DSAML Project (Individual)

**SG United Programme (SGUS): Upskill in Data Science**

**Specialist Diploma in Data Science for Business**

**AY21/22, SGUS Oct Intake 3**

**DECLARATION**

I declare that I am the originator of this work and that all other original sources used in this work have been appropriately acknowledged.

I understand that plagiarism is the act of taking and using the whole or any part of another person’s work and presenting it as my own without proper acknowledgement.

I also understand that plagiarism is an academic offence and that disciplinary action will be taken for plagiarism.

|  |  |
| --- | --- |
|  | I Agree (Please Tick ✓) |

|  |  |
| --- | --- |
| **My Information** | |
| Name (as in matriculation card) | Wong Poh Yeng |
| Admin Number | 2073306I |
| Group (1 or 2) | 1 |

**For Tutor’s Use**

|  |  |
| --- | --- |
| Overall Grade: |  |
| Feedback |  |

Project Background & Aims

BACKGROUND

Crime is one of the biggest and dominating problem in our society. It is an unavoidable part of our lives in this world. Every day we hear about them and some of us even involve in at least one of them throughout our life. As a result, preventing it is a critical task.

Identifying crime patterns and community areas with high rates of violent crime, along with crime event prediction, is critical to societal crime prevention. It can assist law enforcement agencies in developing optimal patrol strategies and concentrating their efforts and resources. As a result, societal security measures will be improved, and crime incidents will be reduced, which will benefit society in a variety of ways.

With today's technology, AI and machine learning can assist in predicting crime patterns and predictive policing in general. We can predict when a violent crime will occur based on the beat, location, and type of crime, and then assign officers to patrol specific areas in the hopes of preventing crimes and making neighborhoods safer.

When an individual is charged with a crime, there are two types: violent crimes and non-violent crimes. The distinction between the two are the severity and sentences.

placing

right resource at right time and right place is the key factor to

overcome this problem

placing

right resource at right time and right place is the key factor to

overcome this problem

placing

right resource at right time and right place is the key factor to

overcome this problem

A crime is considered violent if a deadly weapon was involved, it can be with or without bodily harm. This type of crime includes arson, assault, battery, sexual assault, homicide, intimidation, kidnapping, robbery, sex offense.

Non-violent crimes are usually not severe and do not endanger the lives of others. Common examples are burglary, law violation, damage, trespass, deceptive practice, gambling, human trafficking, interference with public officer, narcotics, non-criminal, obscenity, stalking.

In this project, I will be using the Chicago Crime 2015-2016 dataset. The types of crime are divided into two group – violent and non-violent crime. This dataset will be split into two parts:80% and 20%. The 80% is for training and validation purposes whereas the 20% for test purpose.

OBJECTIVE:

|  |
| --- |
| The goal of this project is to train and predict where the violent crimes are more likely to occur. |

# Description of Dataset

This dataset is obtained from Kaggle. The dataset is sourced from the Chicago Police Department’s CLEAR (Citizen Law Enforcement Analysis and Reporting) system.

It includes reported incidents of crime in the City of Chicago in year 2015-2016.

Columns 21 | Row:234,782

|  |  |
| --- | --- |
| **Data** | **Description** |
| **ID** | Unique identifier for the record. |
| **Case Number** | unique to the incident |
| **Date** | Date when the incident occurred |
| **Block** | address where the incident occurred |
| **IUCR** | The Illinois Unifrom Crime Reporting code |
| **Primary Type** | The primary description of the IUCR code |
| **Description** | The secondary description of the IUCR code |
| **Location Description** | Description of the location where the incident occurred |
| **Arrest** | Indicates whether an arrest was made. |
| **Domestic** | Indicates whether the incident was domestic-related as defined by the Illinois Domestic Violence Act. |
| **Beat** | Indicates the beat where the incident occurred. |
| **District** | Indicates the police district where the incident occurred |
| **Ward** | The ward (City Council district) where the incident occurred |
| **Community Area** | Indicates the community area where the incident occurred |
| **FBI Code** | Indicates the crime classification as outlined in the FBI's National Incident. |
| **X Coordinate** | The x coordinate of the location where the incident occurred |
| **Y Coordinate** | The y coordinate of the location where the incident occurred |
| **Year** | Year the incident occurred. |
| **Latitude** | The latitude of the location where the incident occurred. |
| **Longitude** | The longitude of the location where the incident occurred |
| **Location** | Actual location that allows for creation of maps |

**Observation**

Dates: Date and time is combined into a single column and it is It is in the following format: Y-m-d H:i:s. E.g.: 2015-05-13 23:53:00

Primary Type: 33 unique type of Crime

Location Description: 146 unique count (1,658 missing values)

Beat: Range from 111 to 2535

District: Range from 1 – 31 (1 missing value and some outliers)

Ward: Range 1-50 (14 missing value)

Community Area: Range from 0-77 (44 missing value)

# Data Exploration & Preparation

There are 33 types of crime in Primary Type column. Using Excel's vlookup function, I divided the types of crime into two groups: violent and nonviolent crime.

