CSE 563
Software Requirements and Specifications

Fall 2015

Course Project

Business Proposal
San Francisco Crime Analysis Unit

Japa Swadia
MS in Software Engineering
Arizona State University
ASU ID 1207488616

Contents

Introduction to Business Problem

San Francisco, in northern California, is a city on the tip of a peninsula surrounded by the Pacific Ocean and San Francisco Bay. It's known for its hilly landscape, year-round fog, iconic Golden Gate Bridge, cable cars and colorful Victorian houses. A significant piece of history associated with the city is the island of Alcatraz, that had military and federal prison known for housing notorious criminals back in the sixties.

For a long time now, the city is most popularly known for its tech-scene and well-coined as 'Silicon Valley' across the world. In spite of this, the crime scene in the city is on a rise because of factors such as its metropolitan nature, demographics, income disparities, housing conditions and public transportation channels like the BART, MUNI, etc. The city and county of San Francisco constantly engages in crime control and prevention through various programs for community safety. For achieving crime-free zones, the San Francisco Police Department (SFPD) has a special division- Crime Analysis Unit. They collects data from crime reports in different areas of the city along with other useful informative attributes. This raw data can be gleaned by associated Data Analysts who are interested in gaining actionable information out of this and can consequently be of assistance to the city police in providing 'insights' on future crime occurrences - thereby putting a leash on incidences across the city, making it a safe place to live in.

Since Big Data is on a rise, employing models like these, will be of high significance for governing bodies for efficient public administration. The real challenge lies in coming up with a predictive model that, with a good amount of accuracy, predicts the type of crime that is likely to occur depending on the area in question. Furthermore, the model can be useful in representing the information gained, visually - in the form of data maps that effectively delineate different areas in terms of crime statistics.

Problem and Data Description

This section provides a detailed description of the data set that will be used for analysis of the business problem, along with the problem statement.

Problem Statement

Given a data set of crime records belonging to the city of San Francisco, apply data science techniques in order to predict the category of crime that occurred in a given area and given time; estimate the number of incidences of a particular category by area, and find associations between different crime types in a particular area .

About the Dataset

This dataset contains incidents derived from SFPD Crime Incident Reporting system. The data ranges from 1/1/2003 to 5/13/2015. It has both training and test data. It can be downloaded from the following link:
https://www.kaggle.com/c/sf-crime/data

It is published by SF OpenData, the City and County of San Francisco's official portal for open data. It is further sourced by Kaggle, which hosts online competitions on data analysis. The training data set has **878,050** records and **9** attributes as described below:

1. **Dates** - timestamp of the crime incident
2. **Category** - category of the crime incident (only in train.csv).
3. **Descript** - detailed description of the crime incident
4. **DayOfWeek** - the day of the week
5. **PdDistrict** - name of the Police Department District
6. **Resolution** - how the crime incident was resolved
7. **Address** - the approximate street address of the crime incident
8. **X** - Longitude
9. **Y** - Latitude

Proposed Solution

This section essentially describes some of the techniques that will be employed by data scientists working on this problem statement to achieve the objectives of the business problem.

Objectives

As formulated in the problem statement, the data set is to be analyzed to derive useful 'insights' as explained below:

1. Predict the category of crime based on given location and given time.
2. Estimate number of annual crime incidences of a given category according to location.
3. Finding which crimes occur the most together in a particular area, or which type of crime is most likely to take place given one type of crime in that area. This subproblem can be further extended to determine 'top' crimes in that area, and 'top' days for the crime to occur.

Techniques

Following data mining techniques are being aimed at for solving the business objectives mentioned above. Each of these corresponds to the objective states above.

1. Classification: *Supervised*, Target variable - **Category**
2. Regression: *Supervised*, Target variable - **Category**
3. Co-occurrence grouping/ Association mining: *Unsupervised*, No target variable

Business Understanding

This section contains a set of questions that data analysts should have in mind when considering a data mining project, to have full understanding of the business problem in question.

- **What exactly is the business problem to be solved?**

  As stated in the problem description and statement, the business problem that needs solving is that of San Francisco city's crime occurrences. The past crimes data collected by SFPD has to be analyzed for determining useful information that can serve to aid them in their mission to curb crime incidents across different areas of the city. The goal is to predict type of crime that can occur in an area based on time, estimating number of such crimes and finding associations between different crime types for knowing which ones occur most frequently.

- **Is the data science solution formulated appropriately to solve this business problem?** (NB: sometimes we have to make judicious approximations.)

  I believe the data science solution is formulated appropriately without having to make any approximations. The data presented is pretty straightforward in nature, coming from trustworthy, official source; has well defined attributes to help derive the solution. The data science techniques discussed before (Classification, Regression, Association mining) will be applied on this problem in order to gain necessary information about the amount and type of crimes that can take place in future, as well as a projection of crimes that can occur more frequently.

