SFO Crime Classification

Anusha Ganti
MS in Computer Science
George Mason University
aganti@gmu.edu

Ashok Vardhan Kari
Data Analytics Engineering
George Mason University
akari@gmu.edu

ABSTRACT

From 1934 to 1963, San Francisco was ill-famed for housing some of the world's most notorious criminals on the inevitable island of Alcatraz. San Francisco is the fourth-most populous city in California, after Los Angeles, San Diego and San Jose, and the 13th-most populous city in the United States—with a Census-estimated 2015 population of 864,816. Today, the city is known more for its tech scene than its criminal past. But, with rising wealth inequality, housing shortages, and a proliferation of expensive digital toys riding BART to work, there is no scarcity of crime in the city by the bay. This project is hosted on Kaggle (Home for data science) website. The dataset is provided by SF Opendata [1], the central clearing house for data published by city and county of San Francisco. The goal for project is to predict the category of crime, given time and location it occurred. Also to explore the dataset visually and find interesting patterns of crimes occurred. [2]

Keywords

SFO Crime classification; exploratory data analysis; google vis; Liblinear; binarize categorical predictors; time series trend; data partition; extreme gradient boosting.

1. INTRODUCTION

Criminal activity is inevitable in our lives. By having knowledge of the spatial and temporal patterns of criminal activity, we can take the right measures to eradicate crime in any locality. We did a through exploratory analysis on the dataset to extract interesting patterns. This left us with thorough understanding of data visually providing great insights and remarkable trends. This paper also discusses the models we used for the task and analyze their pros and cons. We employed several models both linear and non-linear, like Naive Bayes, k-NN, logistic regression, SVM and Gradient Tree Boosting to predict the category of the crimes, and the performance of the models shows significant differences. We also submitted our result to Kaggle to see our models' performance.

2. DATA SOURCES

Data is extracted from Kaggle website. Dataset is provided by SF Opendata. Two datasets "Train" and "Test" are provided. Both the datasets have 900,000 records approximately. Train dataset has a column with category of crime which is to be found in the test set using statistical classification techniques. This dataset contains incidents derived from SFPD Crime Incident Reporting system. The data range from 1/1/2003 to 5/13/2015. The training set and test set rotate every week, meaning week 1,3,5,7... belong to the test set, and week 2,4,6,8... belong to the training set. For each row of data, there are 9 columns: [2]

Download data from - https://www.kaggle.com/c/sf-crime/data

Dates: timestamp of the crime incident

Category: category of the crime incident (only in train.csv). This is the target variable we are going to predict.

Descript: detailed description of the crime incident (only in train.csv)

DayOfWeek: the day of the week

PdDistrict: name of the Police Department District

Resolution: how the crime incident was resolved (only in train.csv)
Address: the approximate street address of the crime incident

