

**W. P. CAREY SCHOOL OF BUSINESS HONOR CODE MASTER’S PROGRAMS**

We, the students of the W.P. Carey Master’s programs, are committed to maintaining the highest levels of honesty, integrity, and professional courtesy for ourselves and amongst our classmates. We believe that ethical behavior is not only the cornerstone of a good academic program, but of good business as well. As such, we are steadfast in our commitment to the following principles:

* Academic Integrity – We view the W.P. Carey Academic Integrity Policy as a living, breathing document, not just concepts on a page. As such, we are resolute in ensuring that we abide by the principles contained within, both in letter and in spirit. Furthermore, we will hold our classmates to the same high standards to which we hold ourselves.

* Consideration of Others – As W. P. Carey students, we believe that we should treat others as we wish to be treated. To this end, we will be considerate of classmates, professors, and recruiters in all aspects. Furthermore, we will constructively add to the collaborative environment by helping and encouraging others.
* Professionalism – As W. P. Carey students, we believe that the behaviors we practice today will become habit in the future. To that end, we are committed to act in a professional manner in all school settings, including classes, networking events, and interviews. We will be cognizant of the fact that we are representing the program and our classmates at all times.

Full Name (printed): \_ Ashfaqul Haq \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Signature\_\_\_\_ Ashfaqul Haq \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ Date \_\_\_\_\_04/23/2016\_\_\_\_\_\_\_\_

Full Name (printed): \_\_\_\_ David Dorfman \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Signature\_\_\_\_\_\_\_ David Dorfman \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ Date \_\_\_\_\_04/23/2016\_\_\_\_\_\_\_\_

Full Name (printed): \_\_\_\_ Divya Saini \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Signature\_\_\_\_\_\_ Divya Saini \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ Date \_\_\_\_04/23/2016\_\_\_\_\_\_\_\_

Full Name (printed): \_\_\_\_\_ Shaikh Shiban Qureshi \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Signature\_\_\_\_\_\_\_ Shaikh Shiban Qureshi \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ Date \_\_\_\_\_04/23/2016\_\_\_\_\_\_\_\_

Onsite MSBA Applied Project Report

Spring 2016

W.P. Carey, ASU

|  |  |
| --- | --- |
| Topic | SanFrancisco Crime Classification |
| Team | C-01 |
| Team Members | Shaikh Shiban Qureshi, Divya Saini, David Dorfman, Ashfaqul Haq |
| Client information (if applicable) | NA |

# **Executive Summary**

Kaggle has provided about 12 years of crime reports from all over the San Francisco city. The objective of this project is prediction of crime category, given the date, time, location coordinates, approx. coordinates, name of the Police Department District and details on how the crime was resolved. The training set and test set rotate every week, meaning week 1,3,5,7... belong to test set, week 2,4,6,8 belong to training set with date ranging from 1/1/2003 to 5/13/2015.

The primary beneficiary of this project will be San Francisco Police Department. This project will help them to manage personnel deployment and patrolling operations more effectively. Additionally, the local community will be a great benefactor from this project. Another benefactor will be real estate developers who can use this knowledge to establish more secured housing schemes in crime-prone areas like theft, burglary, etc.

We used two-class and multi-class classification algorithms for this problem using Azure ML and R. We received better results with multi-class classification approach using Azure ML and feature engineering. We observed that since a single address can constitute of multiple (X,Y) coordinates, the Address variable was not of much help in crime category predictions. Address can may be used for positioning of Police force later. We also observed that time information of past crimes is very helpful in prediction of future crimes. We could observe trend in crime occurrence during a day and also over the week and month. The obvious limitation of the approach is that stereotyping the category of crime based on past occurrence of crime is not a practical approach. Crime location may be approximately same more or less but crime category might vary erratically. And there is no obvious trend in crime category observed, which makes the prediction even more difficult. The fundamental assumption that we are making in this problem is that category of the crime happening in future would remain same given the location and time of the crime.

# **Background**

The old methods of blanket crime prevention across the entire city is not working in today’s age of budget cuts and decreased police officers. Therefore, using the analytics to pinpoint exactly what part of the city and what crime segments to focus on enabled the police force to be more effective and prevent more crime with less work and funds than was needed in the past.

The initiative taken in this direction is **Hotspot Policing**. It was introduced through the smart policing initiative which is being introduced in cities by collaboration between the Federal Bureau of Justice Assistance and non-porfit CAN(Center for Naval Analyses) corporation that provides funding, training and technical assistance for data center crime prevention programs across the country.

Due to the success of Smart Police Initiative program at Philadelphia, the department is expanding the number of analysts on the force. Due to these increased efforts with analytics, places that have historically being resistant to efforts in crime prevention have now shown significant reductions in crime with one neighborhood in Philadelphia showing a 20% drop in crime in just three months. Hence, the police department has done recent adoption of more rigorous analytical models such as GIS mapping and predictive analysis which means that police departments can find subtler trends in crime and draw scientific conclusions about their causes.

A specific example occurred in the Germantown in Chestnut Hills neighborhood of Philadelphia. An increase in burglaries in the district were found to be correspondent to truancy rates in the neighborhood schools. This knowledge allowed the police officers to inform the neighbors about the increase in crime and allow them to be vigilant in protecting their homes. There are two things that make this program of Philadelphia different from other police initiatives, one being a partnership with the local university and emphasis on crime science not just analysis.

Why this is important? Some regions in US have shown a steady decline in crime in the past few decades. San Francisco overall crime rate is showing an uneven trend in crime preventions gains. In 2013, Oakland was ranked 3rd highest in violent crimes, however Oakland had lower crime rates in every other category. With such uneven gains, it is obvious that analytics is needed to pinpoint the neighborhoods and the type of crimes that should be focused on the most.

**Figure 1: Statistics on Police Staffing and Crime Rates**

# **Problem Statement**

As in all cities, crime is a reality San Francisco: Everyone who lives in San Francisco seems to know someone whose car window has been smashed in, or whose bicycle was stolen within the past year or two. The city has been battling crime for almost a century now. As the city booms having the tech advantage, there is also an increase in crime ranging from property crimes to thefts and burglaries. This Kaggle problem involves attempting to guess the class of a crime committed within the city, given the time and location it took place. Such studies are representative of efforts by many police forces today: Using machine learning approaches, one can get an improved understanding of which crimes occur where and when in a city — this then allows for better, dynamic allocation of police resources. To aid in the SF challenge, Kaggle has provided about 12 years of crime reports from all over the city — a data set that is pretty interesting to comb through.

