

# San Francisco Crime Resolution Prediction

## Problem and Objectives

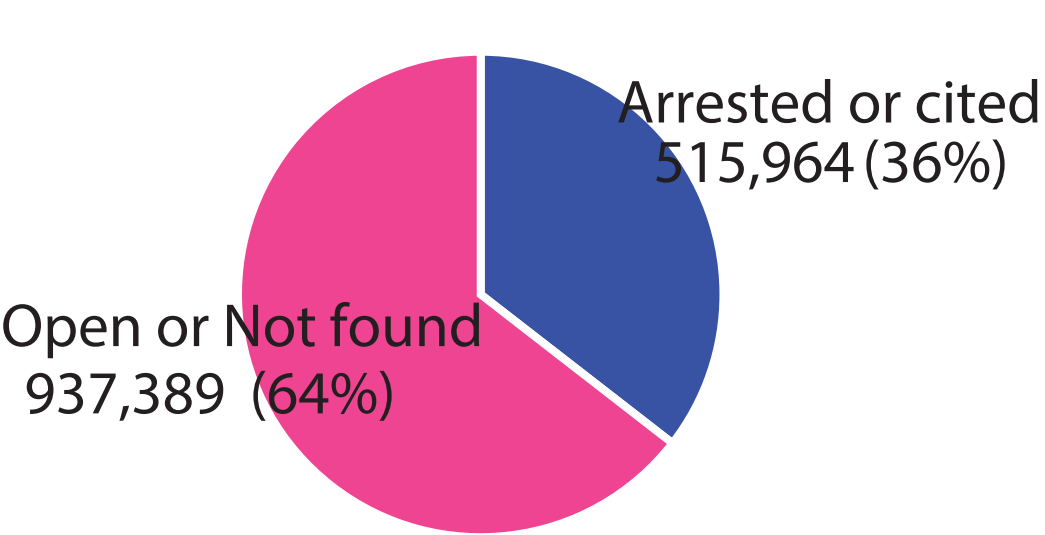
San Francisco is one of the most attractive cities for tourists, investors and entrepreneurs. Thus, the public safety is one of the most important issue people concern about. Except for the number of instances or type of crimes, the resolution performance is also a vital indicator to determine whether it is a safe place to visit. The research is to predict the resolution rate based on crime category, police district and incident time. With this prediction, we could help those who are visiting the city to avoid danger at specific area and time, and therefore ease local authority allocate their resources properly.

## Data Description

The dataset consists of 13 features and has about 1,358,710 crime records in San Francisco from 1/1/2008 to 12/31/2017 (10 years in total). Its columns are crime category, time (day of week, date, time), location (district, address, longitude, latitude), incident descriptions and resolution.

