Crime Classficiation in San Francisco Bay Area

Anindita Mozumder
Computer Science
NCSU
Raleigh, USA
E-mail: amozumd@ncsu.edu

Santhosh K Jagadish
Computer Science
NCSU
Raleigh, USA
E-mail: sjagadi@ncsu.edu

James Morgan
Computer Science
NCSU
Raleigh, USA
E-mail: jamorga3@ncsu.edu

Stuti Nanda
Computer Science
NCSU
Raleigh, USA
E-mail: snanda@ncsu.edu

Nithin R Kotha
Computer Science
NCSU
Raleigh, USA
E-mail: nkotha2@ncsu.edu

Srikar Potta
Computer Science
NCSU
Raleigh, USA
E-mail: spotta@ncsu.edu

Prashant Sharma
Computer Science
NCSU
Raleigh, USA
E-mail: prasha2@ncsu.edu

Abstract— There are several data mining techniques which have been studied and implemented in various projects in organizations and research. Selecting a suitable algorithm still remains a big challenge, especially with the plethora of easy-to-use functions in several data mining packages. In this project, which involves the crime classification in San Francisco, we investigate and experiment several algorithms and evaluate their results in terms of efficiency and correctness in classifying and predicting the crime. We believe that one of the ways to improve the existing efforts taken by the police department and aid in the crime prevention is by the help of crime classification. We have used the data set from kaggle, which consists of 12 years of crime records(reports) of San Francisco Area. The classification model those were investigated were Decision Tree, Random Forest, SVM and Logistic Regression.

Keywords-component; Crime Classification, Neural Network, Decision Tree, SVM, Random Forest

I. INTRODUCTION

San Francisco is currently the cultural, commercial, and financial center of Northern California. Today the city is known more for its tech scene but it has a massive criminal past. The sudden growth in the population has brought an inequality in terms of living, housing shortages leading to no scarcity of crime in the city by the bay. The project aims at predicting the category of crime based on the twelve years of provided past records. A variety of classification techniques like Logistic Regression, Random Forests, Decision Tree, SVM, Bagging Algorithm are applied to predict the category of crime based on time and location and their performance has been juxtaposed to identify the best fit model and showcase the significant differences. Accuracy and Logarithmic loss have been used as performance evaluation measures for the models.

II RELATED WORK AND OUR CONTRIBUTION

One of the most difficult tasks in data mining is choosing the right data mining technique.[6]. The main steps followed in any data-mining project include these high-level tasks[6]

- Developing and understanding the domain.
- Selecting the right dataset and sample.
- Data Cleaning
- Data Reduction
- Choosing the right data mining technique and algorithm.
- Data mining
- Interpreting the mined data
- Consolidating the knowledge

In our project, we have tried some of the well-known classification techniques and compared the results obtained. The techniques and algorithms which have been evaluated are Decision tree, Bagging, Support Vector Machine and Logistic Regression.

The dataset is provided by the SF Open Data, the central clearinghouse for data published by the City and County of San Francisco. This dataset was generated by the Crime Incident Reporting system used by the San Francisco Police department to record all the incidents that were reported from the 1/1/2003 to 5/132015. The train set and the test rotated every week i.e. odd weeks belong to the test set while even weeks belong to the training set.

Following are the data fields present in the dataset:

- Dates: timestamp of the crime incident. This includes the date, month, year and time of the crime.
- DayOfWeek: the day of the week. Sunday to Saturday.

- PdDistrict: name of the Police Department District
- Address: the approximate street address of the crime incident

With the aid of the above mentioned attributes, our model was trained to classify and predict the following attribute:

- Category: category of the crime incident .

The other attributes present in the training dataset were:

- Descript: - detailed description of the crime incident
- Resolution: how the crime incident was resolved

Our problem is a multi-label classification problem. There are many standard methods available to approach classification problem. In this paper we have used Decision Trees, Random Trees, Bagging Ensembles and Logistic Regression to model our problem. Sections III, IV, V and VI briefly describe the various methods.