The objectives of this project include:

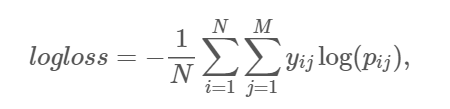
1. Prediction of crime category, given the time and location of crime occurred.
2. Creating rich and comprehensive visualizations to map patterns and trends in the crime data.

The primary beneficiary of this project will be San Francisco Police Department. This project will help them to manage personnel deployment and patrolling operations more effectively. Additionally, the local community will be a great benefactor from this project. Another benefactor will be real estate developers who can use this knowledge to establish more secured housing schemes in crime-prone areas like theft, burglary, etc. The fundamental assumption that we are making in this problem is that category of the crime happening in future would remain same given the location and time of the crime.

Data details:

* Train Data Size -
* Test Data Size – 884262 records
* Number of Crime Categories - 39

For evaluating the submissions, Kaggle is using Multi-Class Logarithmic approach. This is the multi-class version of the Logarithmic Loss metric. Each observation is in one class and for each observation, you submit a predicted probability for each class. The metric is negative the log likelihood of the model that says each test observation is chosen independently from a distribution that places the submitted probability mass on the corresponding class, for each observation.



*where*

* + *N is the number of cases in the test set,*
  + *M is the number of class labels,*
  + *\\(log\\) is the natural logarithm,*
  + *\\(y\_{ij}\\) is 1 if observation \\(i\\) is in class \\(j\\) and 0 otherwise, and*
  + *\\(p\_{ij}\\) is the predicted probability that observation \\(i\\) belongs to class \\(j\\).*

# **Methods**

**Approach till Interim Project Presentation:**

**

We first understood the data by running several visualizations in Tableau and we found out that our data is skewed as most of the crime type belonged to Larceny/Theft as we can see in the word cloud on the left.

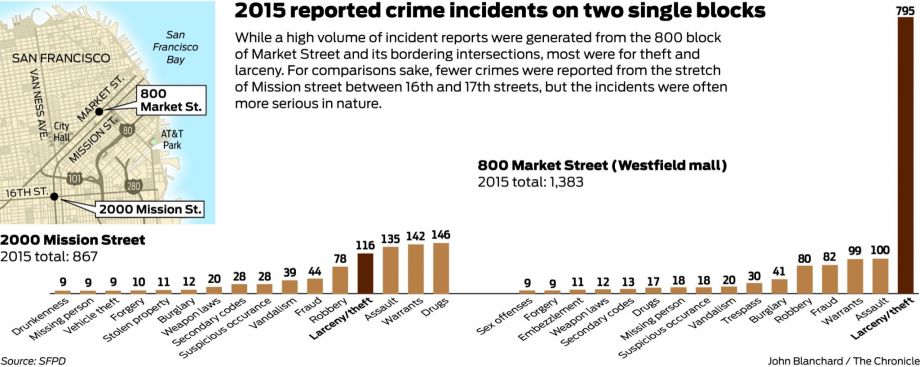
**Figure 2: Crime Category Frequency Word Cloud**

*We first started with a binary class classification approach to solve this problem by converting the training data set in the submission format with column for each crime category and applied two-class classification algorithms like Decision Forest and Support Vector Machine. The reason behind using this approach was that some of the classification algorithms like Support Vector Machine do not support multi-class classification. Also the lower complexity level of two-class classification algorithms gives better performance.*

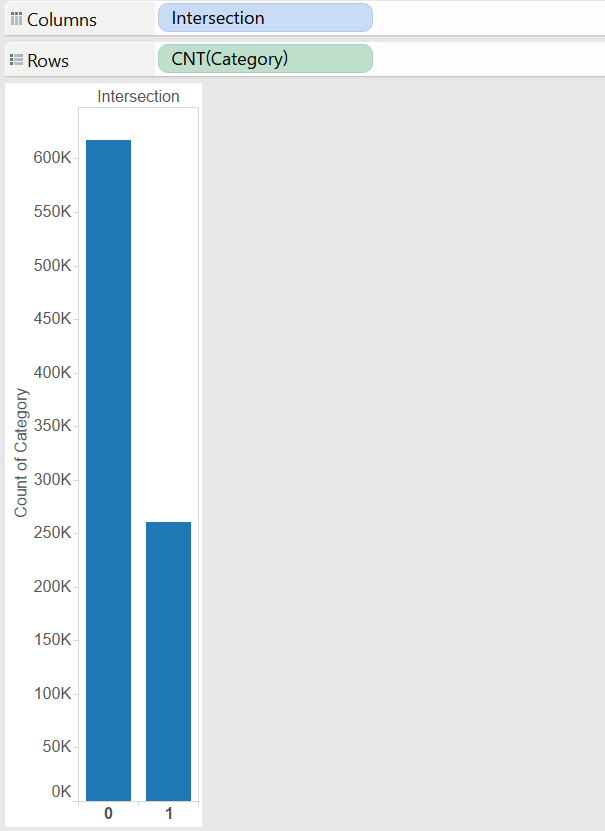
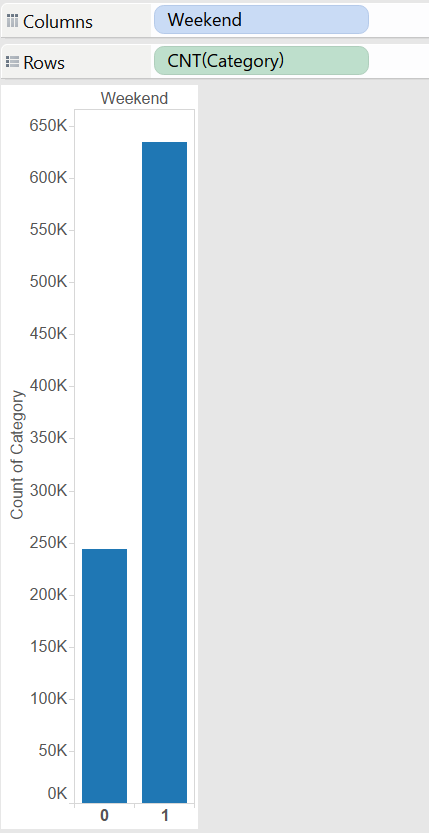
We had used only R to make predictions before interim project presentation. After facing issues with glm(), we implemented glmnet() algorithms using R code to get successful results. Glmnet is for elastic net regression. This penalises the size of estimated coefficients (via a mix of two ways). It tries to explain as much variance in the data through the model as possible while keeping the model coefficients small. Glm doesn't use a penalty term. The effect, as we understand it, that with elastic net you may be accepting some bias in return for a reduction in the variance of the estimator.