## Descriptive Analysis

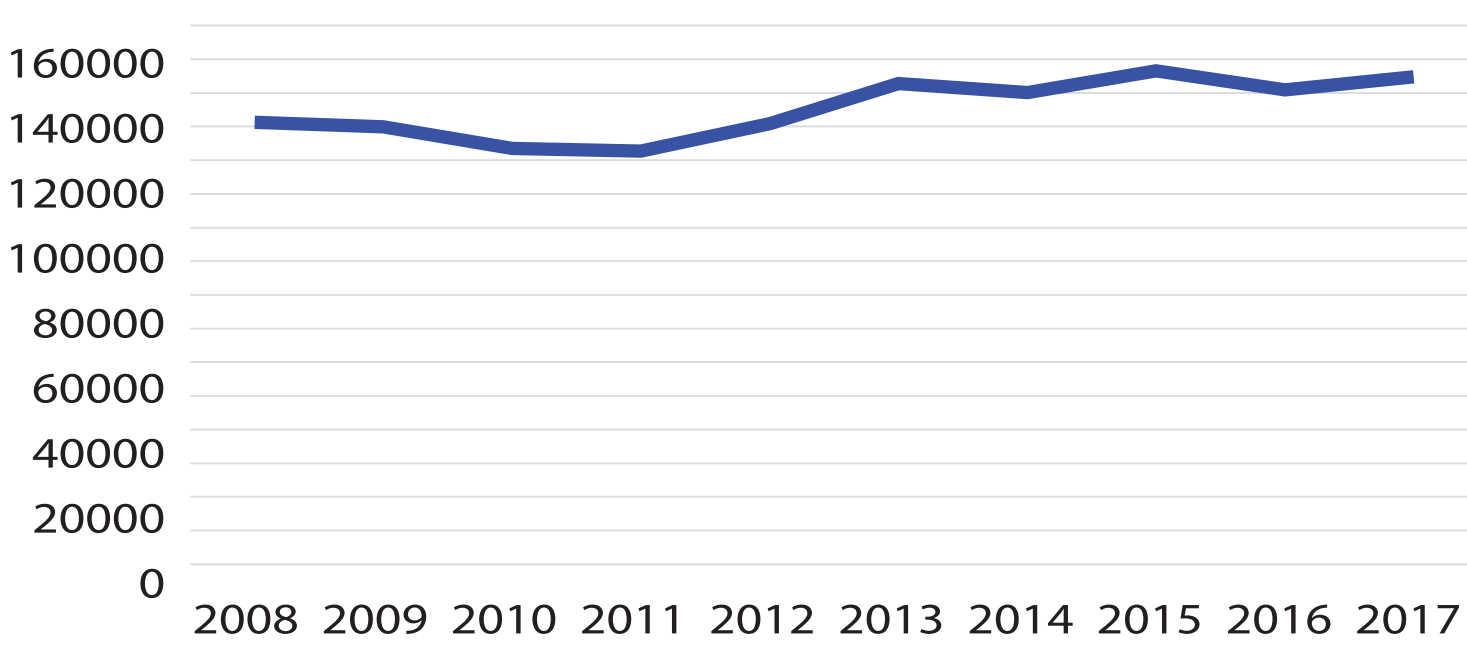
Target Attribute distribution



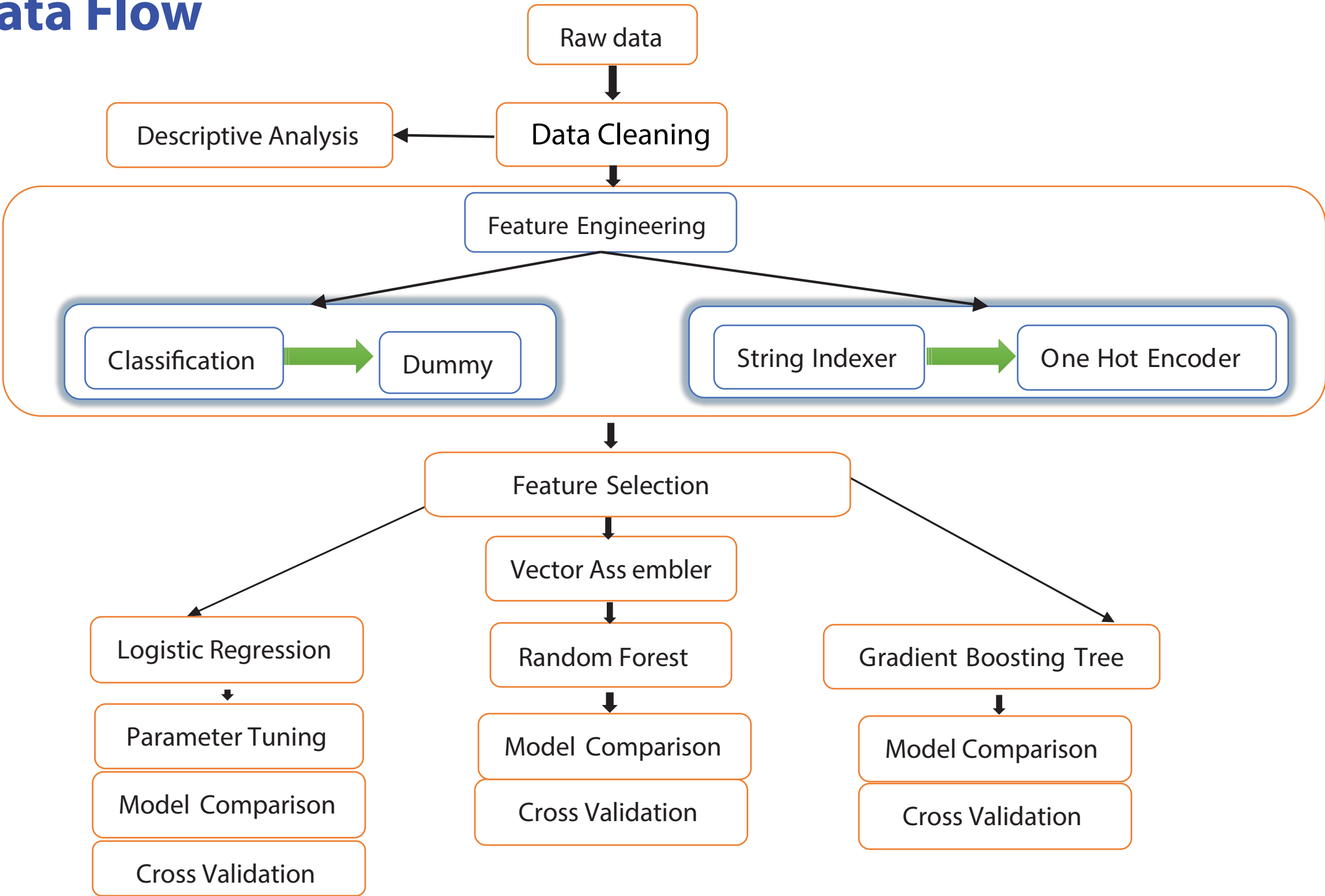
Top 10 Crimes in 10 Years



The Number of Crime from 2008 to 2017



## Data Flow



## Model Description

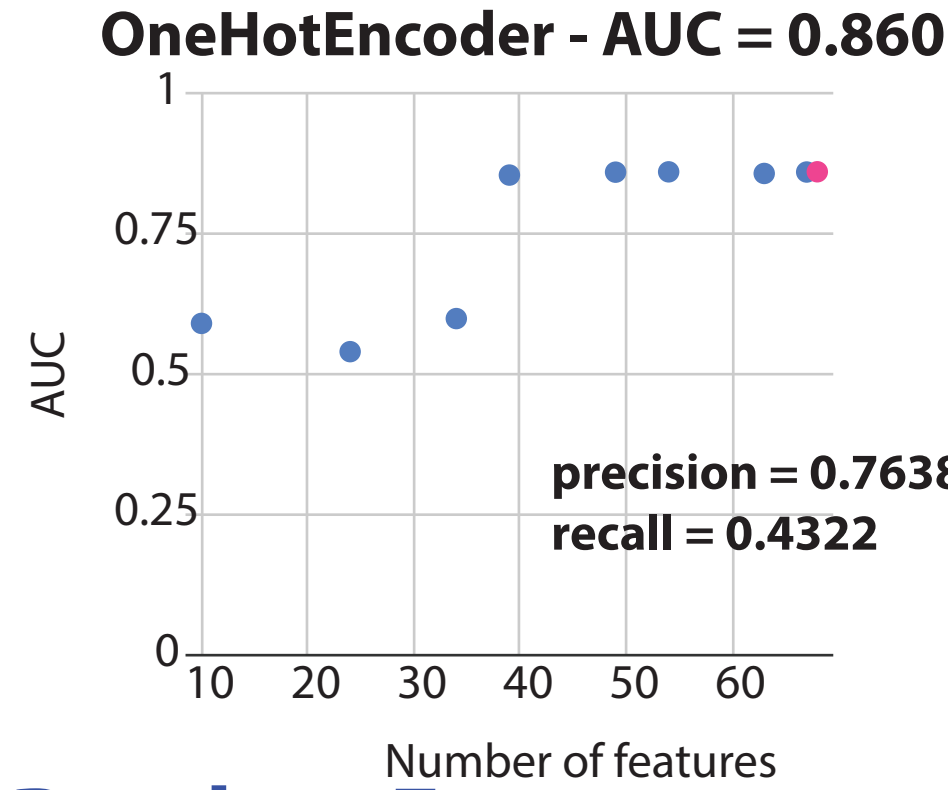
Techniques	Data Pre-processing	Model Name	Parameters	Metrics
Label-dum my variable	Dummy variables for categorical	Random Forest (RF)	numTrees=10, cacheNodeIds = True	AUC
	Group into smaller categories	Gradient-boosted tree classifier	maxIter=10	AUC
	Vector Assembler	Logistic Regression (LR)	maxIter=10; StandardScaler; Regularization (L1, Elastic)	AUC
Python built-in tools	StringIndexer	Random Forest (RF)	numTrees=10, cacheNodeIds = True	AUC
	OneHotEncoder	Gradient-boosted tree classifier	maxIter=10	AUC
	Vector Assembler	Logistic Regression (LR)	maxIter=10; StandardScaler; Regularization (L1, Elastic)	AUC

## Model Details

### OneHotEncoder

M2 : Crime catagory, days of week, locations, month, hour range  
M5 : only Crime category  
M6 : only locations  
M7 : day of week, hour range, months  
M8 : day of week, hour range, crime catagory,month  
M9 : day of week, hour range, locations, month  
M10 : crime catagory, locations  
M13 : crime category, locations, hour range  
M15 : crime category, locations, months, daysOfWeek