A new column, Voilent, was created with Booleen True and False.

After that, I used Tibco Spotfire to clean and transform the data.

Transformation

|  |  |
| --- | --- |
| Columns | Transfromation Details |
| Date | Split the Date and Time into two columns |
| Location Description | Group locations that are similar. From 146 unique locations, I manage to reduce to 100 unique locations.   |  |  | | --- | --- | |  | Example:  All under Airport group | |
| Beat | In Tibco, group the beat into 10 bins. |
| Community Area | In Tibco, group the beat into 7 bins. |

Cleaning

|  |  |
| --- | --- |
| Columns | Work done |
| **ID, Case Number,IUCR, FBI Code,**  **X Coordinate, Y Coordinate , Location, Ward** | Delete all these unwanted columns |
| Location Description | Replace the missing values with “Others” |
| District, Ward, Community Area | Delete all of the missing values since it only a small percentage of total population |
| District | Remove the outliers since it only a small percentage of total population |
| Community Area | Remove Community Area “0” since it only a small percentage of total population. |

The Final Dataset was exported as [Crime\_Full.csv]

Columns 13 | Row:233,183

|  |  |
| --- | --- |
| **Data** | **Description** |
| Time | Time of the incident |
| Block | address where the incident occurred |
| **Primary Type** | The primary description of the IUCR code |
| **Description** | The secondary description of the IUCR code |
| **Location Description** | Description of the location where the incident occurred |
| **Arrest** | Indicates whether an arrest was made. |
| **Domestic** | Indicates whether the incident was domestic-related as defined by the Illinois Domestic Violence Act. |
| **Violent** | Booleen True and False |
| **District** | Police district of the incident |
| **Community Areas** | **Community Areas of the incident** |
| **Year** | Year of the incident |
| **Date** | Date of the incident |
| **Beat** | Beat of the incident |

Using RapidMinar, the dataset was split into training and test sets with an 80:20 ratio.

Summary of files after data preparation:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| No. | Filename | Purpose | Columns | Rows |
| 1 | Crime\_Training.csv | Training | 13 | 186546 |
| 2 | Crime\_Test\_nolabel.csv | Testing | 12\* | 46637 |
| 3 | Crime\_Test\_label.csv | Predicting | 13 | 46637 |

Note\*: Testing data has 1 less column (Target Column)

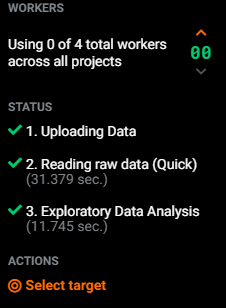
EXPLORATION

1. Import from local file.

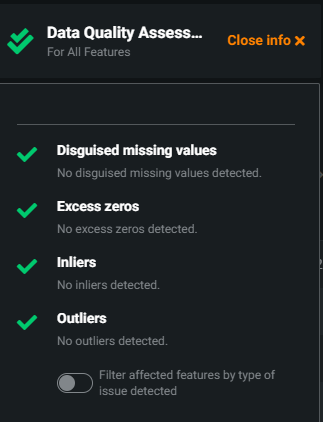
|  |
| --- |
|  |

2. Open and upload the training data file (Crime\_Training.csv).

3. Datarobot automatically performs EDA.



4.Check **the data quality** assessment. Make sure no missing values, excess zeros, inliers and outliers.



Features Engineering

Before running the model, I need to select the target and relevant features.

1. In Spotfire, I have ensured that the data types are correctly classified. The large portion of data transformation and feature engineering are done in Spotfire. There are not many changes that need to be made in Datarobot.
2. Next, I will set Violent the target label. Since I set Violent as Target label, the primary type is an obvious Target Leakage which should not be included in the modelling.

|  |
| --- |
|  |

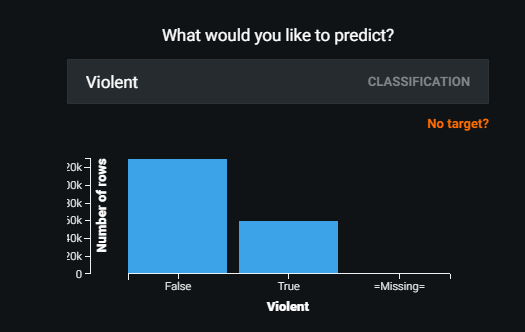
1. The final 6 features are selected as follows. After that, I will create the feature list by clicking +Create feature list.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  |  |  |  |  |

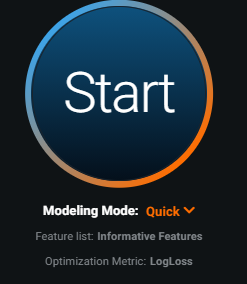
1. Named the new list “myfeaturelist”

# Modelling

1. DataRobot automatically identifies this as a classification problem.



1. Logloss is suggested as the Optimization Metric.



1. Next, I will check “Show Advanced Options – Partitioning”

|  |  |
| --- | --- |
|  |  |

Under partitioning, Stratified is chosen as it ensures similar target distribution across each partition.