We also derived following variables to be used in the model:

* + *Split ‘Dates’ – Day, Month, Year, Hour, Minutes, Seconds – to extract hour of crime time.*
  + *Intersection – 1 at addresses with intersections – some of the addresses have intersection as part of the address, variable that was 1 if the crime was committed at an intersection and 0 if it was not. This is because as per our findings online, we found that crime occurrences are more at intersections.*
  + *Night – 1 between 10 pm to 6 am – this is because we expected a possible rise in crime in the late night hours as that’s the time city is less crowded.*
  + *Week – 1 on weekdays else 0 – this is because we assumed that there is a high possibility of crime on weekend than weekdays as on weekends people are mostly out partying and stay out till very late in the night.*

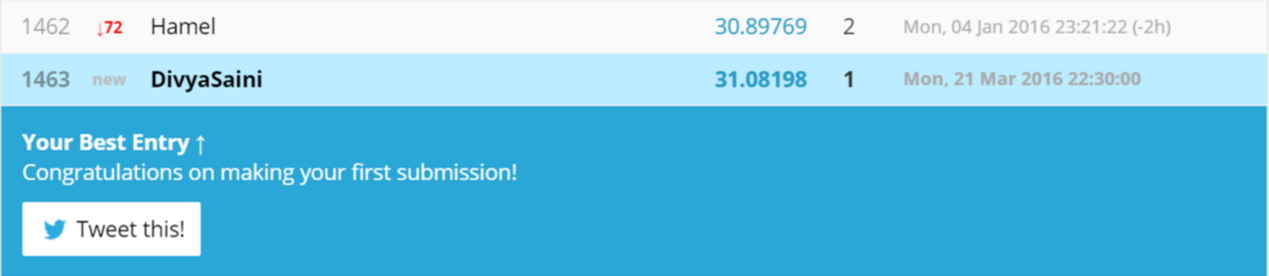


**Figure 3: Crime Pattern at Intersections in SFO**

**Figure 4: Distribution of Crime Over Addresses with Intersection Figure 5: Weekend vs Weekdays Crime Pattern**

*With this approach we received a score of 2.5 on test set but 31.08 on train set suggesting that we might be overfitting our model.*



**Figure 6: First Kaggle Submission**

**Approach after Interim Project Presentation:**

After the feedback we received in interim project presentation and the poor results we received on Kaggle Leaderboard we made significant changes in our model.

We changed our approach from two-class classification approach to multi-class classification. We used Multi-Class Decision Forest, Jungle and Logistic Regression (See Figure 7). We could not successfully run Neural Network as it was taking infinite time to get finished. Out of these, we received best results from Decision Forest, surprisingly, as Decision Jungle are the advanced version of Decision Forest algorithms.

Logistic regression is a well-known method in statistics that is used to predict the probability of an outcome, and is particularly popular for classification tasks. We will briefly discuss Decision Forest and Decision Jungle below.

**Decision Forest:**

The decision forest algorithm is an ensemble learning method for classification. The algorithm works by building multiple decision trees and then voting on the most popular output class. Voting is a form of aggregation, in which each tree in a classification decision forest outputs a non-normalized frequency histogram of labels. The aggregation process sums these histograms and normalizes the result to get the “probabilities” for each label. The trees that have high prediction confidence will have a greater weight in the final decision of the ensemble.

Decision trees in general are non-parametric models, meaning they support data with varied distributions. In each tree, a sequence of simple tests is run for each class, increasing the levels of a tree structure until a leaf node (decision) is reached.

Decision trees have many advantages:

* They can represent non-linear decision boundaries.
* They are efficient in computation and memory usage during training and prediction.
* They perform integrated feature selection and classification.
* They are resilient in the presence of noisy features.

**Decision Jungles:**

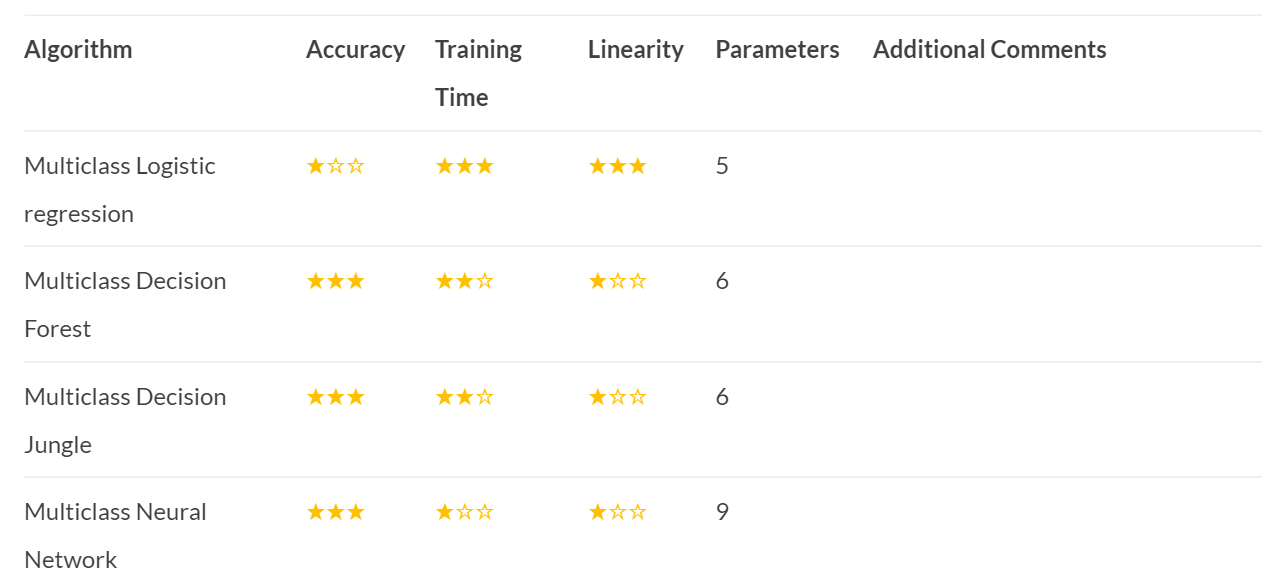
Decision jungles are a recent extension to decision forests. A decision jungle consists of an ensemble of decision directed acyclic graphs (DAGs).