X: Longitude

Y: Latitude

Table 1 : Sample Train dataset

Dates	Category	Descript	DayOfWeek	PdDistrict	Resolution	Address	X	Y
5/13/2015 23:53	WARRANTS	WARRANT ARREST	Wednesday	NORTHERN	ARREST, BOOKED	OAK ST / LAGUNA	-122.426	37.7746
5/13/2015 23:53	OTHER OFFENSES	TRAFFIC VIOLATION ARR	Wednesday	NORTHERN	ARREST, BOOKED	OAK ST / LAGUNA	-122.426	37.7746
5/13/2015 23:33	OTHER OFFENSES	TRAFFIC VIOLATION ARR	Wednesday	NORTHERN	ARREST, BOOKED	VANNESS AV / GR	-122.424	37.80041
5/13/2015 23:30	LARCENY/THEFT	GRAND THEFT FROM LOC	Wednesday	NORTHERN	NONE	1500 Block of LOI	-122.427	37.80087
5/13/2015 23:30	LARCENY/THEFT	GRAND THEFT FROM LOC	Wednesday	PARK	NONE	100 Block of BRO	-122.439	37.77154
5/13/2015 23:30	LARCENY/THEFT	GRAND THEFT FROM UN	Wednesday	INGLESIDE	NONE	0 Block of TEDDY	-122.403	37.71343
5/13/2015 23:30	VEHICLE THEFT	STOLEN AUTOMOBILE	Wednesday	INGLESIDE	NONE	AVALON AV / PER	-122.423	37.72514
5/13/2015 23:30	VEHICLE THEFT	STOLEN AUTOMOBILE	Wednesday	BAYVIEW	NONE	KIRKWOOD AV / I	-122.371	37.72756
5/13/2015 23:00	LARCENY/THEFT	GRAND THEFT FROM LOC	Wednesday	RICHMOND	NONE	600 Block of 47TH	-122.508	37.7766
5/13/2015 23:00	LARCENY/THEFT	GRAND THEFT FROM LOC	Wednesday	CENTRAL	NONE	JEFFERSON ST / L	-122.419	37.8078
5/13/2015 22:58	LARCENY/THEFT	PETTY THEFT FROM LOCI	Wednesday	CENTRAL	NONE	JEFFERSON ST / L	-122.419	37.8078
5/13/2015 22:30	OTHER OFFENSES	MISCELLANEOUS INVEST	Wednesday	TARAVAL	NONE	0 Block of ESCOL	-122.488	37.73767
5/13/2015 22:30	VANDALISM	MALICIOUS MISCHIEF, V/	Wednesday	TENDERLOIN	NONE	TURK ST / JONES	-122.412	37.783
5/13/2015 22:06	LARCENY/THEFT	GRAND THEFT FROM LOC	Wednesday	NORTHERN	NONE	FILLMORE ST / GE	-122.433	37.78435
5/13/2015 22:00	NON-CRIMINAL	FOUND PROPERTY	Wednesday	BAYVIEW	NONE	200 Block of WILI	-122.398	37.72993
5/13/2015 22:00	NON-CRIMINAL	FOUND PROPERTY	Wednesday	BAYVIEW	NONE	0 Block of MENDI	-122.384	37.74319

In the dataset, we have 39 types of crimes among which top 5 are theft, other offences, non-criminal, assault, and drug/narcotic. Top 5 crimes account to 66% of whole records

3. PREPROCESSING

Data preprocessing is a data mining technique that involves transforming raw data into an understandable format. Real-world data is often incomplete, inconsistent, and/or lacking in certain behaviors or trends, and is likely to contain many errors. Data preprocessing is a proven method of resolving such issues. Data preprocessing prepares raw data for further processing.

Steps involved in data preprocessing are:

Data Cleaning: Data is cleansed through processes such as filling in missing values, smoothing the noisy data, or resolving the inconsistencies in the data.

Data Integration: Data with different representations are put together and conflicts within the data are resolved.

Data Transformation: Data is normalized, aggregated and generalized.

Data Reduction: This step aims to present a reduced representation of the data in a data warehouse.

Data Discretization: Involves the reduction of a number of values of a continuous attribute by dividing the range of attribute intervals.

3.1 Splitting timestamp values

Dataset had timestamp values, which cannot be processed by statistical models. So, we split it into Year, Month, Day, Hour.

Table 2: Converted Timestamp values

Dates
2014-04-29 17:27:00
2012-11-28 19:00:00
2006-03-25 19:00:00
2004-09-28 21:55:00
2005-07-25 09:30:00
2011-05-18 21:30:00

Years	Month	DayOfMonth	Hour
2014	04	29	17
2012	11	28	19
2006	03	25	19
2004	09	28	21
2005	07	25	09
2011	05	18	21

3.2 Categorical predictors to binary predictors

Most of our predictors were categorical, which again are difficult to handle with various statistical models, hence we converted all the categorical predictors into binary predictors based on the number of categories present in every categorical predictor. There were no missing/ NA values present in the dataset.