- **What business entity does an instance/example correspond to?**

  *The business entity corresponding to this instance lies in the domain of **Public Safety**.*
  Here's an example record from the data set:
  {5/13/2015 11:53:00 PM, WARRANTS, WARRANT ARREST, Wednesday, NORTHERN, ARREST, BOOKED, OAK ST / LAGUNA ST, 122.4258917, 37.7745986}

- **Is the problem a supervised or unsupervised problem?**

  *The main part of the problem is **supervised**, that is, with a defined target variable.*

—If supervised,
—Is a target variable defined?
*Yes. The target variable being **Category** of crime incident.*
—If so, is it defined precisely?
Yes, the category of crime is well-defined in the dataset, with no missing values. It will be subjected to Classification and Regression techniques in the proposed solution.
—Think about the values it can take.
It will take descriptive values, that can be any of the following: {ASSAULT, WARRANTS, VEHICLE THEFTS, OTHER OFFENCES, LARCENY/THEFT, ROBBERY, VANDALISM, SUSPICIOUS OCC, NON-CRIMINAL, MISSING PERSON, BURGLARY, TRESPASS, SECONDARY CODES, FRAUD, DRUG/NARCOTIC, WEAPON LAWS, FORGERY/COUNTERFEITING, STOLEN PROPERTY, DRUNKENNESS, DISORDERLY CONDUCT, SEX OFFENCE, KIDNAPPING, PROSTITUTION, LOITERING, GAMBLING}

- **Are the attributes defined precisely?**

  All attributes in the data are well defined, with no missing values.
  —Think about the values they can take.
  Refer to dataset description on previous page. The Date attribute takes timestamp values; Category, Description, Address, PdDistrict, Resolution and Day of week takes subjective/descriptive values while the location coordinates (X and Y) take numerical values.

- **For supervised problems: will modeling this target variable improve the stated business problem? An important subproblem? If the latter, is the rest of the business problem addressed?**

  For this business problem, modeling the target variable 'Category' will definitely lead to an improvement in the problem. As discussed before, we wish to gain knowledge of the type of crime that can occur in a neighborhood. Given the historical data, the model aims to find which type of crime is most likely to occur in that area. This information is without a doubt, profitable for the crime-watchers of San Francisco who can take precautionary measures and to an extent 'stop' the incident from occurring. We also aim to estimate the amount of incidences that can occur in a year, which can help SFPD improve their crime statistics efficiently and deploy preventive measures in less time.

- **If unsupervised, is there an "exploratory data analysis" path well defined? (That is, where is the analysis going?)**

A small chunk of this business problem is basically a 'futuristic' endeavor to predict based on a given type of crime, which other crimes are likely to take place in that locality, which areas are most prone to a crime and which days of the week are 'hot-favorite' for criminals. We aim to train the model to discover association patterns among different crime types for determining frequently occurring crimes in a neighborhood. This will involve application of various unsupervised techniques, like association mining and to some extent, clustering. Although unsupervised and purely based on judgements delivered by the model, the exploratory data analysis path is well-defined as we have a firm understanding about the kind of knowledge we wish to gain from the data. The analysis would then proceed by employing relevant co-occurrence grouping technique to derive insights that might be surprising or results that are not quite significant for official usage.

Use Case Specifications

This section contains use cases constructed for the proposed solution.

Use Case 1: Predict the Category of crime based on given area and time.

Objective: The objective of this use case is for the system to predict the type of crime incident that can occur based on a given location and time, using *Classification* technique.
Primary Actor: SFPD Crime Analysis Unit official.
Dependencies: None
Trigger: Predict type of crime in an area at a particular time.
Preconditions:

1. ID is entered in the test data
2. Date and DayOfWeek is entered in the test data
3. PdDistrict is entered in the test data.
4. Address is entered in the test data.
5. X and Y coordinates of the address is entered in the test data.

Post Conditions:

1. SFPD officials feed data into the model/system.
2. System will predict the type of crime based on the data fed.

MAIN SUCCESS SCENARIO

1. SFPD officials will fill in address and time appropriately.
2. Based on the data fed, the model will use classification technique to predict category of crime.

VARIATIONS
Variation ID: *1.1*

1. SFPD officials fed only some of the data features.
2. The system needs more training data to predict results accurately.
3. The system encounters issue in processing large amount of data.

FAILURE VARIATIONS
Variation ID: *1.1-F*

1. The system fails to train the data in the beginning itself.
2. The system is unable to process data and aborts operation.

BUSINESS RULES

1. The addresses entered into the system should match with those in the official city address records, and should be in San Francisco county ONLY.

Use Case 2: Estimate the annual amount of crimes by location.