III. DECISION TREE

Decision trees are popular classification method. This method works on the basis of tree like graph model and leaves of the tree represent the outcome. For large dataset it may overfit if it uses all features, in which case we use pruning to avoid overfitting. It uses the different index for maximizing information gain by selection of features as nodes. In this paper we have used Gini Index for our model.

IV. SVM

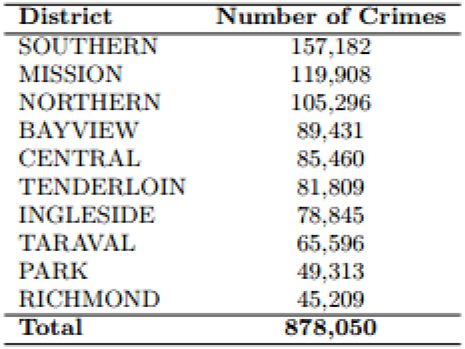
SVM(Support Vector Machine) is a set of supervised learning method. We have used it as a classification algorithm which works on basis of generating a smooth plane or hyperplane separating the classes. It tries to maximize the distance between classes as a cost function. When used with kernel, it can produce high dimensional hyperplanes for intricate dataset. In this paper we have used RBFDOT kernel from kernlab package in R.

V. LOGISTIC REGRESSION

Logistic regression is a classification technique. It assumes there exists a smooth linear decision boundary. In most of the standard algorithms, an inverse logit function is applied to a weighted sum of our features. Then it finds the weights by a maximum likelihood approach. Logit function calculates the probability of an example label, and based on our baseline we can classify that probability as an event or a nonevent.

VI.DECISION TREE WITH BAGGING

Bagging method is a part of the ensemble methods working on the principle of Decision Trees. In bagging methods, we build several instances of an estimator on random subsets of the original training set and then aggregate their individual predictions to form a final prediction. Training set for each individual estimator is a subsample of total training set. These methods are very helpful in reducing variance and maintaining bis-variance tradeoff.



District	Number of Crimes
SOUTHERN	157,182
MISSION	119,908
NORTHERN	105,296
BAYVIEW	89,431
CENTRAL	85,460
TENDERLOIN	81,809
INGLESIDE	78,845
TARAVAL	65,596
PARK	49,313
RICHMOND	45,209
Total	**878,050**

Table 1. Distribution of crime wrt PDDistrict

VII. EXPERIMENT

The experiments were performed on Intel(R) core ™ i5-5200U CPU with an 8GB RAM.

A *Data Collection*

Data is provided from SF Open Data.

B. *Feature Selection and Sampling*

The dataset includes data from 6 Jan 2003 to 13 May 2015 inclusive, and has about 870,000 data points, 6 attributes (Day, Time, Day of Week, District, Address, X and Y coordinates). Using such large datasets needs more computation time and power, but we had limited resources available. Hence, we had to sample the data to obtain a smaller dataset. The most direct way to achieve a scientifically generalizable sample is a form of probability sampling - simple random sample (SRS).

For attribute selection, the following attributes were considered.

- PdDistrict - Name of the Police Department District. Tab1 gives the distribution wrt PdDistrict.
- Dates - Timestamp of the crime incident
- DayOfWeek - The day of the week
- Address - The approximate street address of the crime incident
- X - Longitude
- Y – Latitude

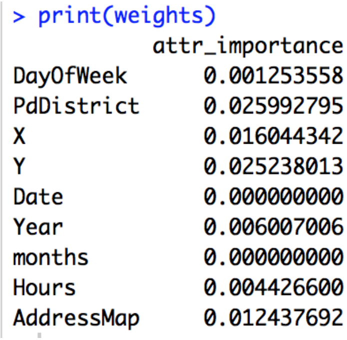
Since date was a time stamp, it was split as

- Date
- Month
- Years
- Hours
- Minutes
- Second

For our algorithms, we selected months, hours, date and year and ignored minutes and second.

Larceny/theft was the single largest type of crime committed. Hence, it was assigned a value 0 and other crimes were assigned a value of 1.