DataRobot will splits your data into a 20% holdout (test) partition and the remaining 80% over five-fold cross-validation (training and validation) partitions. You can change these values from the **Advanced Options** > **Partitioning**

I will run models using Cross-Validation with 5-folds and 20% holdout, as the training data is not very large with only 186,546 rows.

1. Under Show Advanced Options, there are downsampling and feature constraints.

The default is set, but you can change it if you want.

|  |  |  |  |
| --- | --- | --- | --- |
| Downsampling (Default) | Feature Constraints (Defaullt) | | |
|  |  | | |
|  | |  |
|  | |  |

1. Show Advanced Options – Additional (Default)

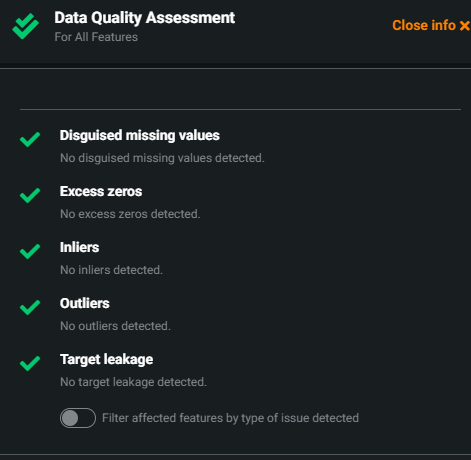
|  |  |
| --- | --- |
|  |  |

Datarobot will recommend and auto select metrics for you. If you want to change selection, you can go to Advanced option> Additional tab. For classification problem, I will use the recommended logloss metric.

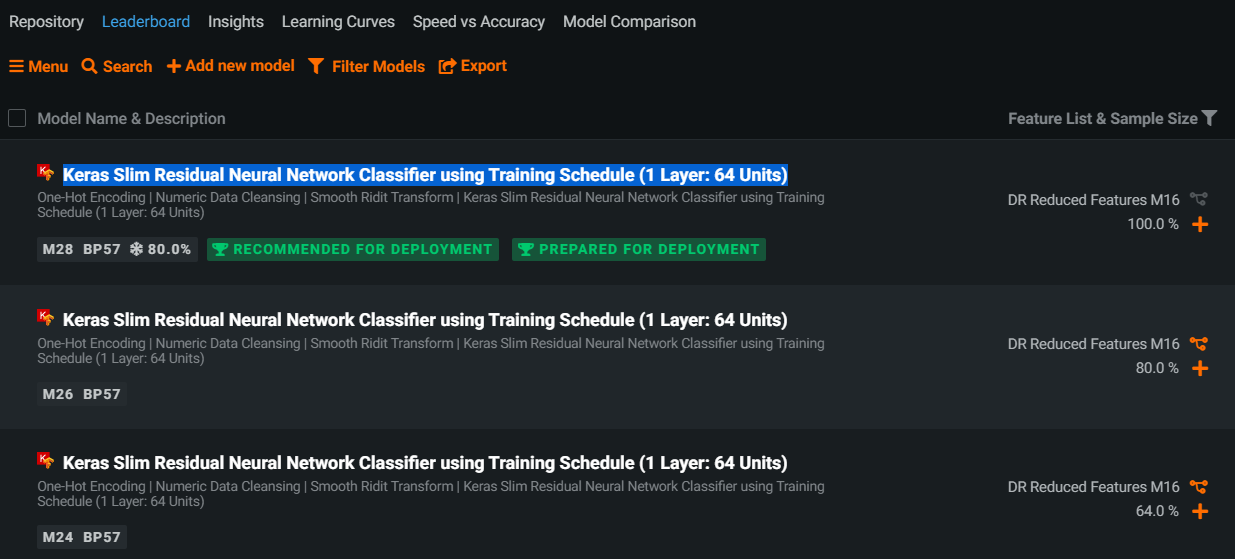
1. Now click on start to run the modelling.

|  |  |
| --- | --- |
|  |  |

1. While DataRobot is generating the models, ensure there is no Target Leakage by looking into Data Quality Assessments.



1. After the models have finished running, I can see 19 models generated. **Keras Slim Residual Neural Network Classifier using Training Schedule (1 layer:64 units)** is recommended for deployment.



