side effects over a large twitter dataset using apache spark. In *Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2017*, pages 915–924. ACM, 2017.

[13] Dan Jurafsky and James H Martin. Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition, 2009.

[14] Brian Kolo. *Binary and Multiclass Classification*. Lulu. com, 2011.

[15] Gabriela Hernandez Larios. Case study report: San Francisco crime classification, 2016.

[16] Andy Liaw, Matthew Wiener, et al. Classification and regression by randomforest. *R news*, 2(3):18–22, 2002.

[17] Shannon J Linning, Martin A Andresen, and Paul J Brantingham. Crime seasonality: Examining the temporal fluctuations of property crime in cities with varying climates. *International journal of offender therapy and comparative criminology*, 61(16):1866–1891, 2017.

[18] Nicholas R Lomb. Least-squares frequency analysis of unequally spaced data. *Astrophysics and space science*, 39(2):447–462, 1976.

[19] Paolo Neirotti, Alberto De Marco, Anna Corinna Cagliano, Giulio Mangano, and Francesco Scorrano. Current trends in smart city initiatives: Some stylised facts. *Cities*, 38:25–36, 2014.

[20] Trung T Nguyen, Amartya Hatua, and Andrew H Sung. Building a learning machine classifier with inadequate data for crime prediction. *Journal of Advances in Information Technology Vol*, 8(2), 2017.

[21] Philip H Swain and Hans Hauska. The decision tree classifier: Design and potential. *IEEE Transactions on Geoscience Electronics*, 15(3):142–147, 1977.

[22] Luca Venturini and Elena Baralis. A spectral analysis of crimes in San Francisco. In *Proceedings of the 2nd ACM SIGSPATIAL Workshop on Smart Cities and Urban Analytics*, page 4. ACM, 2016.

[23] Xiaoxu Wu. An informative and predictive analysis of the San Francisco police department crime data, Master Thesis, 2016.