The basis of our feature selection was that if any feature has high information gain then we can say that feature is important for describing the model and rank high in order with respect to other features. A rule of thumb is if a feature has high variance then there is high probability of that feature being important for model. Fig2 gives the importance of the various attributes.



```
> print(weights)
            attr_importance
DayOfWeek       0.001253558
PdDistrict      0.025992795
X               0.016044342
Y               0.025238013
Date            0.000000000
Year            0.006007006
months          0.000000000
Hours           0.004426600
AddressMap      0.012437692
```

Fig2. Weights of the various attributes

The formula that we got was:

CategoryMap ~ PdDistrict + Y + X + AddressMap + Year + Hours

We performed Chi square test using ANOVA for comparing two different models to identify goodness of fit test. We used a model with all the mentioned parameters and another sub model with following parameters: PdDistrict, Y, X, Hours, AddressMap and Year.

Model 1: CategoryMap ~ PdDistrict + months + Y + X + Year + DayOfWeek + Hours + AddressMap + Date

Model 2: CategoryMap ~ PdDistrict + Y + X + Hours + AddressMap + Year

Hypothesis Test:

H0: The two models are similar.

HA: The two models are different.

We received a p value less than 0.05. Thus making the alternate hypothesis true.

After feature selection we were able to improve accuracy from 22% to 66%.

C. *Algorithms Evaluated*

Logistic Regression and decision tree with bagging were implemented. The results obtained were verified against the standard R packages and Weka. Other functions were executed from standard R package’s functions. Fig3 tabulates the results.

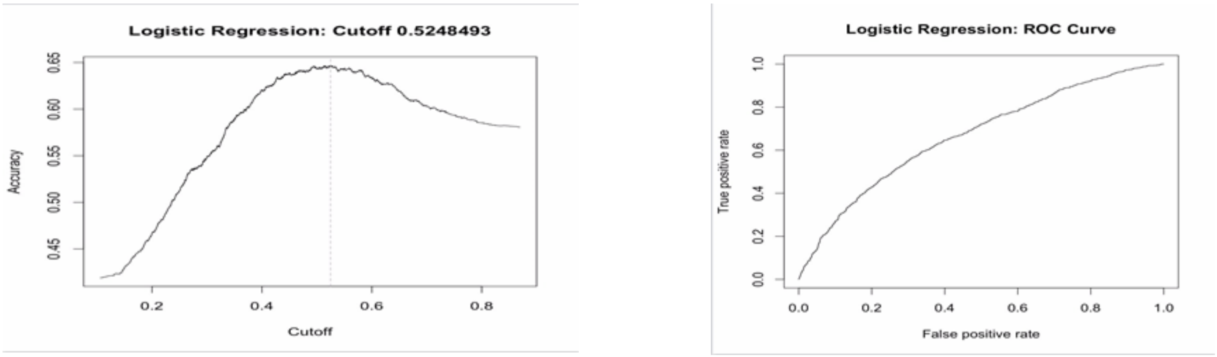


Fig1,2 : ROC Graph: Logistic regression is classifying the probability of theta.

Classifier	Accuracy	Methodology	Top Recall	Top Precision
Decision Tree	**66.06%**	**R/Weka**	**80.92%**	**67.33%**
SVM	**63.09%**	**R**	**75.95%**	**65.85%**
Random Forest	**70.833%**	**R/Weka**	**72.37%**	**80.40%**
Logistic Regression	**62.37%**	**R**	**77.92%**	**40.92%**
Decision Tree with Bagging	**67.10%**	**R**	**80.23%**	**44.52%**

Fig3. Results of our experiments

VIII. CONCLUSION AND FUTURE WORKS

The overall accuracy that we got from the various methods ranged between 65%-70% accuracy which were comparable to the ones obtained from the verifying functions in Weka and R. Although, we obtained satisfactory results, we believe that in future, we may be able to implement more algorithms which will yield better results.

IX ACKNOWLEDGMENTS

We would like to thank Dr. Vatsavai Raju, Krishna Gadiraju and Yihuan Dong for their guidance and encouragement.