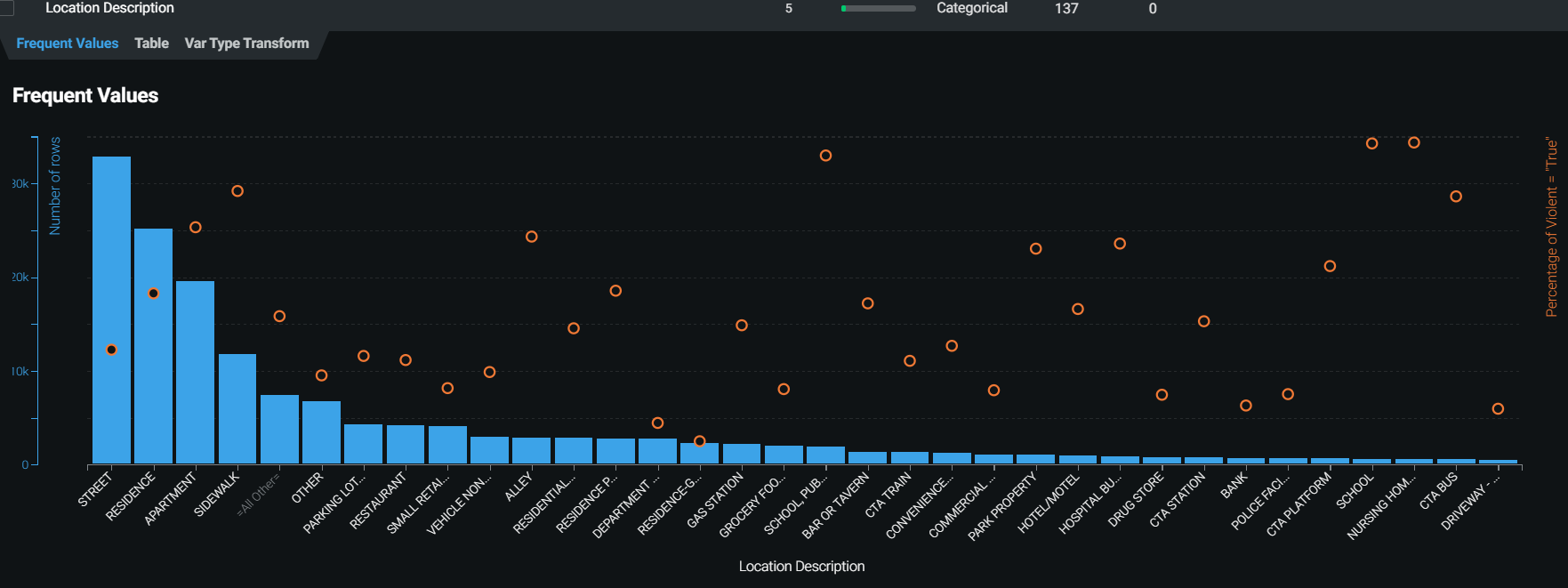
# Interpretation of Results

**Insights discovered in Data Section**

1. After modelling is completed, there is a new feature list called DR Reduced Features M16 was created by DataRobot. It listed out the 6 of the most important features.

