**[San Francisco Crime Classification](https://www.kaggle.com/c/sf-crime)**

**Project status report**

*Asha Mary Thomas - axt143530*

*Biligiri Vasan S R - bxs152830*

*Somasundaram Ardhanareeswaran - sxa146230*

**Description of data:**

The dataset contains incidents derived from SFPD Crime Incident Reporting system. The data ranges from 1/1/2003 to 5/13/2015. The training set and test set rotate every week, meaning week 1,3,5,7 belongs to test set and week 2,4,6,8 belongs to training set.

Given time and location we have to predict the category of crime that occurred.

**Number of instances:**

*Training Data:* **878049**

*Test Data:* **884262**

**Brief description of features:**

*Training Data: 9 features*

* **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 - name of the Police Department District
* Resolution - how the crime incident was resolved
* Address - the approximate street address of the crime incident
* X - Longitude
* Y – Latitude

*Test Data: 6 features*

* Dates - timestamp of the crime incident
* DayOfWeek - the day of the week
* PdDistrict - name of the Police Department District
* Address - the approximate street address of the crime incident
* X - Longitude
* Y – Latitude

**Type of output variables**

Type***: Class Value***

The Output would be one of the category of crime from the training set.

The training set has 40 unique values for category.

**Experimental methodology**

**Training Classifiers:**

Following are the classifiers that we have chosen for training the model:

* Random Forest
* Bagging
* Boosting
  + Decision Trees
  + Support Vector Machines
  + Artificial Neural Networks

**Prevent overfitting:**

After initial analysis of data, we have found that of the 39 categories, around 20-22 categories contribute to about 98% of the data. Hence, we will be ignoring the categories that are less significant while training the data. This would help us in getting a more generalized model.

*Initial Analysis on Data:*

The R code for reading data and counting each category of crime is as follows:

*data <- read.csv("C:/Users/Biligiri Vasan/Desktop/Spring 2016/ML/train.csv",header = T,sep = ",")*

*df <- data.frame(data)*

*data\_table <- data.frame(table(df$Category))*

*data\_ordered <- data\_table[order(data\_table$Freq,decreasing = T),]*

*data\_ordered*

**Calculations:**

If we consider only those categories in which the count is greater than 4000, we have 22 such categories. The total of these categories is 860277. The total count of 39 categories is 878049.

If we calculate the percentage of 22 categories, then:

(860277/878049) \* 100 = 97.97%

We will be only considering the top 22 categories from the below table in the left. The right table (highlighted in grey) has the categories with minimal count

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Category | Count |  | Category | Count |
| LARCENY/THEFT | 174900 |  | RECOVERED VEHICLE | 3138 |
| OTHER OFFENSES | 126182 |  | KIDNAPPING | 2341 |
| NON-CRIMINAL | 92304 |  | DRIVING UNDER THE INFLUENCE | 2268 |
| ASSAULT | 76876 |  | RUNAWAY | 1946 |
| DRUG/NARCOTIC | 53971 |  | LIQUOR LAWS | 1903 |
| VEHICLE THEFT | 53781 |  | ARSON | 1513 |
| VANDALISM | 44725 |  | LOITERING | 1225 |
| WARRANTS | 42214 |  | EMBEZZLEMENT | 1166 |
| BURGLARY | 36755 |  | SUICIDE | 508 |
| SUSPICIOUS OCC | 31414 |  | FAMILY OFFENSES | 491 |
| MISSING PERSON | 25989 |  | BAD CHECKS | 406 |
| ROBBERY | 23000 |  | BRIBERY | 289 |
| FRAUD | 16679 |  | EXTORTION | 256 |
| FORGERY/COUNTERFEITING | 10609 |  | SEX OFFENSES NON FORCIBLE | 148 |
| SECONDARY CODES | 9985 |  | GAMBLING | 146 |
| WEAPON LAWS | 8555 |  | PORNOGRAPHY/OBSCENE MAT | 22 |
| PROSTITUTION | 7484 |  | TREA | 6 |
| TRESPASS | 7326 |  |  |  |
| STOLEN PROPERTY | 4540 |  |  |  |
| SEX OFFENSES FORCIBLE | 4388 |  |  |  |
| DISORDERLY CONDUCT | 4320 |  |  |  |
| DRUNKENNESS | 4280 |  |  |  |

Also, we plan to normalize certain fields (details under Null and Redundant Data section) in our training data to avoid overfitting.

**Testing:**

* The test dataset available does not have the category to which it should belong to.
* So we plan to use the holdout method and split our training data into validation and training set in 80:20 ratio.
* We would be using the validation set to measure the accuracy of our model and do further optimizations in our model.
* Also we plan to do k-fold cross-validation ensure the accuracy in the model
* After the model is created and evaluated, we would use the model to predict the category of crime for the test data.

**Null and Redundant Data:**

* During our initial analysis, we have read the csv files and found that there are no **null values**.
* But we have to make transformations to few data fields while building the model
  + **Dates** – Currently the Dates filed is a timestamp. We intend to split it into Year, Month, Date and Time. We would be using the R package ”lubridate” for this purpose to ensure that no invalid entries are generated while transforming the date.
  + Address – We have to normalize the address field so that the same address represents the same meaning to the classifier.

For eg: “2200 XYZ Street” and “2200 XYZ ST” represents the same address. But the classifier won’t be able to distinguish them. So we will transform it into a normalized form such that both the address means the same to the classifier.

* We plan to avoid using the fields such as “Resolution” which does not presents much useful information in predicting the category of crime and it is also not being used in the test data.

**Visualization:**

We would be using R packages such as ggplot, ggmaps, and histograms to plot the crime data in SF map and visualize results.

**Coding Techniques:**

* We would be using R to code this project.
* Plan to use various machine learning packages such as:
  + e1071
  + random forest,
  + boosting
  + bagging
  + time packages such as lubridate
  + visualization packages such as ggplot, ggmaps