Table 3: Categorical to Binary

Years
2014
2012
2006
2004
2005
2011
2014
2011
2012
2010

Yr.2003	Yr.2004	Yr.2005	Yr.2006	Yr.2007	Yr.2008	Yr.2009	Yr.2010	Yr.2011
0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0
0	0	0	1	0	0	0	0	0
0	1	0	0	0	0	0	0	0
0	0	1	0	0	0	0	0	0
0	0	0	0	0	0	0	0	1
0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	1

3.3 Data Partition (Training/ Test)

We faced couple of issues while partitioning the data into training and test sets. Our class predictor (Category) is very imbalanced. For e.g., one class "Larceny/ Theft" has instances up to 135,000 and another class "TREA" has only 5 instances in the dataset. So while partitioning the data we were missing out information about few classes. To resolve this issue we mostly used the whole dataset to train the model and used cross validation for resampling of dataset. This is also resolved using a function "CreateDataPartition" available in the R package "AppliedPredictiveModelling". This function makes sure that we've information about all the available classes in the class predictor.

4. EXPLORATORY DATA ANALYSIS

Exploratory Data Analysis (EDA) is an approach to analyzing **data** sets to summarize their main characteristics, often with visual methods. A statistical model can be used or not, but primarily EDA is for seeing what the **data** can tell us beyond the formal modeling or hypothesis testing task. Exploratory Data Analysis is a philosophy for data analysis that employs a variety of techniques to maximize insight to a dataset, uncover underlying structure, extract important variables, detect outliers and anomalies, test underlying assumptions, develop parsimonious models and determine optimal factor settings. EDA techniques are often simple which has various techniques of

1. Plotting raw data
2. Plotting simple statistics
3. Positioning such plots to maximize natural pattern recognition abilities, example, multiple plots per page

We extracted very useful insights by doing simple exploratory analysis on the dataset.

4.1 Crimes by its frequency

By plotting crimes by its frequency we understood which crimes are more often committed. If we observe in the plot below, its seen that Larceny/ Theft had occurred more than 130,000 times. [3]

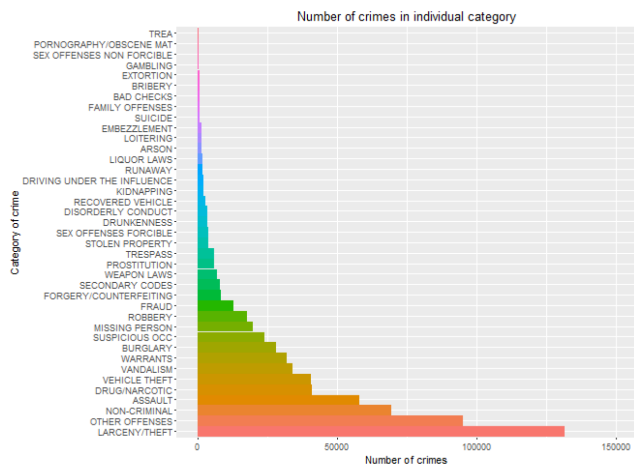

Figure 1: Crimes by its frequency

4.2 Crimes by Day of Week

Crime rates change during the week. More number of crimes are committed on Friday and opposite trend is seen on Sundays.

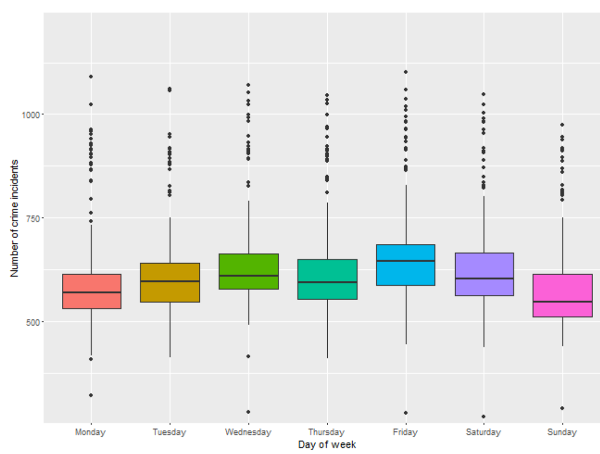

Figure 2: No. of Crimes by Day of Week

4.3 Crimes by Hour of Day

In "Hour of Day" the trend clearly evident that crimes are very lees committed (almost 0) during the night time and more often committed during mid-day and 5:00-6:00PM.