Decision jungles have the following advantages:

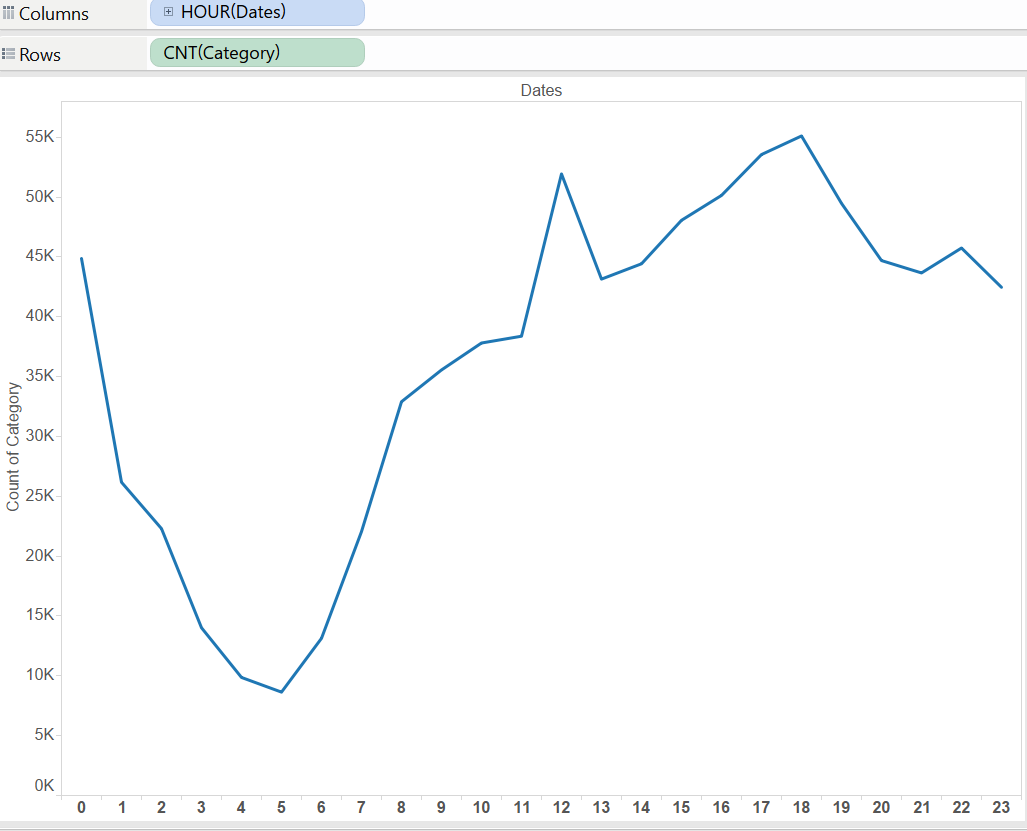
* By allowing tree branches to merge, a decision DAG typically has a lower memory footprint and a better generalization performance than a decision tree, albeit at the cost of a somewhat higher training time.
* Decision jungles are non-parametric models, which can represent non-linear decision boundaries.
* They perform integrated feature selection and classification and are resilient in the presence of noisy features.

We also used a more data driven approach this time for feature engineering. For example, we plotted the crime frequency against day hours for all the years and after observing the pattern, we came up with the following bins (See Figure 8):

* Steep Peak – 5 to 8: Early Morning
* Slow Peak – 8 to 12: Morning
* Short Dip – 12 to 13: Noon
* Slow Peak – 13 to 18: Early Evening
* Slow Dip – 18 to 22: Late Evening
* Sharp Dip – 22 to 5: Night