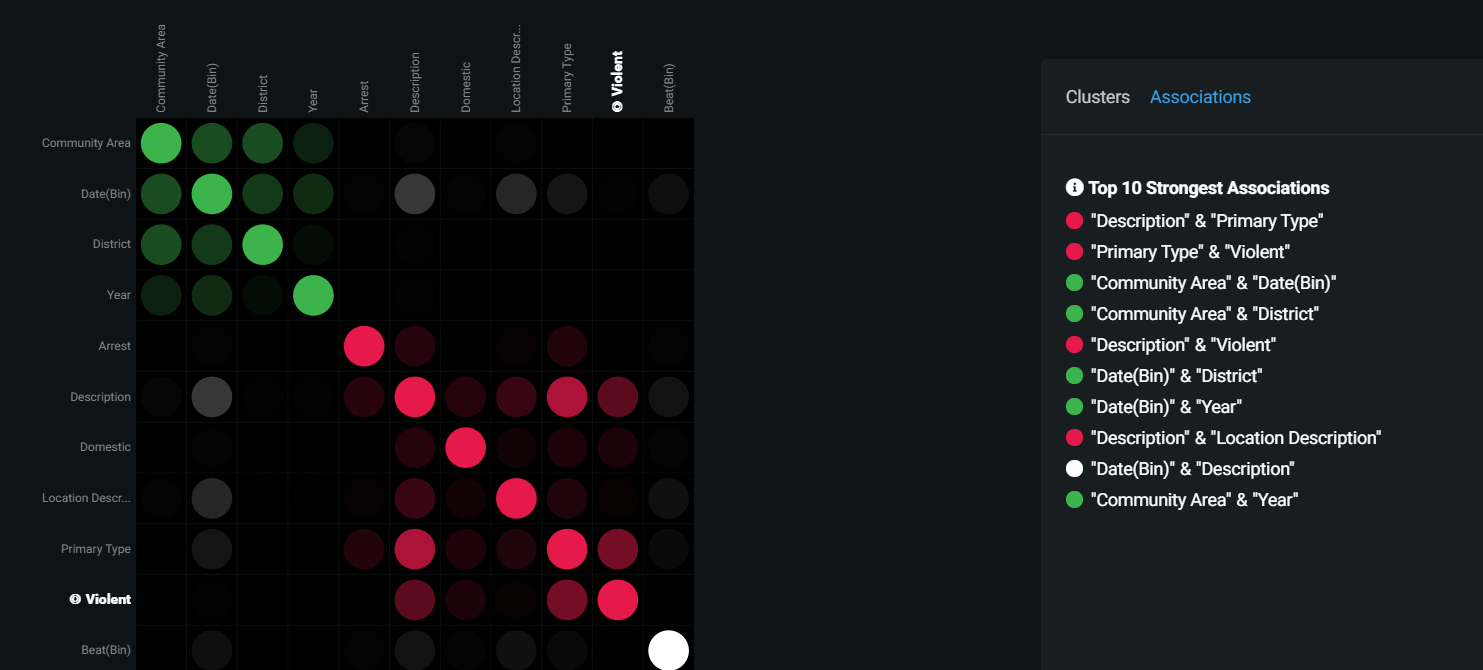
|  |  |
| --- | --- |
|  |  |

1. **Frequency values**

****

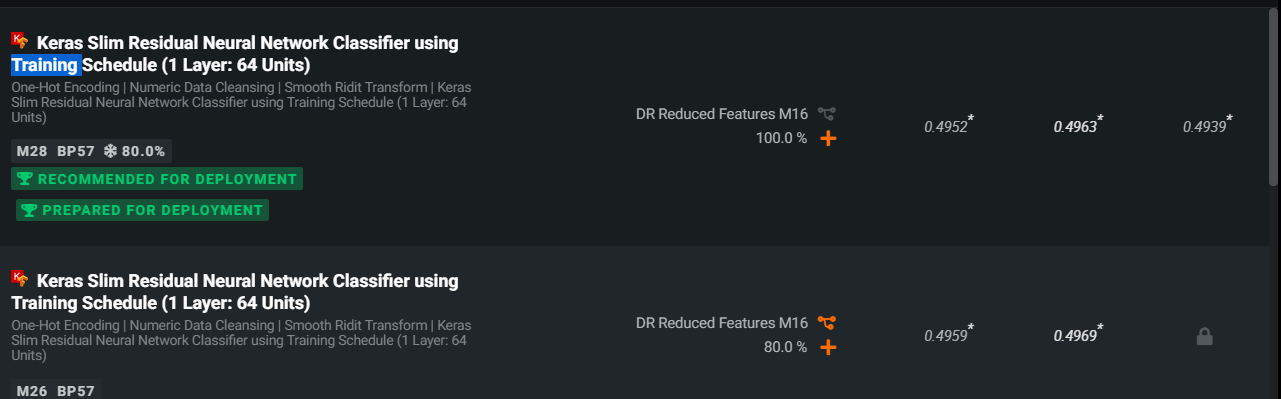
The frequency values in location description column provides me with useful information about where the violent cases are most likely to occur. In this case, it is more common on the street, residence, apartments and sidewalk.

1. **Feature Association**

****

From this Feature Association, I discovered that “Primary Type & Voilent”, “Community Area & District” & “Date & Year” – they are correlated. These features (Primary Type, District & Year) will not be selected when I re-test or run the model.

Interpretation on Model

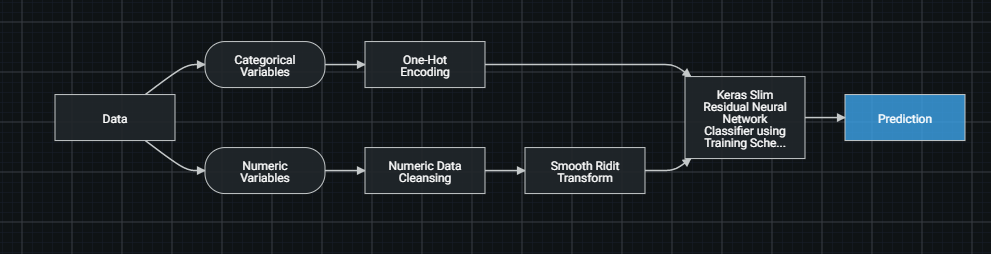


Keras Slim Residual Neural Network Classifier is recommended model for deployment. Let dig deeper into the model.