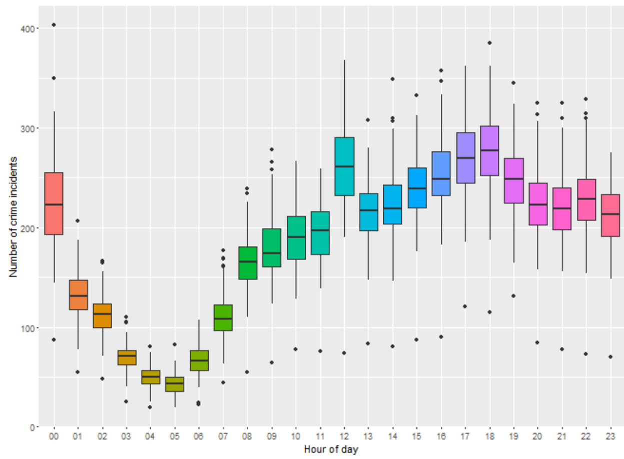

Figure 3: No. of Crimes by Hour of day

4.4 Crimes by Month

Crime rate is highest during October and lowest in December. In this plot crimes seem to follow bimodal pattern with peaks in May and October and valleys in December an August.

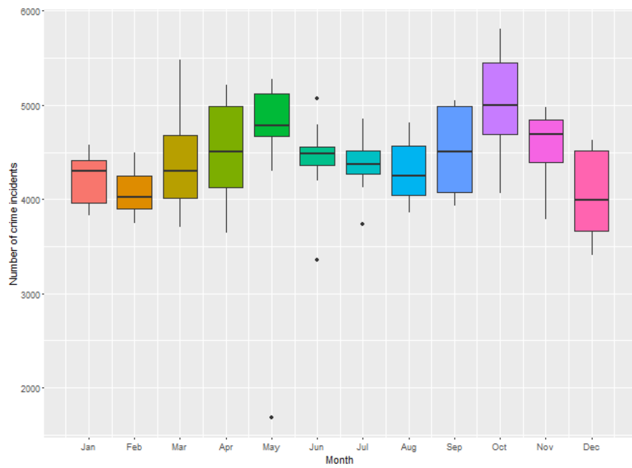

Figure 4: No. of Crimes by Month

4.5 A Consolidated plot

This a Consolidated plot of crimes by "Day of Week", "Hour of Day" and "Month".

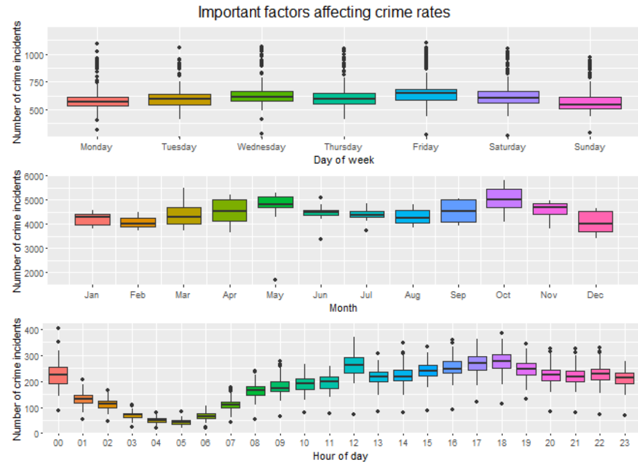

Figure 5: Consolidated plot

4.6 Google Visualization

This is a time series plot developed using google visualization package in R. It shows the trend of crimes committed from 2003-2015. If you observe the trend of Vehicle theft from 2005-2006, there is tremendous reduction in the total no. of crimes committed. That was when anti-theft systems were installed in the cars. If the trend of Larceny/ Theft is observed there is very high rise in the no. of crimes committed from 2011-2014, we did not find a reason for this growth (maybe there is one). [4]

