|  |
| --- |
| **OTAGO POLYTECHNIC AUCKLAND INTERNATIONAL CAMPUS** |
| Progress & Prototyping Report |
| Crime Prediction Using Predictive Analytic |
|  |
|  |
| **Vimita Vaidya (1000063758)** |

**6/28/2020**

|  |
| --- |
|  |

**Graduate Diploma & Bachelor in Information Technology**

# Executive Summary

This report mainly focus on prototype of predictive policing. At the beginning of the document highlight the predictive policing workspace and project structure. Section 4 to 9 in this document is depending on the life cycle step of data science, where the process continues until we get desire score from our model. The main part of this document is modelling section after each new dataset and features score has been obtained and comparing this score with our baseline score.

In the last section appendices, which includes the graphs, pandas-profile report, feature selection score, and research document which I used as a reference for important information and statistic of New Zealand dataset.

We are working on improving the outcome of the experiment. To this end, we apply different machine learning algorithms, cross-validation, parameter tuning, and we would like to concentrate on the TensorFlow and PyTorch deep learning algorithms to tarin our predictive model.

# Table of Contents

[Executive Summary 2](#_Toc44275524)

[Table of Contents 3](#_Toc44275525)

[1 Workspace 5](#_Toc44275526)

[1.1 High-Level Project Structure 5](#_Toc44275527)

[2 Crime Dataset 7](#_Toc44275528)

[3 Baseline Scores for Datasets 9](#_Toc44275529)

[4 Data Visualization 9](#_Toc44275530)

[4.1 Data exploration 9](#_Toc44275531)

[5 Feature Engineering 10](#_Toc44275532)

[6 Modelling 11](#_Toc44275533)

[7 Acquire new dataset 12](#_Toc44275534)

[7.1 Data Visualization 12](#_Toc44275535)

[7.1.1 Data Exploration 13](#_Toc44275536)

[8 Focus on New Zealand (2) Dataset 13](#_Toc44275537)

[9 Acquire More Feature Variables 13](#_Toc44275538)

[9.1 Adding latitude and longitude based on Meshblock 13](#_Toc44275539)

[9.2 Adding demographic features (Age group and Ethnicity) 14](#_Toc44275540)

[9.2.1 Data Pre-Processing 14](#_Toc44275541)

[9.2.2 Perform Modelling 15](#_Toc44275542)

[10 Task in progress 17](#_Toc44275543)

[10.1 Improve predictive modelling cycle 3 17](#_Toc44275544)

[10.2 Deep learning : PyTorch 17](#_Toc44275545)

[11 Conclusion 18](#_Toc44275546)

[13 Appendices 19](#_Toc44275547)

[13.1 Data Visualization result for NZ (1) Dataset 19](#_Toc44275548)

[13.2 Data Visualization result for NZ(2) dataset 20](#_Toc44275549)

[13.3 Pandas profile report for NZ(1) dataset, https://bitbucket.org/vimi\_vaidya/predictive\_policing-vimi/src/master/PP\_NZ\_Data/Output/profile\_report.html 24](#_Toc44275550)

[13.4 Pandas profile report for NZ(2) dataset, https://bitbucket.org/vimi\_vaidya/predictive\_policing-vimi/src/master/PP\_NZ\_Data/Output/profile\_report\_new\_NZ\_Data.html 24](#_Toc44275551)

[13.5 Feature Selection technics result 25](#_Toc44275552)

[13.6 New Zealand Classification report 25](#_Toc44275553)

[13.7 New Zealand crime dataset with latitude and longitude feature 25](#_Toc44275554)

[13.8 Feature selection score by adding Age group and ethnicity in NZ Crime dataset 26](#_Toc44275555)

[13.9 Comparison of Improve cycle 1, 2 with baseline score 27](#_Toc44275556)

**List of Table**

[Table 1. All datasets with column names. 7](#_Toc44275557)

[Table 2 Datasets with baseline scores. 9](#_Toc44275558)

[Table 3 Result of feature engineering 10](#_Toc44275559)

[Table 4 Modelling result after apply feature engineering technic 11](#_Toc44275560)

[Table 5 New Zealand(2) crime dataset features 12](#_Toc44275561)