Objective: The objective of this use case is for the system to estimate the amount of crime incidents that occur annually based on a given location, using *Regression* technique.
Primary Actor: SFPD Crime Analysis Unit official.
Dependencies: Extends Use Case 1.
Trigger: Predict number of crimes occurring in an area annually.
Preconditions:

1. ID is entered in the test data
2. Date and DayOfWeek is entered in the test data
3. PdDistrict is entered in the test data.
4. Address is entered in the test data.
5. X and Y coordinates of the address is entered in the test data.
6. Category of crime is predicted by the model as in Use Case 1.

Post Conditions:

1. SFPD officials feed address data into the model/system.
2. System will estimate the number of crimes occurring annually.

MAIN SUCCESS SCENARIO

1. SFPD officials will fill in address data properly.
2. Based on the data fed and category of crime predicted , the model will use regression technique to project annual amount of crime incidents.

VARIATIONS
Variation ID: *2.1*

1. SFPD officials did not feed data properly.
2. The system needs more training data to regress accurately.
3. The system encounters issue in processing large amount of data.

FAILURE VARIATIONS
Variation ID: *2.1-F*

1. The system fails to train the data in the beginning itself.
2. The system fails to classify the category of crime.
3. The system is unable to process data and aborts operation.
4. If the system crashed, it begins processing again.

BUSINESS RULES

1. The addresses entered into the system should match with those in the official city address records, and should be in San Francisco county ONLY.

Use Case 3: Discover which types of crime can occur most frequently in a given area.

Objective: The objective of this use case is for the system to discover top crime incidents that can occur based on a given location and time, using *Co-occurrence grouping* technique.
Primary Actor: SFPD Crime Analysis Unit official.
Dependencies: None
Trigger: Discover type of crimes that can most frequently co-occur in an area at a particular time.
Preconditions:

1. Training data properly fed into the system.

Post Conditions:

1. SFPD officials feed data into the model/system.
2. System will attempt to discover associations between types of crimes happening based on area data fed.

MAIN SUCCESS SCENARIO

1. SFPD officials will fill in all feature values properly.
2. Based on the data fed, the model will use co-occurrence grouping/association mining algorithm to discover association patterns between types of crime occurrences in an area.

VARIATIONS
Variation ID: *3.1*

4. SFPD officials fed only some of the data features.
5. The system needs more training data to apply unsupervised mining methods accurately.
6. The system encounters issue in processing large amount of data.

FAILURE VARIATIONS
Variation ID: *3.1-F*

1. The system fails to train the data in the beginning itself.
2. The system is unable to process data and aborts operation.
3. The system fails to discover association patterns between crimes.
4. On encountering failure, the system begins analysis from the start.

Data Preparation & Understanding

This phase of the project deals with getting to know more about the data involved in the business problem. It aims to identify information associated with the attributes, values and target variables belonging to the data set in order to build an efficient solution model/system.

(Data set recap : San Francisco Police Department crime records ; detailed description in Phase 1 deliverable.)

Following are the questions which need to be answered for any data science problem, keeping in mind the nature of underlying data:

• Will it be practical to get values for attributes and create feature vectors, and put them into a single table?

Yes. For the data set corresponding to this business problem, it is possible to get values and attributes and create feature vectors to be fed into the model for evaluation. The data set comes from reliable, official source and defined very clearly and precisely - in CSV format that can be converted into an MS Excel table or any other table which can be loaded into the system as feature vector and can be analyzed.

• If not, is an alternative data format defined clearly and precisely? Is this taken into account in the later stages of the project? (Many of the later methods/techniques assume the dataset is in feature vector format.)

There is no need for an alternative format as such, mainly because the data is available 'flexibly'- in CSV format, which can be easily converted into an excel table without any loss of data.

• If the modeling will be supervised, is the target variable well defined? Is it clear how to get values for the target variable (for training and testing) and put them into the table?

The modeling will be supervised for most of the problem, and yes the target variable is clearly-defined, without ambiguities or missing values. The clarity of target variable definition extends to the fact that both training and test data are provided by the source of data. So it will be pretty straightforward to put them into the table (they are already in tabular form) for analysis.

• How exactly will the values for the target variable be acquired? Are there any costs involved? If so, are the costs taken into account in the proposal?

The target variable for the business problem in question is 'Category' of crime occurred in a particular location. The training and test data sets already contain these values, which were acquired from recording of crimes that took place in the city - that is from historical records belonging to SFPD. Values can be added to this dataset once it has been trained over the training set, as and when the crime takes place.

No, there are no costs involved in acquiring the target variable values in this case. The reason being, the records come from an open data set published by the SF city council. Its free and open to interpretation for anyone who wishes to play around with it.

Therefore, no costs are taken into account in the proposal.

• Are the data being drawn from the similar population to which the model will be applied? If there are discrepancies, are the selection biases noted clearly? Is there a plan for how to compensate for them?