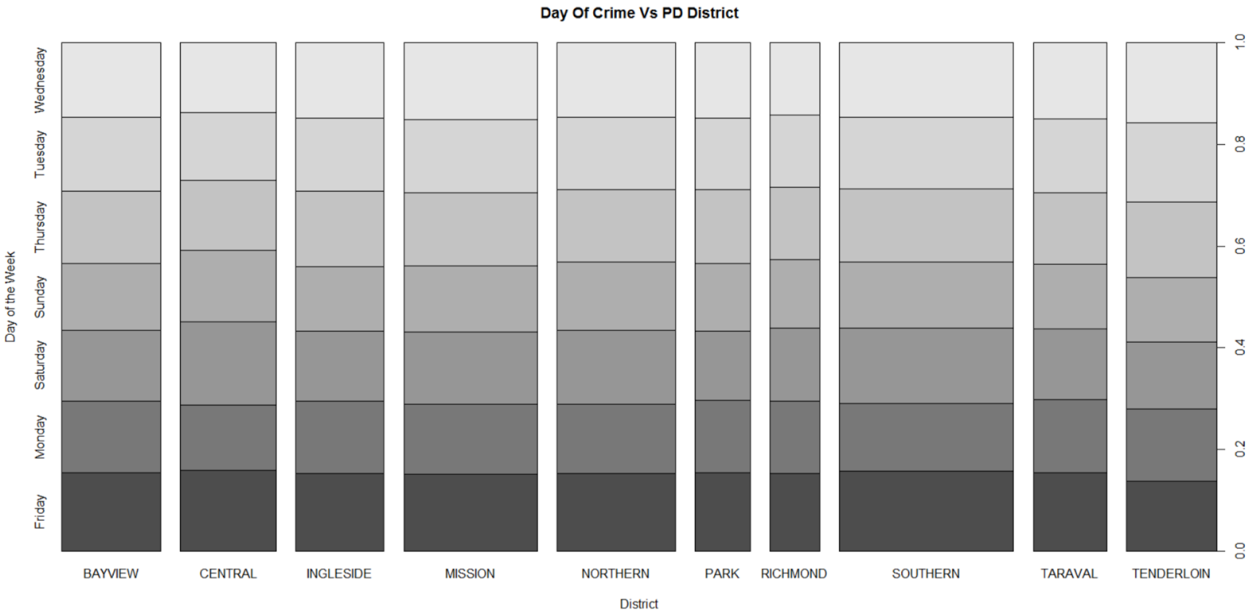
References

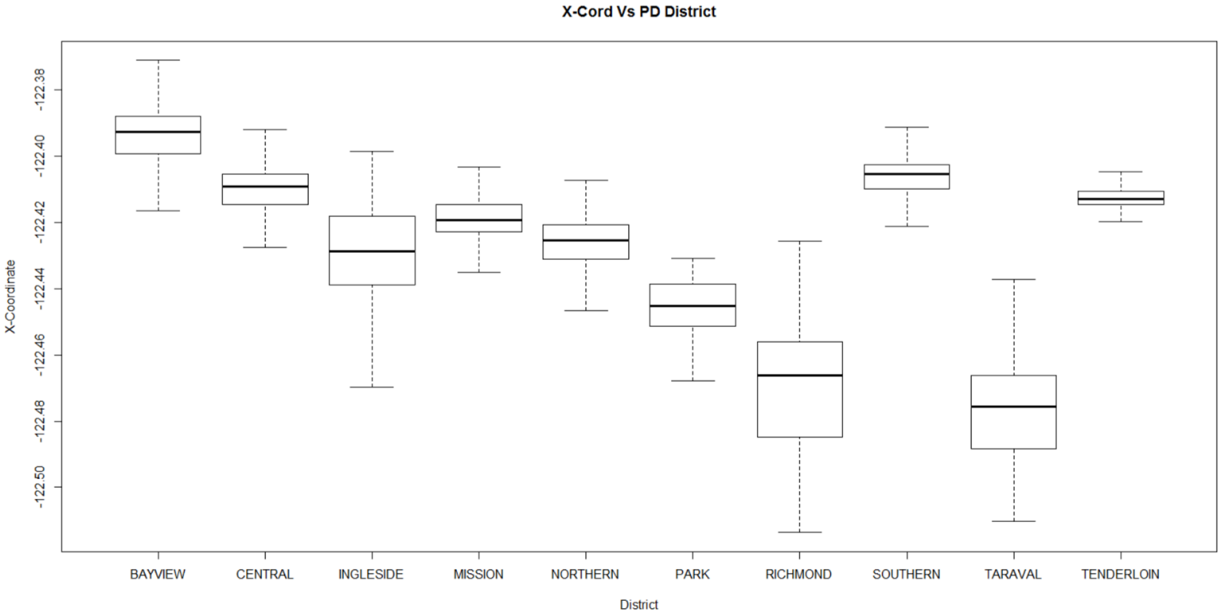
[1]. https://data.sfgov.org/
[2]. https://www.kaggle.com/c/sf-crime
[3]. http://cseweb.ucsd.edu/~jmcauley/cse255/reports/fa15/012.pdf