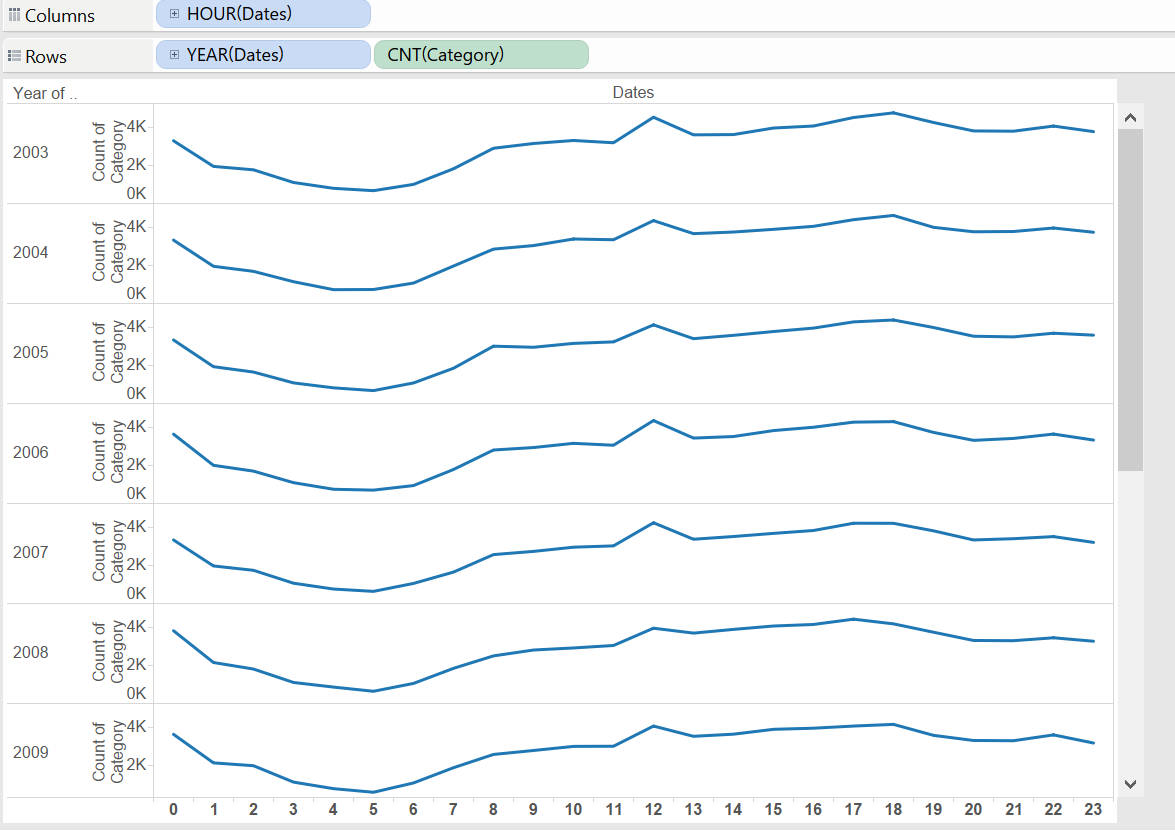


**Figure 7: Comparison of Multi-Class Classification Algorithms**



**Figure 8: Distribution of Crime Frequency Over a Day for all the 12 years**

We also confirmed this pattern by plotting visualization for individual years and we still observed the same pattern (See Figure 9).