Yes, the test data is being drawn from similar population to which the model will be applied. As mentioned in the business proposal, the data set is obtained from kaggle.com and belongs to San Francisco Police Dapartment's Crime Analysis Unit. All data were recorded at the Police Department by officials when the crime occurred. Hence, there are no discrepancies in both training and test data sets. Therefore there is no question of having a compensation scheme.

Modeling

After analyzing the business problem and understanding the data involved comes the Modeling phase where raw data is to be 'modeled' into a framework for generating essential information for solving the business problem. A model was created in **R** statistical programming language for analyzing our data and applying techniques listed in the previous sections. Following are the questions which need to be addressed in order to develop an efficient model:

• **Is the choice of model appropriate for the choice of target variable?**
—Classification, class probability estimation, ranking, regression, clustering, etc.

The target variable here is Category of crime. The model chosen applies classification technique on the training set, by first attempting a few varieties of algorithms available. Since the target variable is of categorical nature, we applied Random Forest classification, Support Vector Machine (SVM) classifier and K- means classification. Out of these, SVM worked the best for classifying test data records into appropriate crime categories. Neural-network technique was time-intensive and low on accuracy. K-means failed to evaluate 'distances' between data points as a consequence of data being extensively categorical and textual, while random forest could not perform at all because of Category containing too many levels. Text classification was also attempted, but as the category is to be determined based on Address and day/date, it would not be feasible to use Text classification as there is no semantics associated which could predict category based on 2 to 3 independent attributes containing more than 5000 levels. SVM was, therefore, deemed fit for the problem on hand. Linear Regression was also achieved on Category against day of week as well as by year. Although this technique is the standard regression technique and reaps useful insight into the crime data, it is practically not advisable for large data sets like the one in question. Logistic regression and Poisson regression are not taken into consideration since there are no binary decisions to be made and there are no continuous variables here, respectively. Unsupervised technique of Association rule discovery was applied using Apriori algorithm in lieu of finding support and confidence of different attributes affecting crime in the city.

The model is sufficiently appropriate for training this data in the initial learning phases of the business problem solution. However, since there is a lot of textual data and no numerical data involved, the most appropriate modeling would be one based on text mining techniques.

• **Does the model/modeling technique meet the other requirements of the task?**
—Generalization performance, comprehensibility, speed of learning, speed of application,

amount of data required, type of data, missing values?

Yes. The task basically aims to predict category of crime and apply unsupervised learning. Since the size of data set is really large not all classification techniques could reap actionable results. Moreover, as mentioned before, all data is explicitly categorical and textual which leaves no scope for numerical or binary predictions which are simpler to compute. SVM performed with high accuracy in predicting category of crime based on attributes like day and address, also with a better speed than others.

—Is the choice of modeling technique compatible with prior knowledge of problem (e.g., is a linear model being proposed for a definitely nonlinear problem)?

As explained in the preceding question, choice of modeling technique plays a critical role in analyzing data in the best way possible by the virtue of generating actionable insights. For each of the three techniques used (Classification, Regression, Association rule mining), different algorithms were tried based on nature of target variable and independent variables to make it compatible with the problem to be solved and type of data.

• Should various models be tried and compared (in evaluation)?

Yes, I believe different models should be tried and compared against each other in order to determine which yields the best results in terms for the business problem in question, just like the current phase in which different methods were tried and compared - only to learn the most appropriate ones.

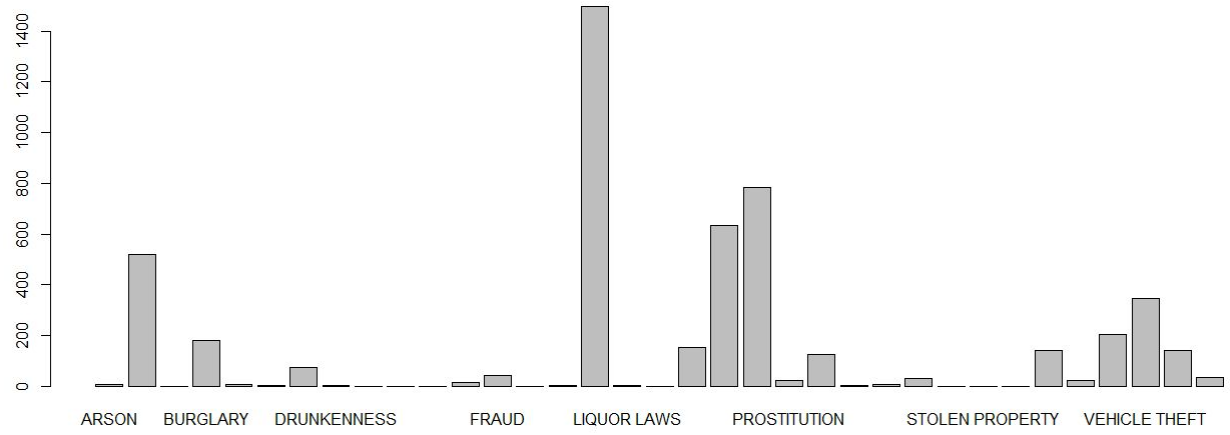
• For clustering, is there a similarity metric defined? Does it make sense for the business problem?