1. **Logloss**

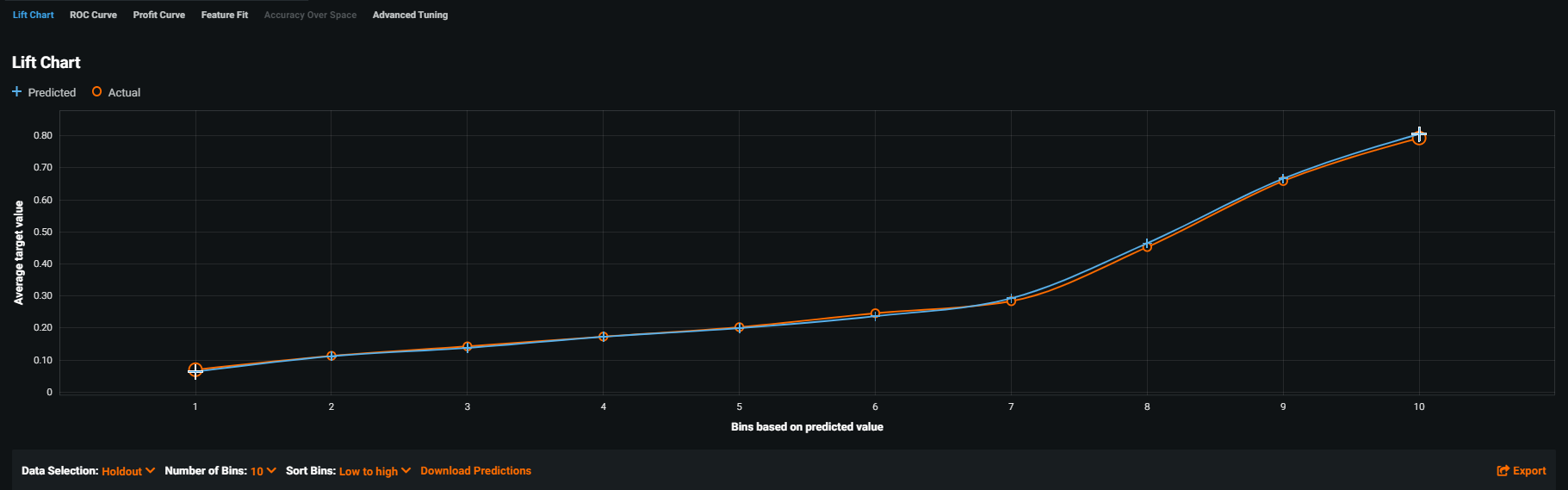
The logloss results in Validation, Cross Validation and Holdout sets show below 0.5 (perfect classifier logloss=0). This rate is still considered acceptable. Even though the model is not fully trained, it is still better than random guess.

1. **Describe – Blueprint**



As per the blueprint, some feature engineering was performed prior to generating the final prediction output from the model.

1. **Lift Chart**



The orange and blue lines in the lift chart cross each other. There are almost no gaps between the prediction and the actual. This indicates that the model is fairly accurate and it is not consistently overestimating or underestimating.

1. **ROC Curve**

|  |
| --- |
|  |

**ROC Curve**

It has curve well above random line. It shows high optimum point at 0.6 with low Fallout rate at 0.18.

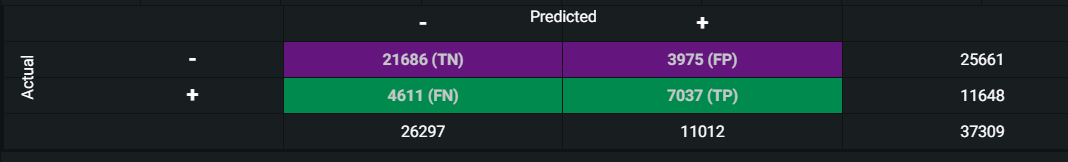
AUC curve is used to measure the predictive accuracy of a model. If the model with AUC of 0.5 or less is worthless. The graph above shows AUC of 0.7811 which indicated that it is a highly acceptable model.

**Prediction Distribution**

This graph shows the distribution of actual values in relation to the threshold. Purple color is classified as a “false” – [FP,TN]. Whereas Green color to the right is classified as a “true”- [TP,FN]. A good Prediction distribution graph will show no overlapping.

Purple and green overlap in my model when the threshold is set to 0.5. This is because there are a lot of false cases in the dataset.

1. **Confusion Matrix**

****

When “Voilent Crime” is set as target. The Confusion Matrix is interpreted as follow:

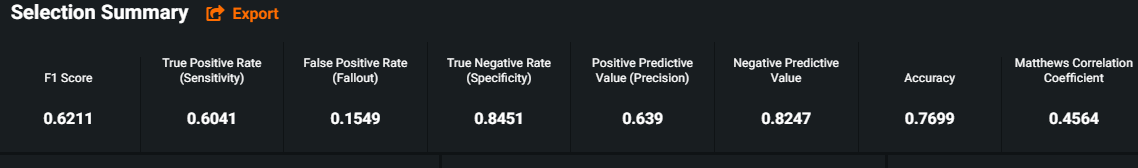
|  |  |  |  |
| --- | --- | --- | --- |
| **Actual** |  | **Predicted** | |
| **Non-voilent Crime** | **Voilent Crime** |
| **Non Voilent** | TN =True Negative  **Non-Voilent Crime** the model predicts as **Non-Voilent Crime**. | FP = False Positive (False alarm)  **Non-Voilent Crime** the model predicts as **Voilent Crime**. |
| **Voilent** | FN = False Negative(Miss Out)  **Voilent Crime** the model predicts as **Non-Voilent Crime.** | TP = True Positive  **Voilent Crime** the model predicts as **Violent Crime**. |

**True Negative(TN**)=21,686 has a decent high value and **True Positive(TP)** shows 7,037, given the fact that it is not very high, it is still higher than False Positive (FP) and False Negative (FN). It means that the model can predict correctly whether it is a violent or non-violent crime.