[Table 6 With latitude and longitude feature 14](#_Toc44275562)

[Table 7 Modelling result comparison 15](#_Toc44275563)

[Table 8 Extra Tree classifier result 25](#_Toc44275564)

[Table 9 SelectKBest result 25](#_Toc44275565)

[Table 10 SelectKBest result with age group and ethnicity feature 26](#_Toc44275566)

**List of Figure**

[Figure 1 Hierarchy of New Zealand crime category 8](file:///D:\Level%207%20GDIT\Block%204&5\Mini%20Project\Phase%202%20Study%20different%20machine%20learning%20algo\Documentation\History%20of%20work\checkpoint3.docx#_Toc44275567)

[Figure 2 pandas-profiling result 10](#_Toc44275568)

[Figure 3 NZ Crime and Safety Survey report findings. Retrieved from http://www.justice.govt.nz/assets/Documents/Publications/NZCASS-201602-Main-Findings-Report-Updated.pdf 14](#_Toc44275569)

[Figure 4 extra tree classifier result with age group and ethnicity feature 26](#_Toc44275570)

# Workspace

* The repository is located at <https://bitbucket.org/vimi_vaidya/predictive_policing-vimi/src/master/>
* In my main repository includes 3 different projects,
  + 1. Building\_Modeling : include the codes and result of Denver dataset.
  + 2. PP\_NZ\_Data : include the analysis, codes, and result of New Zealand dataset only.
  + 3. Predictive\_Policing: Final main directory of code where the structure of workspace and program organized.

## High-Level Project Structure

The project consists of data, source code and output folders. The below project structure describes the different structure and project that I worked on.

* PP\_NZ\_Data

At the beginning I worked on Denver dataset. However, our main goal is to focus on New Zealand dataset.

|  |  |  |
| --- | --- | --- |
| Folder | File | Description |
| Dataset | /[NZ datasets](https://bitbucket.org/vimi_vaidya/predictive_policing-vimi/src/master/PP_NZ_Data/Dataset/NZ%20datasets/)  /[crimes-in-boston](https://bitbucket.org/vimi_vaidya/predictive_policing-vimi/src/master/PP_NZ_Data/Dataset/crimes-in-boston/)  /[denver-crime-data](https://bitbucket.org/vimi_vaidya/predictive_policing-vimi/src/master/PP_NZ_Data/Dataset/denver-crime-data/" \t "_self)  /[london-crime](https://bitbucket.org/vimi_vaidya/predictive_policing-vimi/src/master/PP_NZ_Data/Dataset/london-crime/)  /[sanfranciso-crime-dataset](https://bitbucket.org/vimi_vaidya/predictive_policing-vimi/src/master/PP_NZ_Data/Dataset/sanfranciso-crime-dataset/) | Store all datasets raw data, transformed and clean CSV file |
| Output | .png image file , .txt model result file | Store output in the format of an image file and text file |
| [Predictvie\_policing](https://bitbucket.org/vimi_vaidya/predictive_policing-vimi/src/master/PP_NZ_Data/Predictvie_policing/) | [Load\_Dataset](https://bitbucket.org/vimi_vaidya/predictive_policing-vimi/src/master/PP_NZ_Data/Predictvie_policing/Load_Dataset/)  [Explore\_Dataset(Descriptive\_Analysis)](https://bitbucket.org/vimi_vaidya/predictive_policing-vimi/src/master/PP_NZ_Data/Predictvie_policing/Explore_Dataset(Descriptive_Analysis)/)  [Data\_Transform](https://bitbucket.org/vimi_vaidya/predictive_policing-vimi/src/master/PP_NZ_Data/Predictvie_policing/Data_Transform/)  [Model](https://bitbucket.org/vimi_vaidya/predictive_policing-vimi/src/master/PP_NZ_Data/Predictvie_policing/Model/)  [Aquire\_new\_feature](https://bitbucket.org/vimi_vaidya/predictive_policing-vimi/src/master/PP_NZ_Data/Predictvie_policing/Aquire_new_feature/) | The folder includes the different python code file. |

* Predictive\_Policing

This directory is my leading directory when we start working on New NZ dataset.