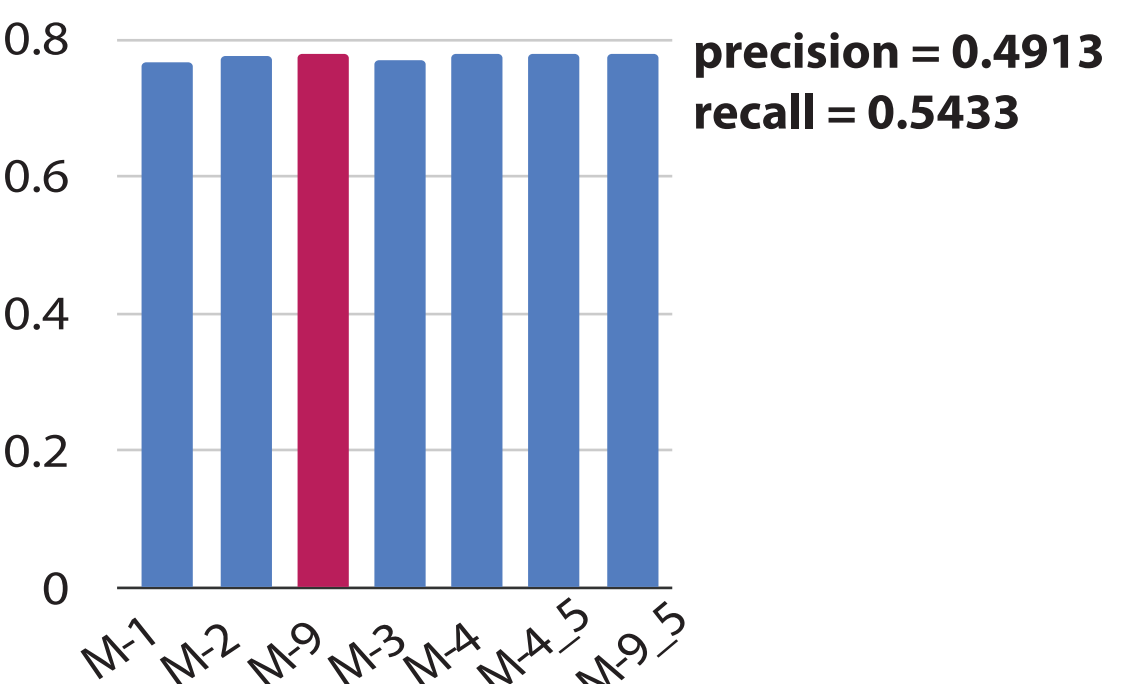
### Logistic Regression



### Label-Dummy Variable

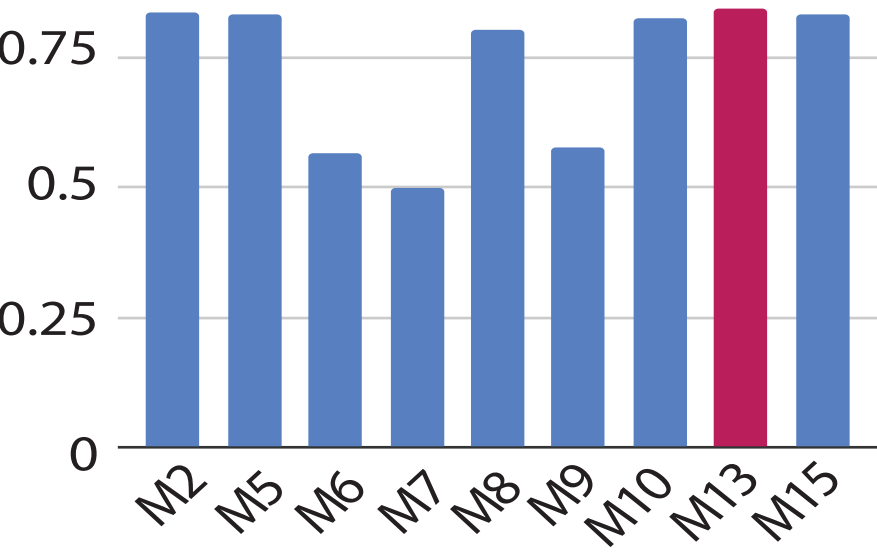
M-1 : only crime catagory  
M-2 : crime category, locations  
M-9 : crime category, locations, month, hour level  
M-3 : crime category, days, month, hour level  
M-4 : crime category, days, month, hour level, location  
M-4\_5 : crime category, days, month, hour level, location  
M-9\_5 : crime category, locations, month, hour level