Clustering is not used for this model. However, we have used Association rule mining, a type of unsupervised learning for discovering association rules between different attributes. This is mainly to determine which factors (from the given data set) affect crimes the most.

Model Evaluation and Testing

This section provides details of the model in terms of its classification accuracy and other parameters, as well as visual representation of the results.

Plots of crime levels:

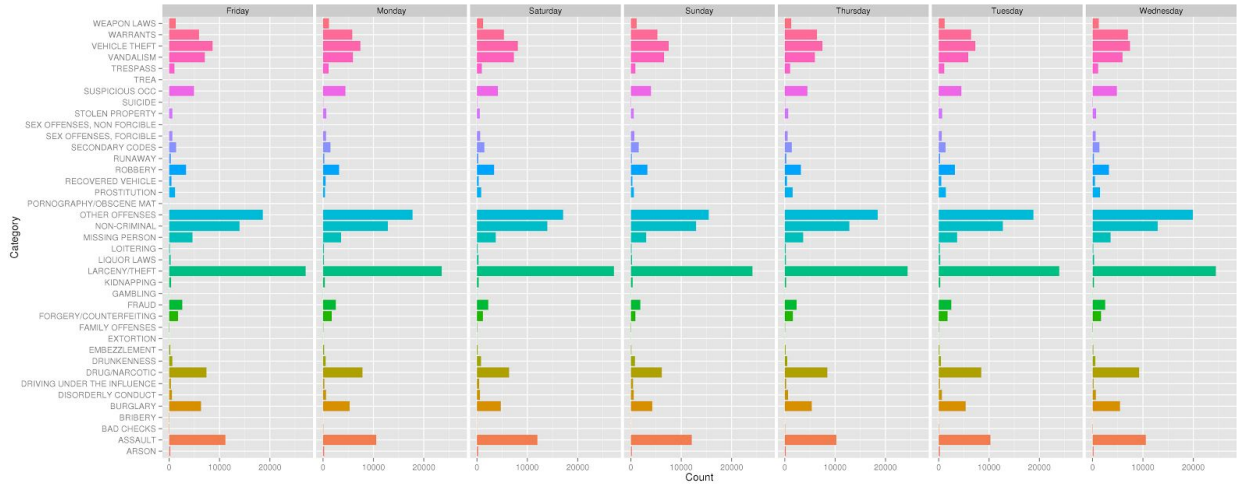


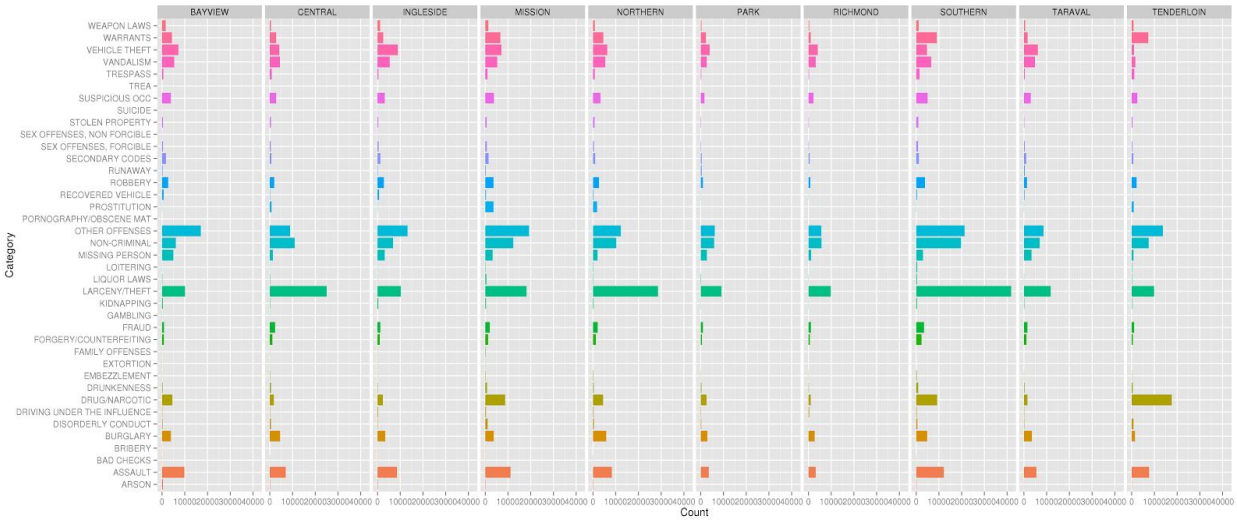
Confusion Matrix results:

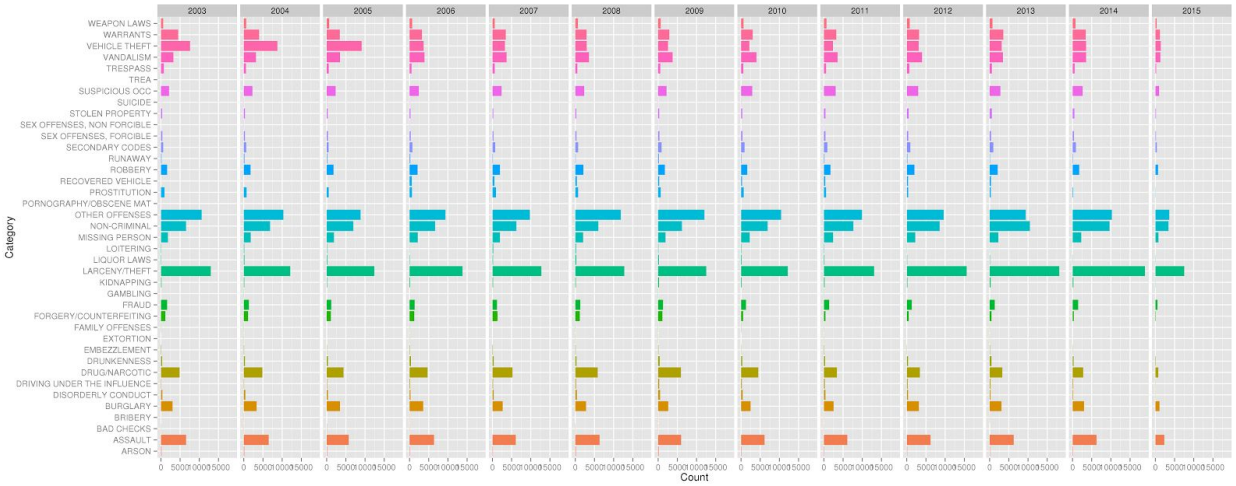
```
Overall Statistics

               Accuracy : 0.8536
                 95% CI : (0.8435, 0.8633)
    No Information Rate : 0.2758
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.8301
 Mcnemar's Test P-Value : NA
```

```
Statistics by Class:

                     Class: ARSON Class: ASSAULT Class: BRIBERY Class: BURGLARY Class: DISORDERLY CONDUCT Class: DRIVING UNDER THE INFLUENCE
Sensitivity                0.7000         0.8899         0.0000          0.8213                    0.3571                            0.3333
Specificity                1.0000         0.9713         1.0000          0.9979                    0.9998                            1.0000
Pos Pred Value             1.0000         0.7476            NaN          0.9444                    0.8333                            1.0000
Neg Pred Value             0.9994         0.9893         0.9994          0.9923                    0.9982                            0.9988
Prevalence                 0.0020         0.0872         0.0006          0.0414                    0.0028                            0.0018
Detection Rate             0.0014         0.0776         0.0000          0.0340                    0.0010                            0.0006
Detection Prevalence       0.0014         0.1038         0.0000          0.0360                    0.0012                            0.0006
Balanced Accuracy          0.8500         0.9306         0.5000          0.9096                    0.6785                            0.6667
                     Class: DRUG/NARCOTIC Class: DRUNKENNESS Class: EMBEZZLEMENT Class: EXTORTION Class: FAMILY OFFENSES
Sensitivity                        0.5094              0.1579              0.0000              NA                0.0000
Specificity                        0.9959              1.0000              1.0000               1                1.0000
Pos Pred Value                     0.7297              1.0000                 NaN              NA                   NaN
Neg Pred Value                     0.9894              0.9968              0.9998              NA                0.9998
Prevalence                         0.0212              0.0038              0.0002               0                0.0002
Detection Rate                     0.0108              0.0006              0.0000               0                0.0000
Detection Prevalence               0.0148              0.0006              0.0000               0                0.0000
Balanced Accuracy                  0.7527              0.5789              0.5000              NA                0.5000
                     Class: FORGERY/COUNTERFEITING Class: FRAUD Class: GAMBLING Class: KIDNAPPING Class: LARCENY/THEFT Class: LIQUOR LAWS
Sensitivity                             0.6190        0.4494              NA             0.1250              0.9884            0.2857
Specificity                             0.9998        0.9994               1             0.9996              0.9633            1.0000
Pos Pred Value                          0.9286        0.9302              NA             0.6000              0.9111            1.0000
Neg Pred Value                          0.9984        0.9901              NA             0.9958              0.9954            0.9990
Prevalence                              0.0042        0.0178               0             0.0048              0.2758            0.0014
Detection Rate                          0.0026        0.0080               0             0.0006              0.2726            0.0004
Detection Prevalence                    0.0028        0.0086               0             0.0010              0.2992            0.0004
Balanced Accuracy                       0.8094        0.7244              NA             0.5623              0.9758            0.6429
                     Class: LOITERING Class: MISSING PERSON Class: NON-CRIMINAL Class: OTHER OFFENSES Class: PROSTITUTION Class: ROBBERY
Sensitivity                    0.5000                0.9177              0.9485               0.9143              0.8261         0.7660
Specificity                    1.0000                0.9986              0.9859               0.9479              0.9996         0.9967
Pos Pred Value                 1.0000                0.9539              0.9021               0.7079              0.9048         0.8710
Neg Pred Value                 0.9998                0.9973              0.9929               0.9877              0.9992         0.9932
Prevalence                     0.0004                0.0316              0.1204               0.1214              0.0046         0.0282
Detection Rate                 0.0002                0.0290              0.1142               0.1110              0.0038         0.0216
Detection Prevalence           0.0002                0.0304              0.1266               0.1568              0.0042         0.0248
Balanced Accuracy              0.7500                0.9581              0.9672               0.9311              0.9128         0.8813
```