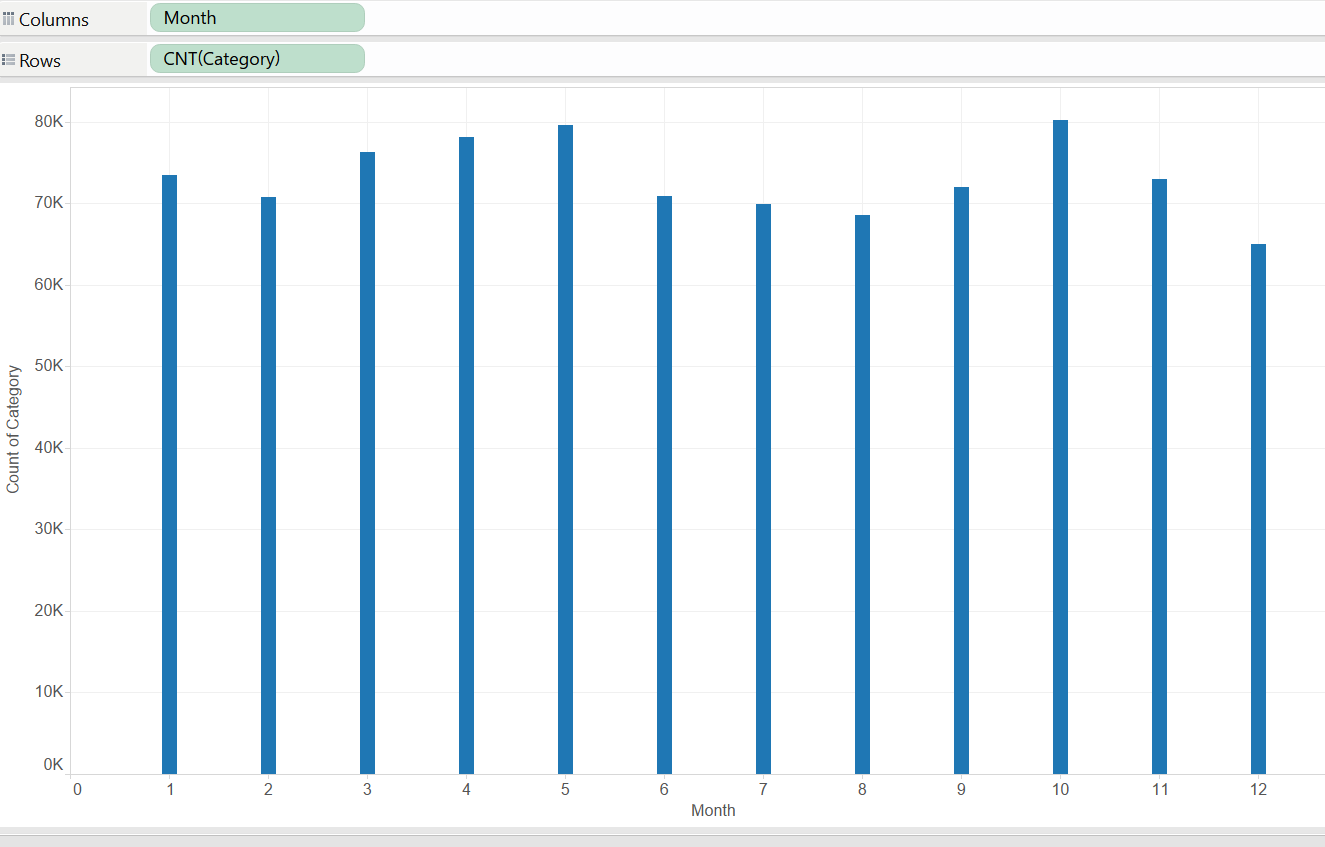


**Figure 9: Distribution of Crime Frequency Over a Day for Individual Years**

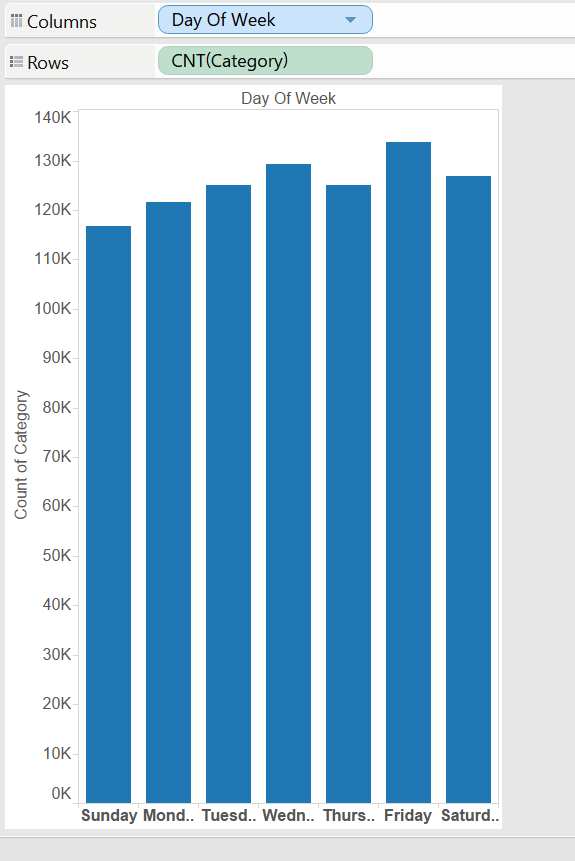
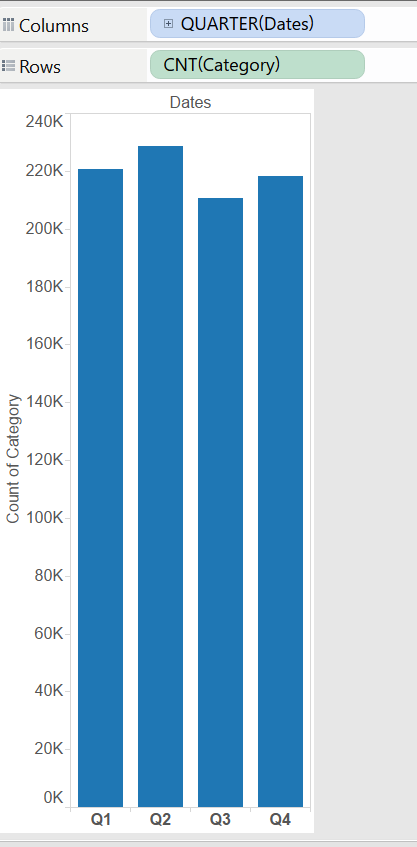
We further observed seasonality in crime frequency over a year by plotting the frequency against months and used month as another feature (See Figure 10).

We also decided to use weekdays as a feature for our model rather than the weekend feature earlier as we observed correlation between crime frequency and weekdays (See Figure 11).

However, we did not observe any significant seasonality over different quarters (See Figure 12).

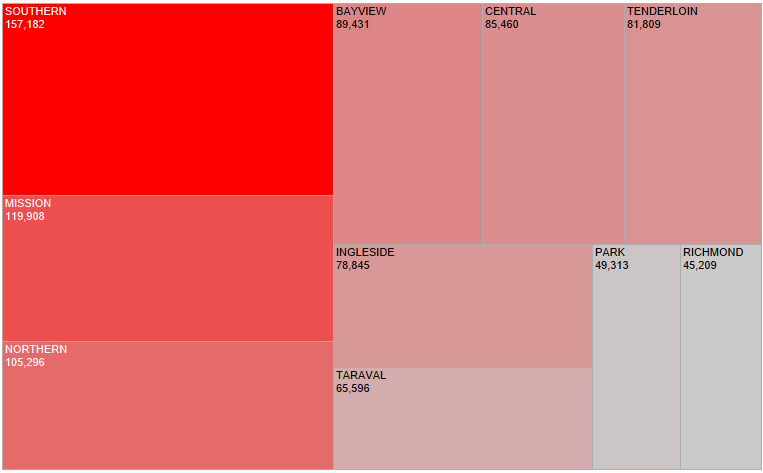


**Figure 10: Distribution of Crime over Months**

**Figure 11: Distribution of Crime over Weekdays Figure 12: Distribution of Crime over the four Quarters**

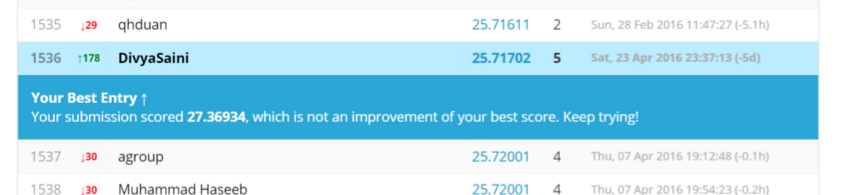
We could not draw any significant pattern from the Police District given, also Police districts were very few and would have reduced the granularity of information given our large data size.



**Figure 13: Heat Map of Police Districts and Crime Frequency**

For the same reason, we dropped the Address variable and used (X,Y) coordinates only for location as (X,Y) coordinates are more specific to the crime location.

After making all these changes we received an improved score 25.7.