**(Predictive Policing)**

├──README.md <- The top-level README for developers using this project.

├── data

│ ├── external <- Data from third party sources.

│ ├── interim <- Intermediate data that has been transformed.

│ ├── processed <- The final, canonical data sets for modelling.

│ └── raw <- The original, immutable data dump.

│

├── docs <- A default documents which define the project details

│

├── models <- Save Trained model in .pkl extension

├── references <- Data dictionaries, manuals, and all other explanatory materials.

│

├── reports <- Generated analysis as HTML format output of panadas profiling.

│ └── image <- Generated output in the image to be used in reporting

│

├── requirements.txt <- The requirements file for reproducing the analysis environment,

│

│

├── src <- Source code for use in this project.

│ ├── \_\_init\_\_.py <- Makes src a Python module

│ │

│ ├── data <- Scripts to load or read data

│ │ └── read\_dataset.py <- code to read CSV file

| | └── prepare\_dataset.py <- code to prepare a dataset

| | └── clean\_dataset.py <- code to clean CSV file

│ │

| |

│ │

│ ├── features <- Scripts to turn raw data into features for modelling

│ │ └── feature\_engineering.py <- code of different feature engineering techniques

│ │ └── feature\_selection.py <- code of different feature selection techniques

│ |

│ ├── models <- Scripts to train models and then use trained models to make

│ │ │ predictions

│ │ ├── predict\_model.py <- code of different modelling techniques

│ │ └── train\_model.py <- code of split data into train, validation, and test

│ │

│ └── visualisation <- Scripts to create exploratory and results-oriented visualisations

│ │ └── visualize.py <- code which will plot graphs to evaluate modelling results

│ │

└── ├──split\_data.py <- Scripts to separate data into train and test

├──run.py <- Final Scripts to run all functions in one file

# Crime Dataset

There are 5 crime datasets provided for this project. For the New Zealand crime dataset, we need to acquire a new crime dataset because the first dataset does not contain the Day of Week and Hour of Day columns.

The table below shows all 6 datasets with their columns. Most of them have some common information such as day, month, year, time, and location except for London and New Zealand (1) dataset that they do not have day and time values. This missing information would impact the accuracy of the prediction.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Boston** | **Denver** | **London** | **San Francisco** | **New Zealand (1)** | **New Zealand (2) (nz\_victim\_timeplace.csv)** |
| INCIDENT\_NUMBER | INCIDENT\_ID | lsoa\_code | IncidntNum | Age Group 5Yr Band | ANZSOC Division |
| OFFENSE\_CODE | OFFENSE\_ID | borough | Category | ANZSOC Division | ANZSOC Group |
| OFFENSE\_CODE\_GROUP | OFFENSE\_CODE | major\_category | Descript | ANZSOC Group | ANZSOC Subdivision |
| OFFENSE\_DESCRIPTION | OFFENSE\_CODE\_EXTENSION | minor\_category | DayOfWeek | ANZSOC Subdivision | Area Unit |
| DISTRICT | OFFENSE\_TYPE\_ID | value | Date | Person/Organisation | Table 1 |
| REPORTING\_AREA | OFFENSE\_CATEGORY\_ID | year | Time | Table 1 | Location Type |
| SHOOTING | FIRST\_OCCURRENCE\_DATE | month | PdDistrict | Selected Period | Locn Type Division |
| OCCURRED\_ON\_DATE | LAST\_OCCURRENCE\_DATE |  | Resolution | Previous Period | Meshblock |
| YEAR | REPORTED\_DATE |  | Address | Age Group | Number of Records |
| MONTH | INCIDENT\_ADDRESS |  | X | % Variance | Occurrence Day Of Week |
| DAY\_OF\_WEEK | GEO\_X |  | Y | Ethnicity | Occurrence Hour Of Day |
| HOUR | GEO\_Y |  | Location | Ethnic Group | Territorial Authority |
| UCR\_PART | GEO\_LON |  | PdId | Months Ago | Victimisations |
| STREET | GEO\_LAT |  |  | Mop Division | Weapon |
| Lat | DISTRICT\_ID |  |  | Mop Group | Year Month |
| Long | PRECINCT\_ID |  |  | Mop Subdivision | Year Month (copy 2) |
| Location | NEIGHBORHOOD\_ID |  |  | Method of Proceeding | Month Year |
|  | IS\_CRIME |  |  | Number of Records |  |
|  | IS\_TRAFFIC |  |  | Police Area |  |
|  |  |  |  | Police District |  |
|  |  |  |  | Proceedings |  |
|  |  |  |  | Year Month |  |
|  |  |  |  | Ytd Month |  |

Table 1. All datasets with column names.