### label-dummy - AUC = 0.764



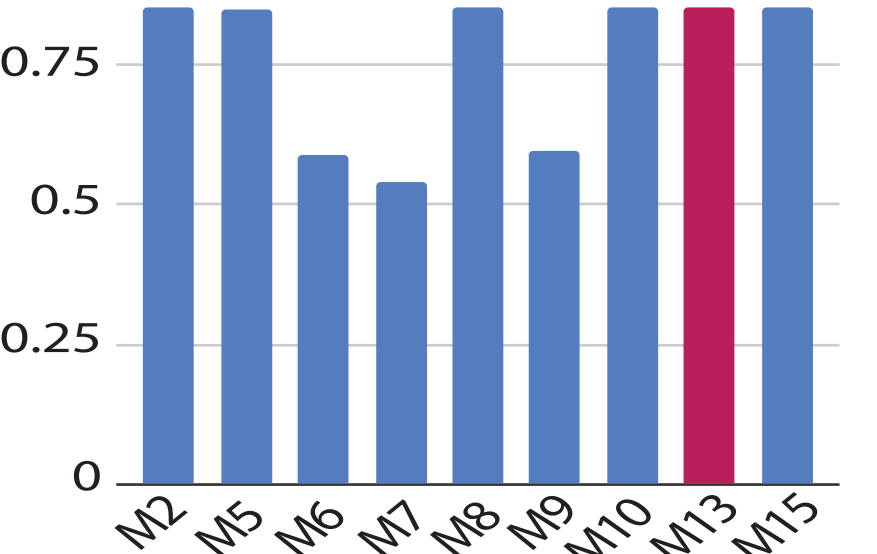
## Random Forest

### OneHotEncoder - AUC = 0.846



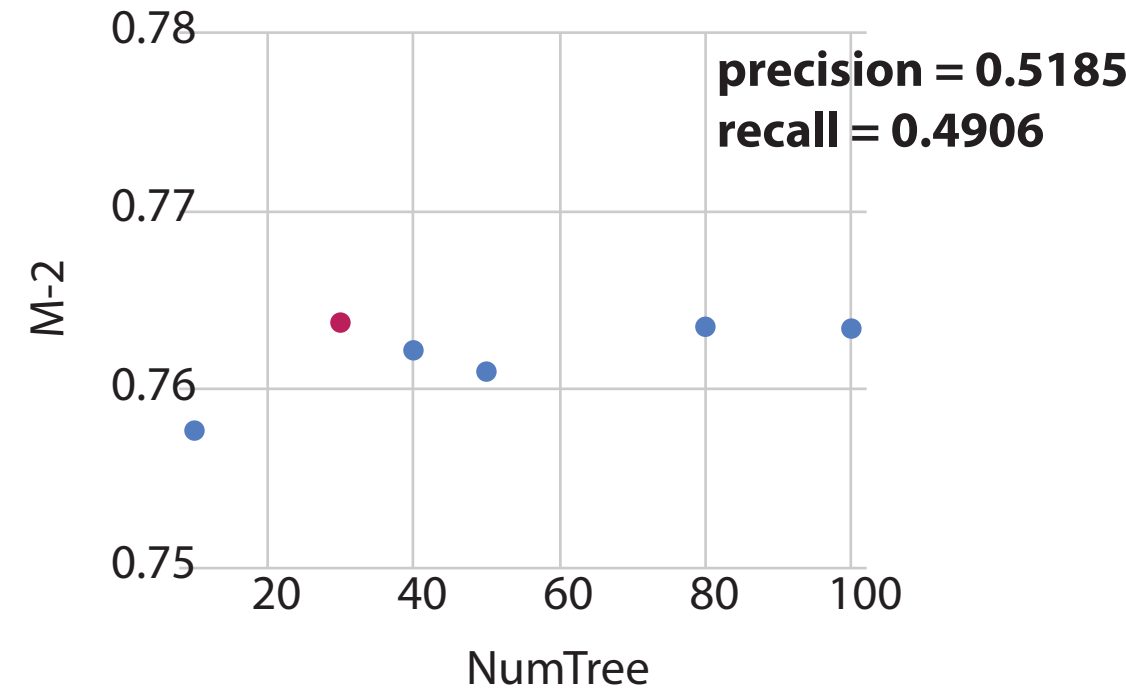
## GBTClassifier

### OneHotEncoder - AUC = 0.854

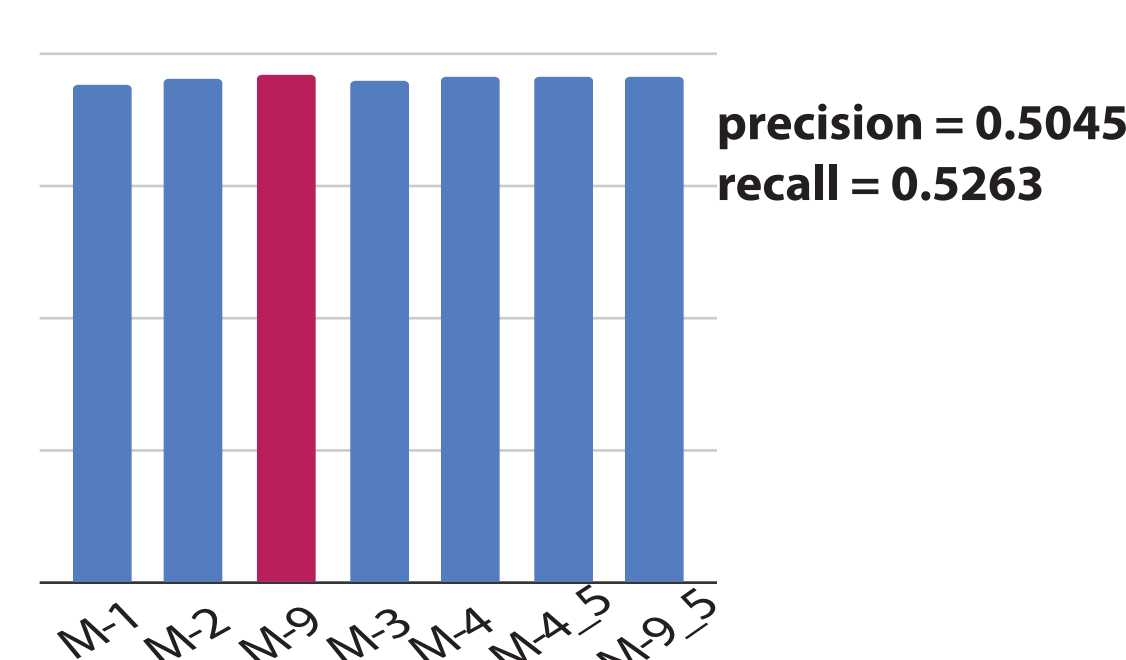


### label-dummy - AUC = 0.764

Random Forest Number of Tree vs. AUC