Plots of crime rates by day, district and year:







Linear regression model results:

```
Call:
lm(formula = Category ~ DayOfWeek, data = train_data)

Coefficients:
       (Intercept)     DayOfWeekMonday   DayOfWeekSaturday     DayOfWeekSunday   DayOfWeekThursday    DayOfWeekTuesday
           16.3578             1.2599             1.0218             0.5190             0.8881             0.2130
DayOfWeekWednesday
            0.7995
```

```
                                                Estimate Std. Error   t value Pr(>|t|)
(Intercept)                                    2.015e+03  4.130e-11 4.879e+13   <2e-16 ***
train_data$CategoryASSAULT                    -1.857e-22  4.177e-11 0.000e+00    1.000
train_data$CategoryBRIBERY                    -1.817e-22  8.597e-11 0.000e+00    1.000
train_data$CategoryBURGLARY                   -1.871e-22  4.228e-11 0.000e+00    1.000
train_data$CategoryDISORDERLY CONDUCT         -1.960e-22  5.407e-11 0.000e+00    1.000
train_data$CategoryDRIVING UNDER THE INFLUENCE -1.919e-22  6.000e-11 0.000e+00    1.000
train_data$CategoryDRUG/NARCOTIC              -1.874e-22  4.320e-11 0.000e+00    1.000
train_data$CategoryDRUNKENNESS                -1.908e-22  5.102e-11 0.000e+00    1.000
train_data$CategoryEMBEZZLEMENT               -2.077e-22  1.370e-10 0.000e+00    1.000
train_data$CategoryFAMILY OFFENSES            -1.712e-22  1.370e-10 0.000e+00    1.000
train_data$CategoryFORGERY/COUNTERFEITING      -1.784e-22  5.018e-11 0.000e+00    1.000
train_data$CategoryFRAUD                      -1.868e-22  4.356e-11 0.000e+00    1.000
train_data$CategoryKIDNAPPING                 -1.814e-22  4.915e-11 0.000e+00    1.000
train_data$CategoryLARCENY/THEFT              -1.859e-22  4.145e-11 0.000e+00    1.000
train_data$CategoryLIQUOR LAWS                -1.708e-22  6.436e-11 0.000e+00    1.000
train_data$CategoryLOITERING                  -1.627e-22  1.012e-10 0.000e+00    1.000
train_data$CategoryMISSING PERSON             -1.873e-22  4.258e-11 0.000e+00    1.000
train_data$CategoryNON-CRIMINAL               -1.884e-22  4.164e-11 0.000e+00    1.000
train_data$CategoryOTHER OFFENSES             -1.883e-22  4.164e-11 0.000e+00    1.000
train_data$CategoryPROSTITUTION               -1.836e-22  4.947e-11 0.000e+00    1.000
train_data$CategoryROBBERY                    -1.872e-22  4.274e-11 0.000e+00    1.000
train_data$CategoryRUNAWAY                    -1.796e-22  7.726e-11 0.000e+00    1.000
train_data$CategorySECONDARY CODES            -1.835e-22  4.450e-11 0.000e+00    1.000
train_data$CategorySEX OFFENSES FORCIBLE      -1.845e-22  4.683e-11 0.000e+00    1.000
train_data$CategorySTOLEN PROPERTY            -1.870e-22  4.915e-11 0.000e+00    1.000
train_data$CategorySUICIDE                    -1.896e-22  1.370e-10 0.000e+00    1.000
train_data$CategorySUSPICIOUS OCC             -1.870e-22  4.254e-11 0.000e+00    1.000
train_data$CategoryTRESPASS                   -1.857e-22  4.629e-11 0.000e+00    1.000
train_data$CategoryVANDALISM                  -1.867e-22  4.220e-11 0.000e+00    1.000
train_data$CategoryVEHICLE THEFT              -1.875e-22  4.191e-11 0.000e+00    1.000
train_data$CategoryWARRANTS                    4.757e-11  4.235e-11 1.123e+00    0.261
train_data$CategoryWEAPON LAWS                -1.882e-22  4.483e-11 0.000e+00    1.000
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.306e-10 on 4968 degrees of freedom
Multiple R-squared:    0.5, Adjusted R-squared:  0.4969
```