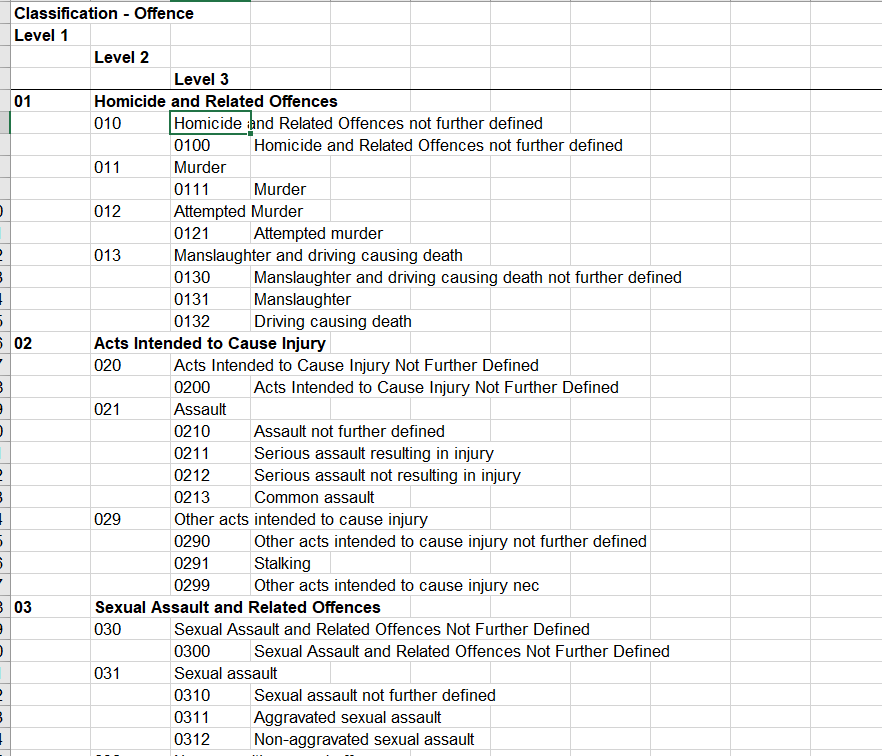
* In the above table, features are highlighted with three different colours, which are separated in the different group. The yellow highlighted features represent the features which describe the category of crime, blue highlighted features give location-related information and green describe the time information.
* After analysing above dataset with their different features, we noticed that for the New Zealand dataset mainly in the category of crime has different behaviour than another dataset which is based on classification structure provided by New Zealand Police[[1]](#footnote-1).
* The New Zealand Dataset has a different classification of crimes that has a hierarchy of three levels, according to ANZSOC(Australian and New Zealand Standard Offence Classification) StatsNZ.

Figure 1 Hierarchy of New Zealand crime category

# Baseline Scores for Datasets

We quickly test the dataset with various algorithms to generate baseline scores. The scores are very low, which could be because of the quality of the feature variables, the number of categories in the response variable, number of records, and imbalanced data.

Table 2 Datasets with baseline scores.

|  |  |  |  |
| --- | --- | --- | --- |
| **Algorithms** | **Denver** | **New Zealand (1)** | **New Zealand (2) (nz\_victim\_timeplace.csv)** |
| LogisticRegression | 0.26 | 0.16 | 0.57 |
| DecisionTreeClassifier | 0.33 | 0.10 | 0.54 |
| XGBClassifier | 0.40 | 0.18 | 0.63 |

# Data Visualization

**Main code:** Data\_Exploration.py

**Input:** /PP\_NZ\_Data/data/Dataset/NZ datasets/ Consolidated\_NZ\_Dataset.csv

**Output:**

1. Bar graphs with counts of all feature and response variables and pie chart to see the ethnicity percentage. Please see [appendix 10.1](#_Data_Visualization_result)

## Data exploration

For more understanding of each feature, we use the pandas-profiling tool. This tool is used for a quick data analysis which gives an overview of the dataset, each variable information with distinct count and missing values.