### label-dummy - AUC = 0.766



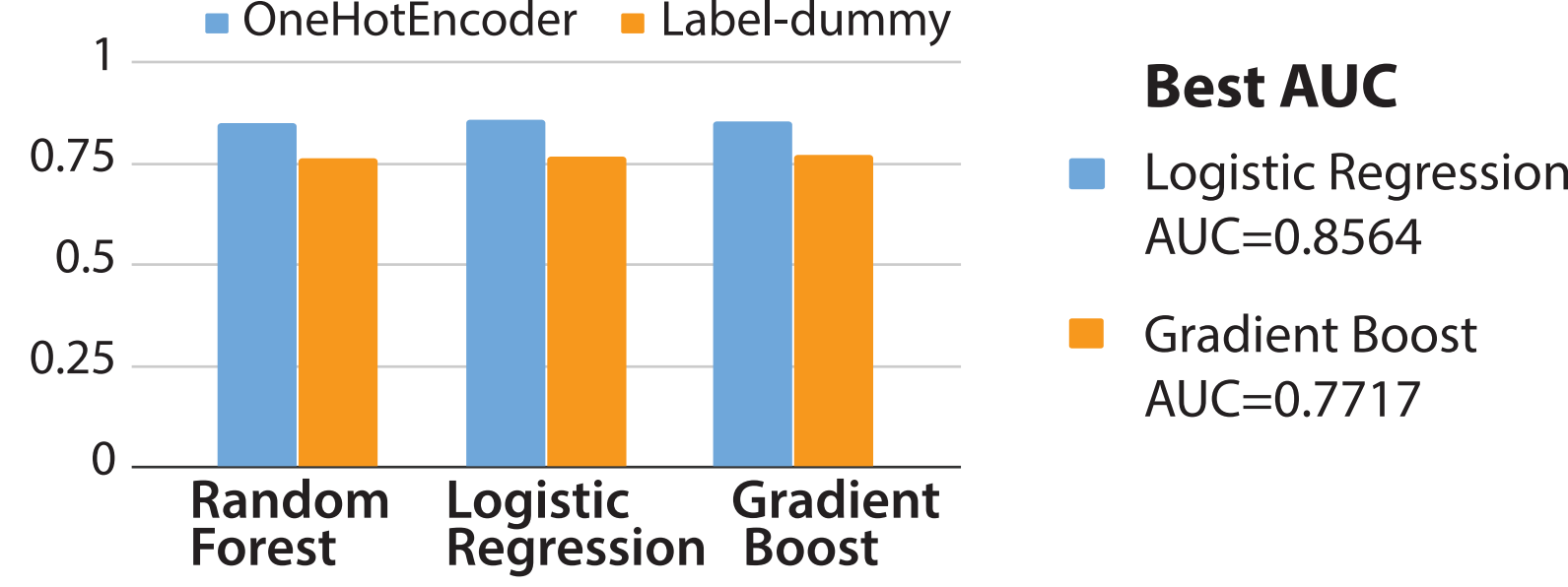
## Model Comparison Metrics

Leave-one-out cross validation is used in the experiment. Data points were splited by year, treated as time-series data.

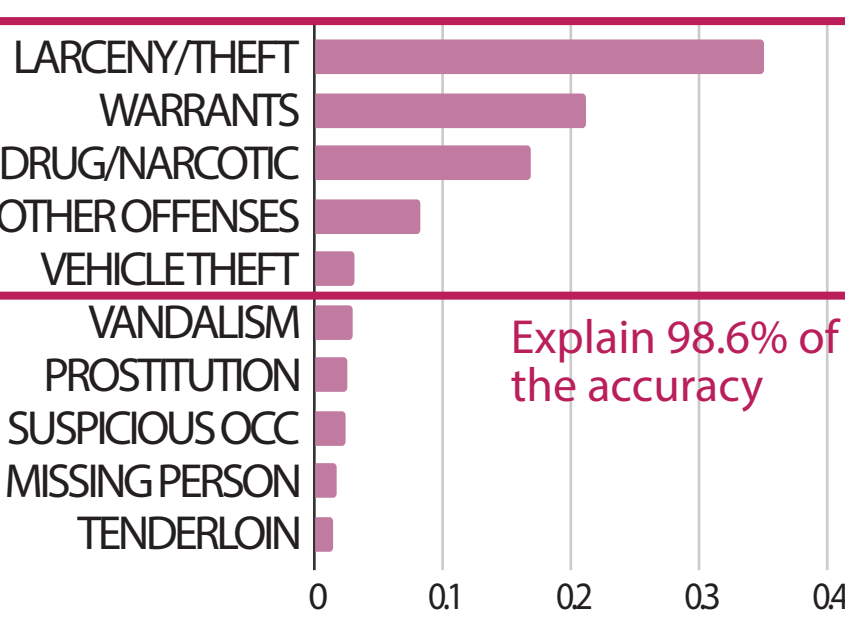
- **Training dataset** are data points with year 2008 to year 2013 (60% = 841,044)