**False Positive(FP**) - It is a false alarm, but not as serious as False negative (FN). Only additional manpower or cost required. The model shows 3,975 cases which are still considered acceptable.

**False Negative(FN)** can cost lives when violent cases being missed out. The model shows 4,611 (12.35%) cases which is a point of concern.

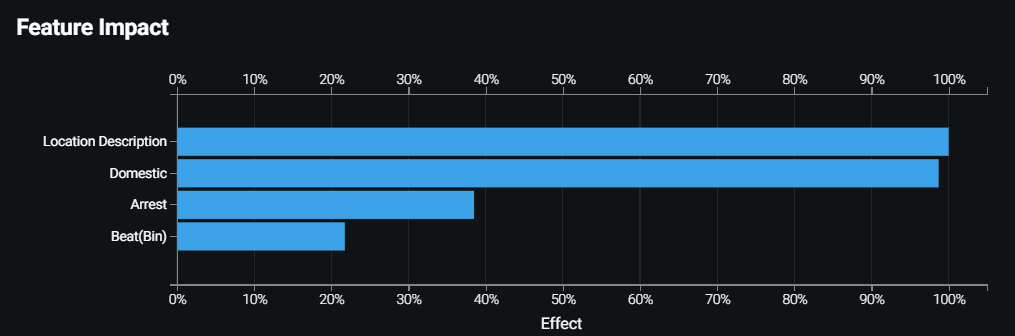
1. **Sensitivity(Recall) & Accuracy**

****

The reasons for using Sensitivity and Accuracy as performance metrics for this model is as follow:

1. **Sensitivity/Recall** – used when you do not want to miss any violent crime. Therefore, False Negative can be as low as possible. We can compromise low precision, but high recall. In this model, the Recall rate is 0.6041 which is considered acceptable as it is above 0.5.
2. **Accuracy** can also be used when FN & FP counts are close. With the accuracy score of 0.7699, it means that the model is being trained correctly.
3. **F1 Score** isoverall measure the accuracy of the model. It is used when we have an uneven class distribution. Given that the F1 score of 0.6211 which is considered as fair (above 0.5). It means that the false positive and false negative are not that high.

**Feature Impact**



Location Description, Domestic and Arrest are the 3 features that have the most impact on the “target”. It appears that there is a connection between the location of violent crime and the likelihood of domestic violence.

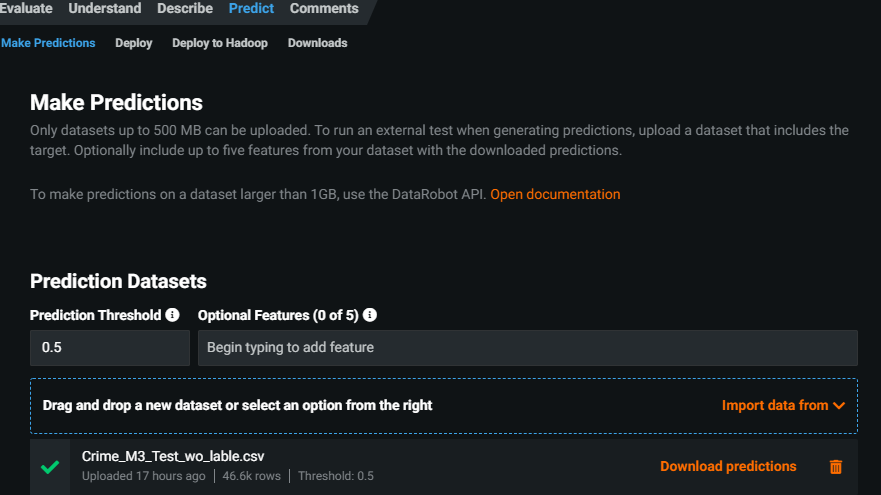
1. Prediction Explanation



# The prediction explanation emphasizes the significance of the three factors that will influence the prediction's outcome.

Prediction

Finally, I will use the training model to predict using my test dataset.