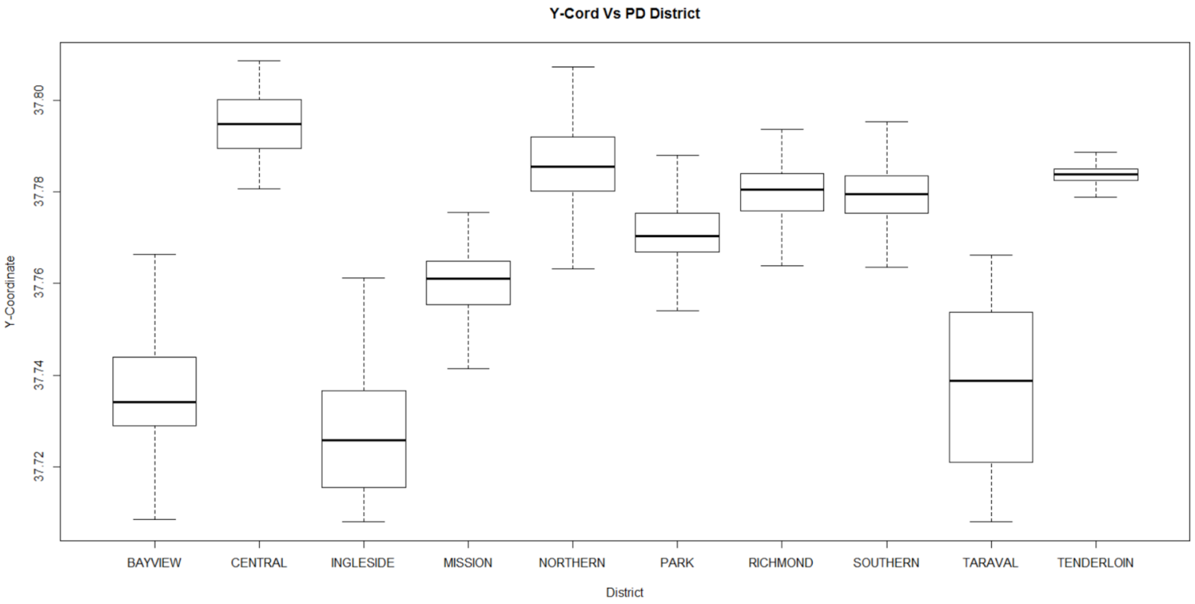
[4]. http://www.fairfaxcounty.gov/demogrph/pdf/samplingprocedures.pdf
[5]. http://cs229.stanford.edu/proj2015/254_report.pdf
[6]. Choosing the Right Data Mining Technique: Classification of Methods and Intelligent Recommendation Karina Giber et al.