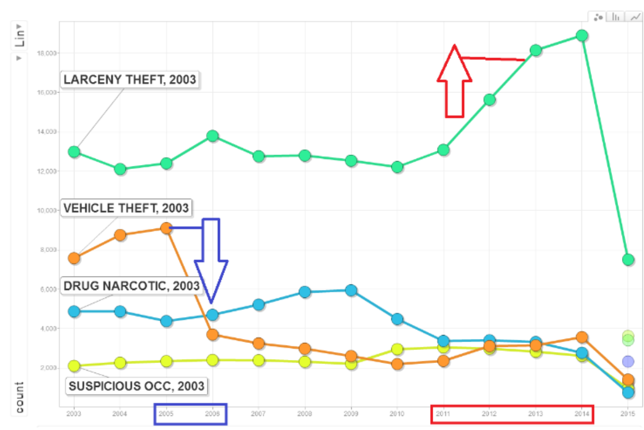

Figure 6: Time series trend in crimes

4.6.1 Bar Plot

A frequency bar plot of top ten committed crimes. This plot shows the crimes committed in descending order. This plot for the year 2007. Its seen that in any given year Larceny/ Theft is the most often committed crime.

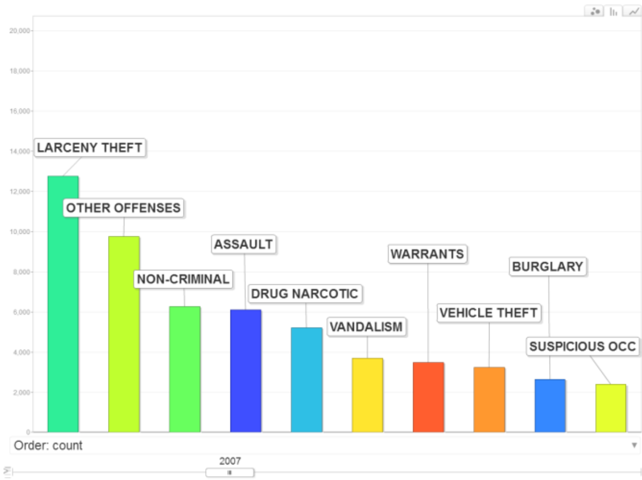

Figure 7: No. of Crimes in the year 2007

4.6.2 Line Chart

A line chart that shows the trends of top ten crimes from 2003-2015. As seen in the chart all the crimes see a down trend from the year 2014-2015, but its with the data that's available for the year 2015. We've records only till May2015.

Assault – Its one crime that has a straight line trend over the years. There is no rise/ fall in the total no. of crimes committed.

Non-Criminal – These crimes have also seen an upward trend over the years.

[4]

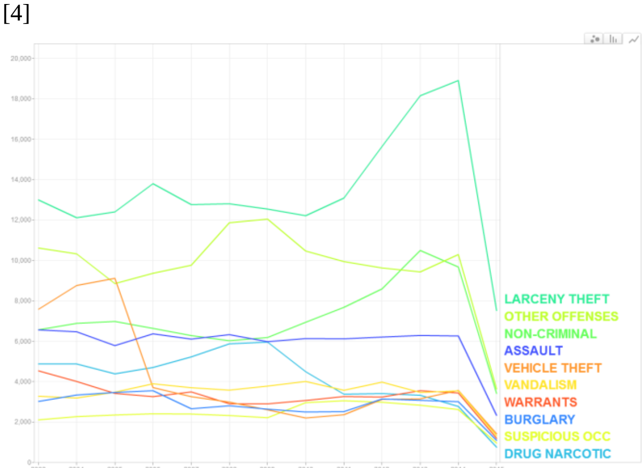

Figure 8: Line Chart of top ten crimes from 2003-2015

5. FEATURE ENGINEERING

We struggled a lot to figure the right set of predictors to train any classification model. After a lot of deliberation, we figured out few combinations of predictors to train the model. First, we used only three numerical predictors (x, y, hour) which are normalized. Second, we used six normalized numerical predictors (x, y, year, hour, month, day). Third, we created 196 binary predictors using 5 categorical predictors.

