Crime Prediction and Classification in San Francisco City

Addarsh Chandrasekar, Poorna Kumar, Abhilash Sunder Raj
CS 229 Project, Stanford University

Introduction

We are dealing with the problem of crime classification in San Francisco city.

The aim of our project is to analyze crime data collected from San Francisco over the past twelve years, and understand if there is a relationship between the type of crime that occurs and the time and place it occurs.

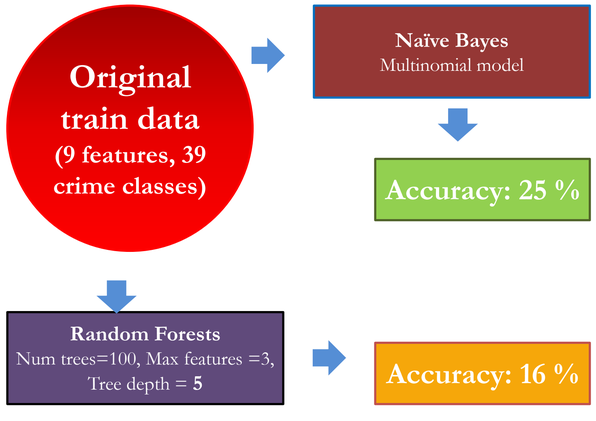
Initial Approach

- **Training dataset :** **9** features,,
  39 crime categories
  878049 training examples

- **Sample train data:**

Dates	Category	Descript	DayOfWeek	PdDistrict	Resolution	Address	X	Y
13-05-2015 23:53	WARRANTS	WARRANT ARREST	Wednesday	NORTHERN	ARREST, BOOKED	OAK ST / LAGUNA ST	-122.426	37.7746
13-05-2015 23:53	OTHER OFFENSES	TRAFFIC VIOLATION ARREST	Wednesday	NORTHERN	ARREST, BOOKED	OAK ST / LAGUNA ST	-122.426	37.7746
13-05-2015 23:33	OTHER OFFENSES	TRAFFIC VIOLATION ARREST	Wednesday	NORTHERN	ARREST, BOOKED	VANNESS AV / GREENWICH ST	-122.424	37.80041
13-05-2015 23:30	LARCENY/ THEFT	GRAND THEFT FROM LOCKED AUTO	Wednesday	NORTHERN	NONE	1500 Block of LOMBARD ST	-122.427	37.80087

- **Modelling and testing original data**



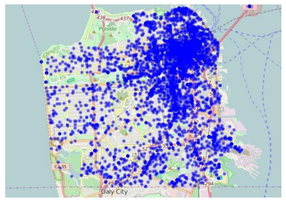
Collapse Classes

Classifying 39 classes all at once was not feasible

We collapsed all crimes into 2 major classes:

Blue Collar Crimes
Larceny, Assault, Theft, Burglary etc.

White Collar Crimes
Fraud, Forgery, Extortion, Bribery etc.



Data Enriching

- Since we had limited features in the original dataset new features were taken from the census data to further enrich our training set

- Added features include Mean Family Income, Population count, Minority Population etc. to the original training data

- Census data:

Tract Code	Tract Income Level	Distressed or Under	Tract Median Family Income %	2015 FFIEC Est. Median Family Income	2015 Est. Tract Median Family Income	2010 Tract Median Family Income	Tract Population	Tract Minority %	Minority Population	Owner Occupied Units	1- to 4- Family Units
101.00	Moderate	No	69.03	$96,900	$66,890	$64,886	3739	53.92	2016	260	424
102.00	Upper	No	154.01	$96,900	$1,49,236	$1,44,750	4143	18.95	785	943	918
103.00	Upper	No	136.32	$96,900	$1,32,094	$1,28,125	3852	39.25	1512	505	1282
104.00	Middle	No	103.39	$96,900	$1,00,185	$97,179	4545	37.01	1682	783	1580
105.00	Upper	No	160.39	$96,900	$1,55,418	$1,50,750	2685	38.36	1030	414	148
106.00	Low	No	27.68	$96,900	$26,822	$26,024	3894	65.82	2563	208	927

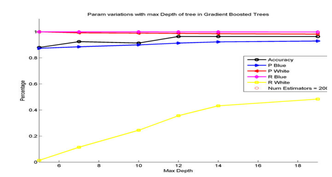
- The training data after enriching has 19 new features (an increase of 10)

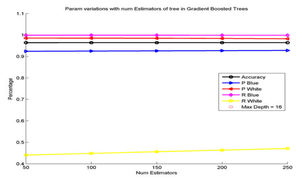
Results

The final training data is run using 3 different learning algorithms:

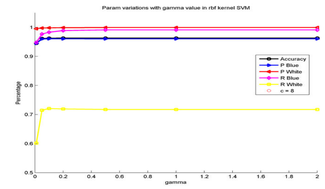
- Random Forests
- Gradient Boosted Decision Trees
- Support Vector Machines

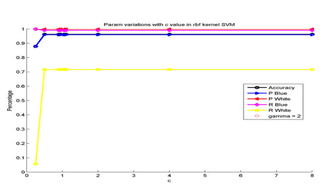
- Gradient Boosted Decision Trees



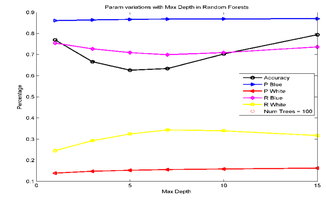


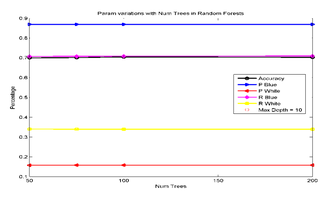
- 'RBF' Kernel SVM





- Random Forests





Random Forests
Accuracy = 79.18%
P_Blue = 0.87, P_White = 0.16
R_blue = 0.74,,R_white = 0.31

Gradient Boosting
Accuracy = 96.4%
P_Blue = 0.96, P_White = 0.97
R_blue = 0.99,R_white = 0.73

'RBF' SVM
Accuracy = 96%
P_Blue = 0.96, P_White = 0.99
R_blue = 0.99,R_white = 0.71