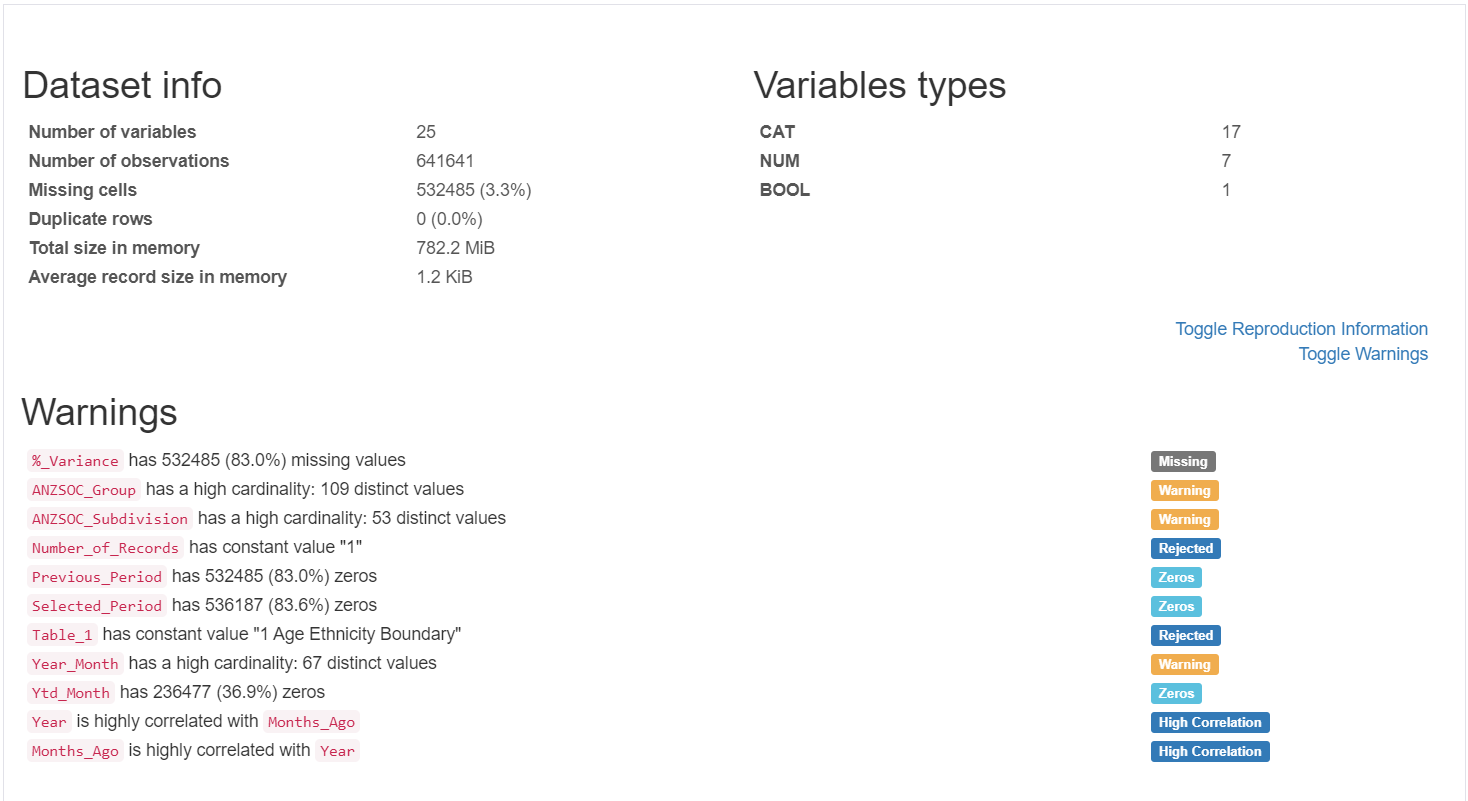


Figure 2 pandas-profiling result

* To see the whole pandas-profiling report please see [appendix 10.3](#_Pandas_profile_report)