6. STATISTICAL MODELS

Before we go into details of all the statistical model applied on the dataset, we'll discuss about the ones that we submitted solutions on Kaggle.

6.1 Kaggle Score

Our current position on the leaderboard of Kaggle is 728/ 1979.

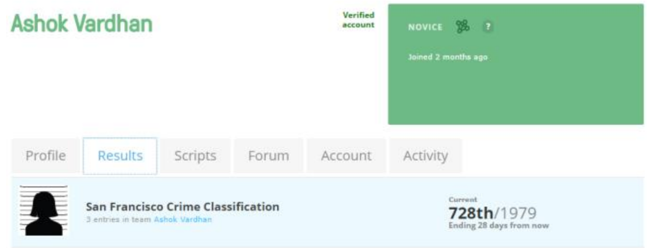

Figure 9:Our Overall position in the competition

Our highest score was 2.55788 (rank – 728/ 1979), where top score as of now was 1.9593 (rank – 1/ 1979). We got this score with model built using "L2 Regularized logistic regression (Primal)", more about it later.

Figure 10: Top Scores on Kaggle

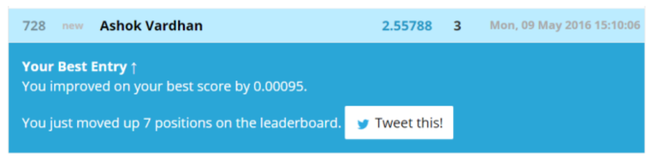

Figure 11:L2-regularized logistic regression (primal)

Our second highest score was 2.55883 with a rank of 732/ 1979. This was with model "L2-regularized logistic regression (dual)", more about it later.

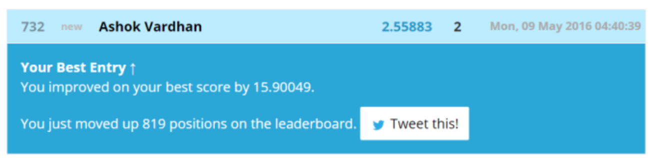

Figure 12: L2-regularized logistic regression (dual)

6.2 Linear Models

KNN – Started with basic classification model KNN (K-Nearest Neighbors) with three normalized numerical predictors (x, y, hour). We trained the model using a training set with 685,000 (approx.) instances. Applied the model on a test set with 225,000 (approx.) and got an accuracy of 59%. But when we submitted the output of model on Kaggle we got a score of 15.69 (Rank – 1558/ 2000).

Used KNN with another set of predictors (x, y, year, hour, month, day). But the accuracy turned even bad.

Regularized Logistic Regression – We used L2 (primal) with 5 numerical predictors. There is a function called "Liblinear" in Liblinear package of R using which we've trained this model. But the accuracy of the model still did not improve. [5]

Regularized Logistic Regression (Primal) – In this model we've used all the 196 binary predictors to train the model and predicted the category of crime using the test set/ Accuracy still did not improve but solution submitted on Kaggle gave a decent score of

2.55788 with rank of 728/1979. We feel this is good model with good predictive ability of category of crime.