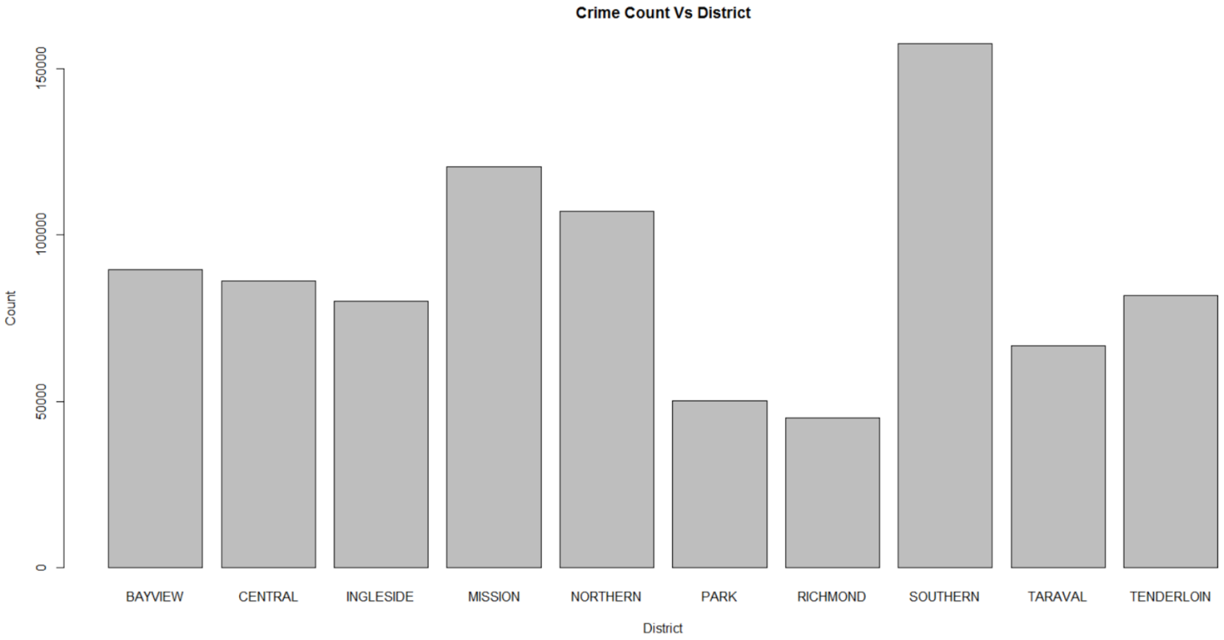
IAppendixI.

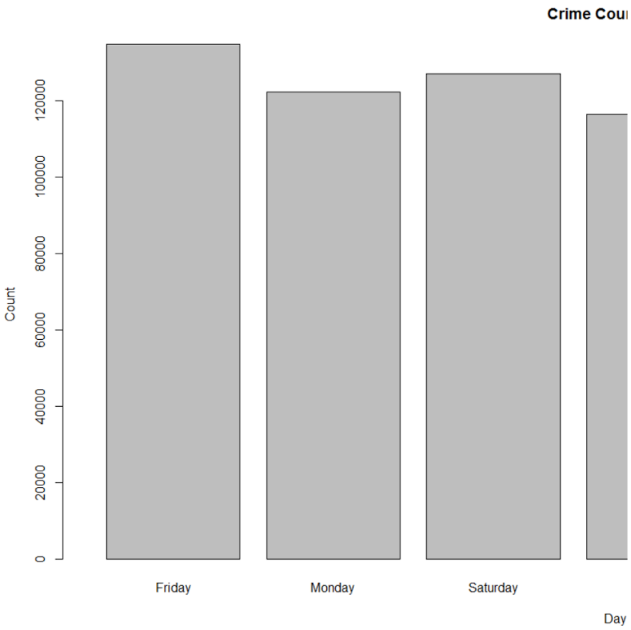
DATASET AND EDA






Y-Cord Vs PD District
Y-Coordinate
37.72
37.74
37.76
37.78
37.80
BAYVIEW
CENTRAL
INGLESIDE
MISSION
NORTHERN
PARK
RICHMOND
SOUTHERN
TARAVAL
TENDERLOIN
District


Crime Count Vs District
Count
0
50000
100000
150000
BAYVIEW
CENTRAL
INGLESIDE
MISSION
NORTHERN
PARK
RICHMOND
SOUTHERN
TARAVAL
TENDERLOIN
District


Crime Cou
Count
0
20000
40000
60000
80000
100000
120000
Friday
Monday
Saturday
Day