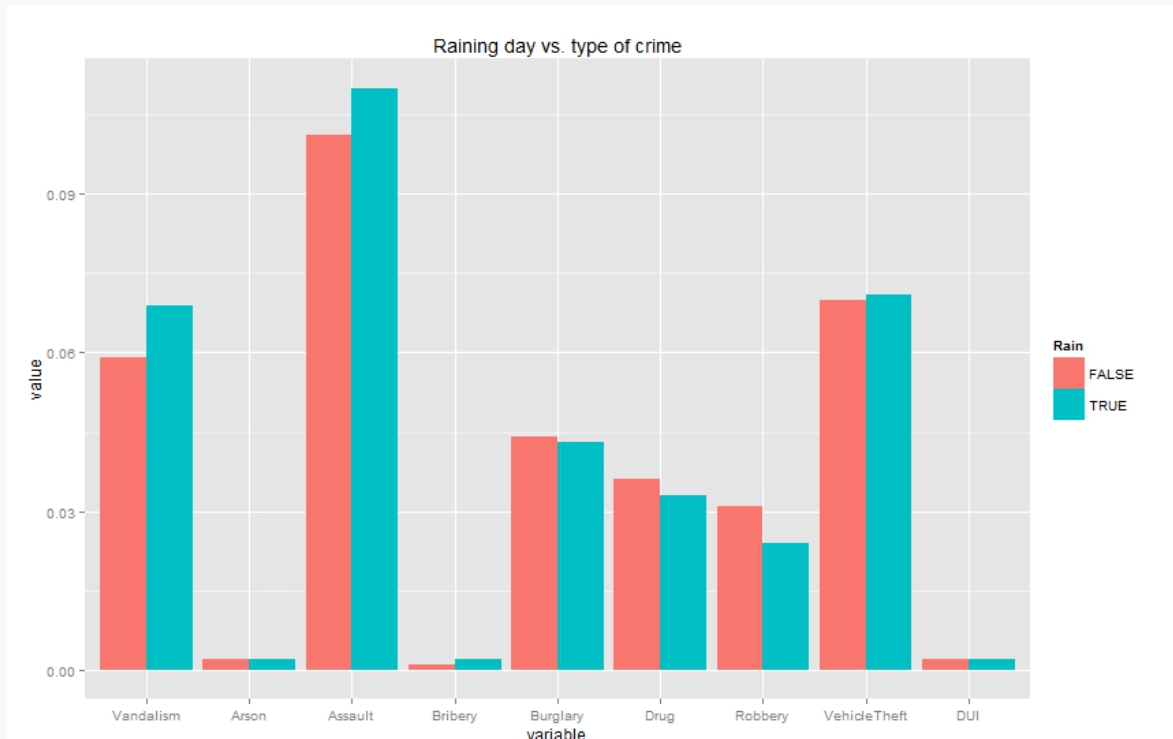


**Figure 14: Last Kaggle Submission**

# **Results and Conclusions**

The problem of predicting crime appears like a simple application of classiﬁcation algorithms, however the exploration of the data unearthed some interesting trends such as the relationship between crime and the hour of the day. While the dataset gives us the police districts, we have attempted to improve on this by using the X and Y coordinates as a grid to create ﬁner location features. Our ﬁnal model represents our efforts to create a model which makes use of both the time and location features present in the original dataset, as well as using the Multi-Class Decision Forest which is easy to understand and yet “complex” enough to make use of the larger feature set in our model. This model also runs relatively quickly on the large dataset. Our group has enjoyed working on this dataset and will be exploring some of the ideas discussed for our future Kaggle submissions. It will be interesting to see how the Kaggle prize-winning models will inﬂuence future police work, and whether this will have the effect of helping to lower the crime rate in San Francisco. However, the significant assumption as we make in any other prediction problem is correlation between past occurrence and future expected occurrences. In case of crime prediction also, there can be a lot of noise factors depending on social situations and weather conditions that may lead to deviation from the predcitions.

Further, since crimes are social event, we also came across interesting correlation between crime type and weather of the day (rainy/sunny), we believe if explored further, this could be an interesting correlation for accurate crime predictions.



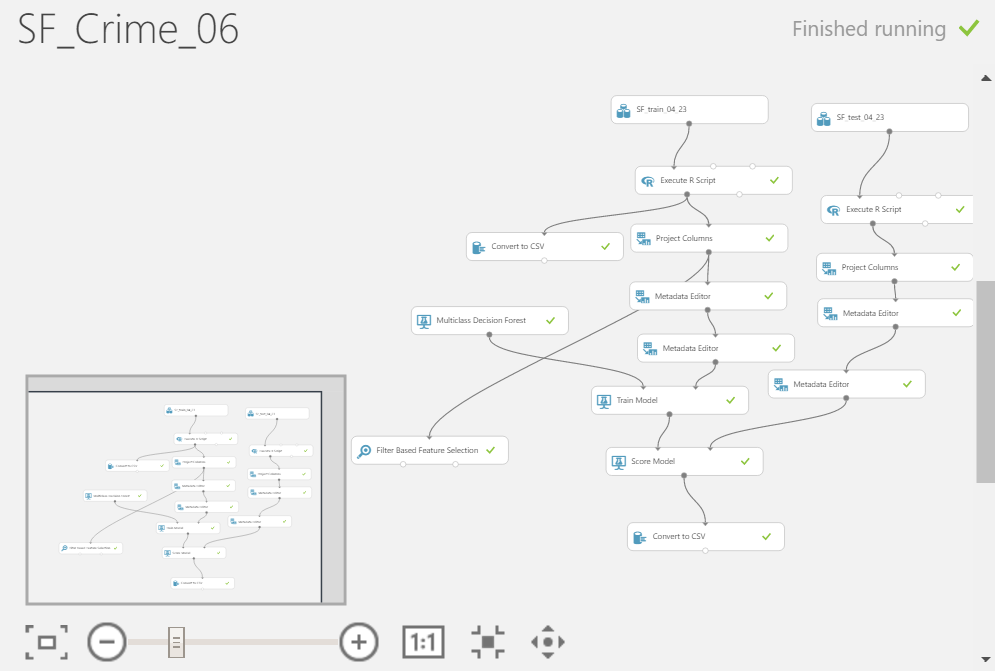
**Figure 15: Relation between Rain Occurrence and Crime Category Type**

# **References**

1. <https://shapeofdata.wordpress.com/2013/06/04/multi-class-classification/>
2. [*http://technical.ly/philly/2014/02/18/philadelphia-police-smart-policing-crime-scientists/*](http://technical.ly/philly/2014/02/18/philadelphia-police-smart-policing-crime-scientists/)
3. [*http://www.forbes.com/sites/emc/2014/06/03/data-analysis-helps-police-departments-fight-crime/#244bd59043e8*](http://www.forbes.com/sites/emc/2014/06/03/data-analysis-helps-police-departments-fight-crime/#244bd59043e8)
4. [*https://nycdatascience.com/correlation-between-weather-condition-and-the-type-of-crime/*](https://nycdatascience.com/correlation-between-weather-condition-and-the-type-of-crime/)
5. [*http://www.sfchronicle.com/crime/article/Dubious-distinction-S-F-s-most-crime-ridden-6767550.php*](http://www.sfchronicle.com/crime/article/Dubious-distinction-S-F-s-most-crime-ridden-6767550.php)
6. [*http://www.sfchronicle.com/crime/article/U-S-crime-drops-again-but-gains-uneven-in-Bay-6536129.php*](http://www.sfchronicle.com/crime/article/U-S-crime-drops-again-but-gains-uneven-in-Bay-6536129.php)
7. [*https://msdn.microsoft.com/en-us/library/azure/dn905853.aspx*](https://msdn.microsoft.com/en-us/library/azure/dn905853.aspx)
8. [*http://scottge.net/2015/07/04/ml101-how-to-choose-a-machine-learning-algorithm-for-multi-class-classification-problems/*](http://scottge.net/2015/07/04/ml101-how-to-choose-a-machine-learning-algorithm-for-multi-class-classification-problems/)