- **Validation dataset** are data points with year 2014 to year 2016 (30% = 457,536)
- **Test dataset** are data points with year 2017 and after (10% = 154,773)

## Results - Prediction performance

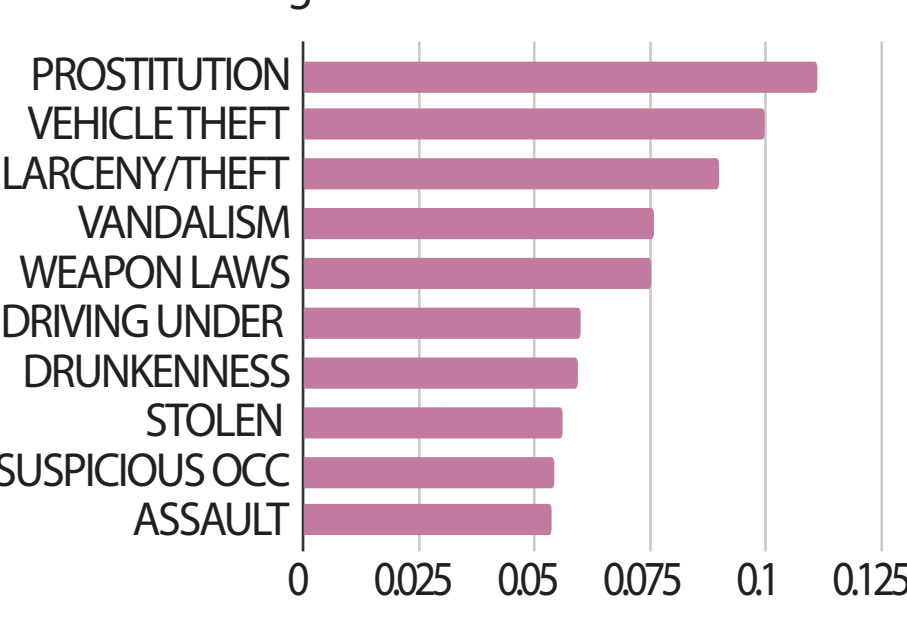
### Model Generalized Performance (test data)