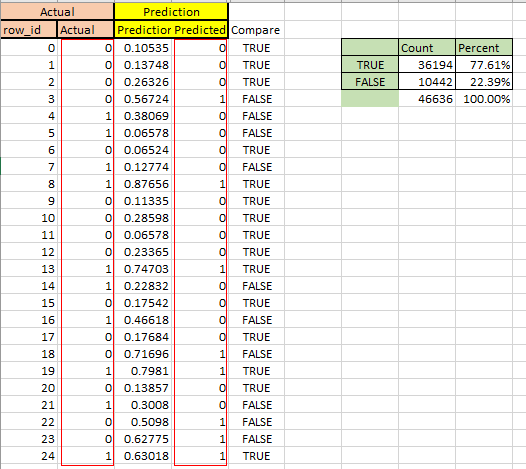


1.Under predict, upload the “Crime\_Test\_wo\_lable.csv” -> click compute.

2.Once complete, download the prediction.csv file.

3.Open the prediction.csv using Excel. Copy all the 3 columns in the csv.file.

4.Paste the 4 columns into the table of “Crime\_Test\_label.csv” file.



5. A new column [Compare] was add to compare Actual and Prediction Column.

6. The formula is set in Excel, and it shows 36,194 ‘True’ cases for a 77.61% accuracy in predicting violent case.

# Findings

1. **Reduce Features**

The selection of features is important, and reducing irrelevant features will increase model performance and avoid overfitting.

My original dataset has 13 features in this project. By reducing the numbers of features to 6, the Logloss decreased from 0.502 to 0.4952 , the F1 score increased from 0.6162 to 0.6211 and Sensitivity increased from 0.5957 to 0.6041.

1. **Reduce Classes**

Reduce the numbers of classes on the Beat, time (e.g., 1-3am, 8-10pm), and Community Area can improve performance and accuracy - Logloss has dropped from 0.4952 to 0.3694 and time spent running the model is reduced.

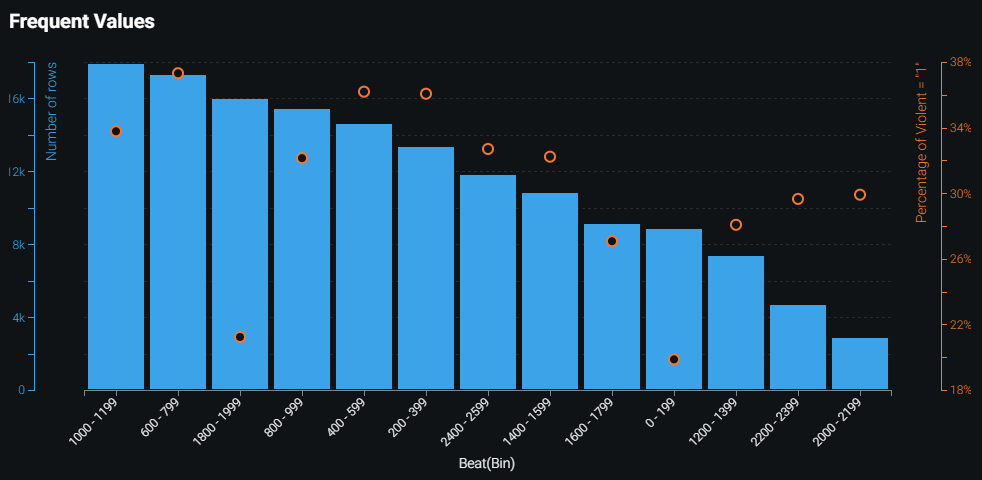
# Recommendations

It is inspiring to see Datarobot predict with 77.61% accuracy. Given more time, I believe this accuracy could be improved further by obtaining relevant dynamic features. Following by re-train the model using newer, more informative data.

One of the areas I would like to dig deeper into when I have more time is grouping the date and predict whether or not any violent crime will occur based on day (eg mon or Sunday). Based on my curiosity, I believe it plays an important part in crime prediction. I am unable to apply in this project due to time constraints and the dirty dataset provided.

**Analyses and Recommendations**

1. As you can see, using Datarobot to explore the Chicago Crime Dataset can help in the identification of crime patterns/types and locations.



A bar chart can be used to predict which beat area and location, such as an apartment or street, the brutal crime is most likely to occur. This crime classification is helpful. It can forecast which areas will be crime hotspots. Police could be warned in this way, and they could, for example, increase surveillance and patrolling in this area.

1. **Analyses & Recommendation**

As previously stated in the Confusion Matrix, I am more concerned about the False negative because it has the potential to be costly to individuals and the community.

A false negative indicates that there is a violent crime but the model predicted as non-voilent crime. It could have happened for a variety of reasons, including human error in incorrectly classifying the type of crime or bias imposed on policymakers.

In this project, both FN and FP have the same equal weight. There is a trade-off between the two. When one goes up, the other goes down.

These two types of errors are very difficult to implement in the justice system. Determining the risk of violence in each situation can be challenging if law enforcement officers do not understand the models.

To produce accurate results, law enforcement and transparency of predict policing models are required.

\*\*\*\*\* END OF REPORT \*\*\*\*\*