Evaluation and Deployment

After the proposed solution is modeled using different techniques articulated in the business understanding phase, it is time for data scientists and/or analysts to perform evaluations of the model before deploying it. The model is compared against several metrics depending upon the data set and requirements of the solution system.

• **Is there a plan for domain-knowledge validation?**
—Will domain experts or stakeholders want to vet the model before deployment?
If so, will the model be in a form they can understand?

The domain here is public safety, which is of utmost importance for SF city crime-fighters and the governing body. Therefore there needs to be a plan in place for validating the model results as the stakeholders/domain experts in this case would want to vet the model before deployment to ensure proper results for determining type of crime. Also, since the As the model is based on R platform, it will require a data scientist or analyst to explain the model to the concerned authorities. The data scientist should make them understand the techniques involved and algorithms employed for training their existing data and applying the model to the incoming test data. A user guide could come in handy.

Once the results are evaluated and reviewed by human experts for consistency and accuracy, then only we can say that domain-knowledge has been validated against the data science model.

• **Is the evaluation setup and metric appropriate for the business task? Recall the original formulation.**
—Are business costs and benefits taken into account?
Business costs and benefits associated to the business problem are taken into consideration in the initial phases of the project. These are under the jurisdiction of SF city governing council.

—For classification, how is a classification threshold chosen?
The default threshold of 50% or 0.5 is chosen for classification.

—Are probability estimates used directly?
We are not estimating any probabilities here; it is a pure classification implementation.

—Is ranking more appropriate (e.g., for a fixed budget)?

Ranking would not be appropriate in this case, as it is not practical or wise to rank different categories of crime. They can occur at any place at any time, and could be of any nature. Agreed, in a given area based on the day of week we can rank crimes according to their occurrence frequency.

—For regression, how will you evaluate the quality of numeric predictions? Why is this the right way in the context of the problem?

In this problem, regression is used mainly to get an estimate of number of crime incidents occurring by each day of the week. it is plotted to provide a visual representation that is more clear. In the context of the problem, this is correct as it gives substantial information to the city police officials to act and strategize upon.

• **Does the evaluation use holdout data?**
—Cross-validation is one technique.

The evaluation uses a 10-fold cross-validation technique for SVM classifier.

• **Against what baselines will the results be compared?**
—Why do these make sense in the context of the actual problem to be solved?

The model is first trained on the training set provided, that is the historical data of crime records supplied by the SFPD. The model is tested on incoming test data with an accuracy of 85%, the highest achievable value in this case. The model is tried out again on the training set to identify the error rate and matching the results with the actual crime categories. This will give the officials an idea of specificity of the model and they would know when to take a leap of faith in the model. It makes sense in the current context to try out the model back on the training set itself to expect the possibility of crimes in an area.

—Is there a plan to evaluate the baseline methods objectively as well?

The baseline methods will be evaluated by data scientists and analysts by plotting the model predictions and analyzing precision, recall, F-measure, error rate, specificity and accuracy obtained from confusion matrices of the classifier. After extensive study of these factors as well as manual observations and comparisons of classified values, these methods can be evaluated objectively by the data folks and not the SF Police whose knowledge and experience in crime related matters might affect the judgement.

• For clustering, how will the clustering be understood?

Clustering was not used for this business problem.

• Will deployment as planned actually (best) address the stated business problem?

Yes. The planned deployment after rigorous testing and evaluation rounds will best address the business problem if the crime rate starts dipping after using the model.

• If the project expense has to be justified to stakeholders, what is the plan to measure the final (deployed) business impact?

The final business impact could be well measured by the observed level of crime occurrences after going live with the model. If the crime categories are classified as accurately as possible, and the police officials are achieving success in curbing them, then the project has a positive impact on the community and business problem in general, and the related expense can be justified to stakeholders without any qualms.

References

1. San Francisco Crime Classification [Competition Data]. Retrieved from https://www.kaggle.com/c/sf-crime
2. SF Open Data [Data Set]. Retrieved from https://data.sfgov.org/
3. Provost, F & Fawcett, T. *Data Science for Business.* Sebastopol, CA: O'Reilly
4. SFPD information, http://www.sanfranciscopolice.org/
5. Teetor, P. *R Cookbook*: O'Reilly
6. Google.com for all R- related questions