### Top10 most important features - Random Forest -

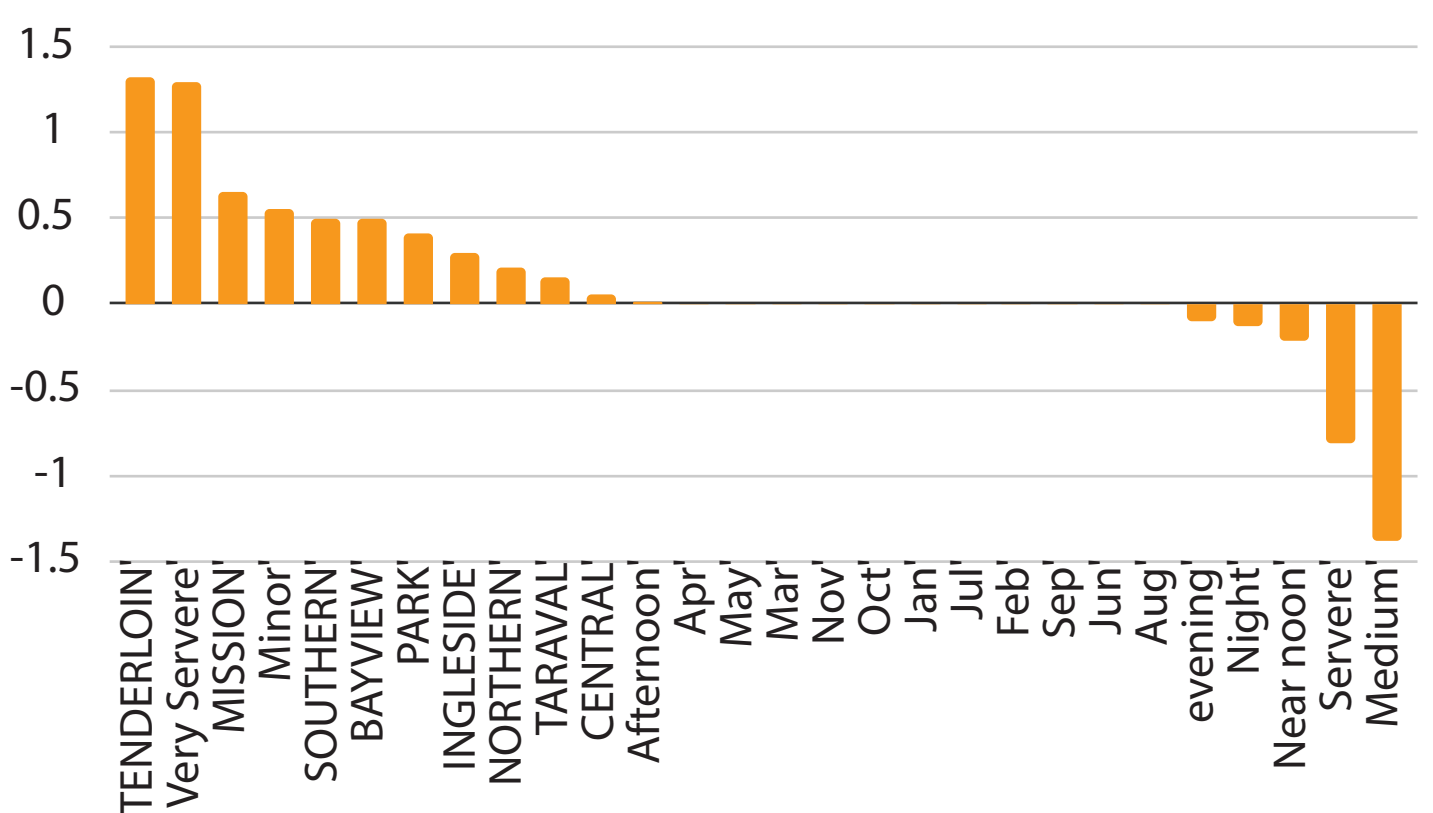


### Top 10 important features - gradient boost -



## Inference

### Coefficient of Logistic Regression



## Conclusion

Medium type of crime (include theft, missing person, sex offenses etc.), crime happened at Richmond and crimes happened near noon are the least likely to be solved. Therefore, in order to increase the rate of resolution, we suggest that the local authority to allocate more resources at Richmond and Central (top two low resolution place) and arrange more police patrol during 9-12 will increase the rate of resolution.