# **Appendix – Reproduction of Results**

* The training and testing data set can be downloaded from - <https://www.kaggle.com/c/sf-crime/data>. It also has a sample submission file which shows the format of the submission file to be submitted on Kaggle.
* Tools/Platform used – Azure ML, R and Tableau
* We are not able to provide the link for published experiment from Azure ML here as the dataset size crosses 100MB. Hence we are providing a screenshot of the experiment here.



**Figure 16: Experiment Snapshot from Azure ML**

* After running the Azure ML experiment, download the csv with scored labels and copy pasted scored label column in the SamplesSubmission.csv file, at the end of all the columns. Use the R script below for this updated SamplesSubmission.csv file. And after running the script below, delete the first column from the output file created and change the headers same as the original submission file. Now, “Submission05.csv” file is ready for submission on Kaggle leaderboard.

data05 = read.csv("Submission05Old.csv")

str(data05)

table(data05$Scored.Labels)

data05$ARSON=ifelse(data05$Scored.Labels=="ARSON",1,0)

data05$ASSAULT=ifelse(data05$Scored.Labels=="ASSAULT",1,0)

data05$BAD.CHECKS=ifelse(data05$Scored.Labels=="BAD CHECKS",1,0)

data05$BRIBERY=ifelse(data05$Scored.Labels=="BRIBERY",1,0)

data05$BURGLARY=ifelse(data05$Scored.Labels=="BURGLARY",1,0)

data05$DISORDERLY.CONDUCT=ifelse(data05$Scored.Labels=="DISORDERLY CONDUCT",1,0)

data05$DRIVING.UNDER.THE.INFLUENCE=ifelse(data05$Scored.Labels=="DRIVING UNDER THE INFLUENCE",1,0)

data05$DRUG.NARCOTIC=ifelse(data05$Scored.Labels=="DRUG/NARCOTIC",1,0)

data05$DRUNKENNESS=ifelse(data05$Scored.Labels=="DRUNKENNESS",1,0)

data05$EMBEZZLEMENT=ifelse(data05$Scored.Labels=="EMBEZZLEMENT",1,0)

data05$EXTORTION=ifelse(data05$Scored.Labels=="EXTORTION",1,0)

data05$FAMILY.OFFENSES=ifelse(data05$Scored.Labels=="FAMILY OFFENSES",1,0)

data05$FORGERY.COUNTERFEITING=ifelse(data05$Scored.Labels=="FORGERY/COUNTERFEITING",1,0)

data05$FRAUD=ifelse(data05$Scored.Labels=="FRAUD",1,0)

data05$GAMBLING=ifelse(data05$Scored.Labels=="GAMBLING",1,0)

data05$KIDNAPPING=ifelse(data05$Scored.Labels=="KIDNAPPING",1,0)

data05$LARCENY.THEFT=ifelse(data05$Scored.Labels=="LARCENY/THEFT",1,0)

data05$LIQUOR.LAWS=ifelse(data05$Scored.Labels=="LIQUOR LAWS",1,0)

data05$LOITERING=ifelse(data05$Scored.Labels=="LOITERING",1,0)

data05$MISSING.PERSON=ifelse(data05$Scored.Labels=="MISSING PERSON",1,0)

data05$NON.CRIMINAL=ifelse(data05$Scored.Labels=="NON-CRIMINAL",1,0)

data05$OTHER.OFFENSES=ifelse(data05$Scored.Labels=="OTHER OFFENSES",1,0)

data05$PORNOGRAPHY.OBSCENE.MAT=ifelse(data05$Scored.Labels=="PORNOGRAPHY/OBSCENE MAT",1,0)

data05$PROSTITUTION=ifelse(data05$Scored.Labels=="PROSTITUTION",1,0)

data05$RECOVERED.VEHICLE=ifelse(data05$Scored.Labels=="RECOVERED VEHICLE",1,0)

data05$ROBBERY=ifelse(data05$Scored.Labels=="ROBBERY",1,0)

data05$RUNAWAY=ifelse(data05$Scored.Labels=="RUNAWAY",1,0)

data05$SECONDARY.CODES=ifelse(data05$Scored.Labels=="SECONDARY CODES",1,0)

data05$SEX.OFFENSES.FORCIBLE=ifelse(data05$Scored.Labels=="SEX OFFENSES FORCIBLE",1,0)

data05$SEX.OFFENSES.NON.FORCIBLE=ifelse(data05$Scored.Labels=="SEX OFFENSES NON FORCIBLE",1,0)

data05$STOLEN.PROPERTY=ifelse(data05$Scored.Labels=="STOLEN PROPERTY",1,0)

data05$SUICIDE=ifelse(data05$Scored.Labels=="SUICIDE",1,0)

data05$SUSPICIOUS.OCC=ifelse(data05$Scored.Labels=="SUSPICIOUS OCC",1,0)

data05$TREA=ifelse(data05$Scored.Labels=="TREA",1,0)

data05$TRESPASS=ifelse(data05$Scored.Labels=="TRESPASS",1,0)

data05$VANDALISM=ifelse(data05$Scored.Labels=="VANDALISM",1,0)

data05$VEHICLE.THEFT=ifelse(data05$Scored.Labels=="VEHICLE THEFT",1,0)

data05$WARRANTS=ifelse(data05$Scored.Labels=="WARRANTS",1,0)

data05$WEAPON.LAWS=ifelse(data05$Scored.Labels=="WEAPON LAWS",1,0)

write.csv(data05, "Submission05.csv")