Regularized Logistic Regression (Dual) – The application of this model is same as previous one with one difference in the arguments passed to build the model. [6]

Linear Models			
Model	**Scaled Variables**	**Accuracy**	**Kaggle Score**
KNN	X, Y, Hour	59%	15.69
KNN	X, Y, Year, Hour, Month, Day	25%	NA
Regularized Logistic Regression	X, Y, Year, Hour, Month, Day	22%	NA
Regularized Logistic Regression (Dual)	196 Binarized Variables	23%	2.5583
Regularized Logistic Regression (Primal)	196 Binarized Variables	22%	2.55788

Figure 13: Performance of Linear Models

6.3 Non-Linear Models

Extreme Gradient Boosting – XGBoost is a famous gradient boosting algorithm. This is referred as blackbox because of its intrinsic structure of algorithm. It builds m thousands of decision trees to find the most useful predictors and use only them to build the model. It's very complicated and difficult to understand by a human reader. This model gave us an accuracy of 34% (improved) from other models. [7]

We've also applied other famous classification methods such as Neural Networks, Naïve Bayes, Support Vector Machine but all the models ran for more than 10hours, so we had to manually stop the models. [8]

Non-Linear Models			
Model	**Variables**	**Accuracy**	**Kaggle Score**
Extreme Gradient Boosting	196 Binarized Variables	34%	NA
Neural Networks	196 Binarized Variables	Ran forever	NA
Naïve Bayes	196 Binarized Variables	Ran forever	NA
Support Vector Mach	196 Binarized Variables	Ran forever	NA

Figure 14: Performance of Non-Linear Models

7. CONCLUSIONS

The problem looked like a simple application of classification algorithm but while exploring the data we could uncover some interesting trends. Crime being a society based issue, by observing this data we can predict various occurrences of crime and hence relevant measures can be taken in order to reduce them. By using different visualization techniques, we discovered several patterns in the occurrence of crimes. We figured which crimes are more often committed. On which weekday more no. of crimes is committed. At what time the trend of crime goes up/ down. And using google vis we understood the trend of top ten crimes over the years and also found good reasoning for the rise/ fall of crimes.

We think there is no relationship between dates, time to the category of crime. Given date, time and place of a crime incident it's difficult to predict the category of crime. We did not find good relationship between the predictors and the response predictor.

8. INDIVIDUAL CONTRIBUTIONS

Tasks	Anusha	Ashok
Presentation	Prepared	-NA-
Data Collection	-NA-	From Kaggle
Preprocessing	Time Stamp Values	Categorical to Binary
EDA	Bar Plots	Google Vis
Statistical Models	KNN, Regularized Logistic (Primal), Naïve Bayes.	Regularized Logistic (Dual), Neural Networks, SVM

9. REFERENCES

[1] SFO Gov, "SF Open Data," [Online]. Available: https://data.sfgov.org/.

[2] Kaggle, "San Francisco Crime Classification," Kaggle, 02 June 2015. [Online]. Available: https://www.kaggle.com/c/sf-crime.

[3] V. Yadav, "Udacity_projects," 30 Jan 2016. [Online]. Available: https://github.com/vxy10/Udacity_projects/blob/master/P4_SFO_CrimeAnalysis/main_p4_at_v2_submission.Rmd.

[4] CRAN, "googleVis examples," cran-r, [Online]. Available: https://cran.r-project.org/web/packages/googleVis/vignettes/googleVis_examples.html.

[5] T. Helleputte, "Package 'LiblineaR'," 30 1 2015. [Online]. Available: https://cran.r-project.org/web/packages/LiblineaR/LiblineaR.pdf.

[6] K.-W. C. Rong-En Fan, "LIBLINEAR: A Library for Large Linear Classification," 3 July 2015. [Online]. Available: http://www.csie.ntu.edu.tw/~cjlin/papers/liblinear.pdf.

[7] T. Chen, "Understanding XGBoost Model on Otto Dataset," Kaggle, 1 April 2015. [Online]. Available: https://www.kaggle.com/tqchen/otto-group-product-classification-challenge/understanding-xgboost-model-on-otto-data/comments.

[8] hxd1011, "XGBoost Parameters," 2 Feb 2014. [Online]. Available: https://github.com/dmlc/xgboost/blob/master/doc/parameter.md.