# Feature Engineering

**Main code:** Data\_Transformation.py

**Input**: /PP\_NZ\_Data/data/Dataset/NZ datasets/ Clean\_Data/NZ\_Clean\_Data\_Set.csv

**Output:** Statistics, New transformed csv file

Outcome:

1. Extract unique label from the dataset.
2. Replace each variable by its count (Feature Engineering: Count/ Frequency encoding)
3. Store data into CSV file for modelling

Table 3 Result of feature engineering

|  |  |
| --- | --- |
| Identify unique labels | Crime\_Type : 53 labels  Age\_Group : 19 labels  Nationality : 11 labels  Police\_Area : 40 labels  Month : 12 labels  Year : 7 labels |
| Result by performing Count encoding technic. |  |
| NZ\_Transformed\_Data\_Set.csv | <https://bitbucket.org/vimi_vaidya/predictive_policing-vimi/src/master/PP_NZ_Data/Dataset/NZ%20datasets/Clean_Data/> |

# Modelling

**Main code:** Old\_NZ\_modeling.py

**Input:** /PP\_NZ\_Data/data/Dataset/NZ datasets/ Clean\_Data/NZ\_Transformed\_Data\_Set.csv

**Output:**

1. Accuracy score.
2. Classification Report
3. Confusion matrix

Table 4 Modelling result after apply feature engineering technic

|  |  |
| --- | --- |
| Algorithm | After apply count frequency encoding feature engineering technic result |
| Logistic Regression | 0.11 |
| XGBoost | 0.14 |
| Decision Tree | 0.13 |

* For the NZ(1) dataset, we do not have Location and time features, and after performing feature engineering technics, the accuracy score is low.
* To apply count frequency encoding technic which does not expand the feature space, however, If 2 labels appear the same amount of times in the dataset, that is, contain the same number of observations, they will be merged: may lose valuable information

# Acquire new dataset

For New Zealand(1) dataset, we do not have more information for location and time data, and the accuracy score is very low due to the quality of features and unbalanced data.

Dataset is located at,

|  |  |
| --- | --- |
| Nz\_victim\_timeplace.csv | <https://teams.microsoft.com/_#/school/files/General?threadId=19%3A3342246000ff4fb79193257873a76452%40thread.tacv2&ctx=channel&context=nz_victim_timeplace&rootfolder=%252Fsites%252FPredictivePolicing%252FShared%2520Documents%252FGeneral%252FDatasets%252FNZ%2520datasets%252Fnz_victim_timeplace> |

* Below table describe the features of NZ\_victim\_timeplace.csv. The yellow highlighted feature describes the category of crime, green represent the location information and blue give the information about time variables.

Table 5 New Zealand(2) crime dataset features

|  |  |  |
| --- | --- | --- |
| ANZSOC Division |  | |
| ANZSOC Group | Category of Crime | |
| ANZSOC Subdivision |  | |
| Area Unit | |  | |
| Table 1 | |  | |
| Location Type | Location info | |
| Locn Type Division |  | |
| Meshblock |  | |
| Number of Records |  | |
| Occurrence Day Of Week | | |
| Occurrence Hour Of Day | | |
| Territorial Authority |  | |
| Victimisations |  | |
| Weapon | |  | |
| Year Month | Time info | |
| Year Month (copy 2) |  | |
| Month Year |  | |

## Data Visualization

**Main code:** Data\_Acquisition.py

**Input:** /PP\_NZ\_Data/data/Dataset/NZ datasets/ Nz\_victim\_timeplace.csv

**Output:**

* Bar graphs with counts of all feature and response. Please see [appendix 10.2](#_Data_Visualization_result_1)

### Data Exploration

We performed pandas-profiling report to understand variables in detail. To view the pandas profile report for New Zealand(2) dataset, please visit [appendix 10.4](#_Pandas_profile_report_1)

# Focus on New Zealand (2) Dataset

The difference between New Zealand (1) and New Zealand (2) dataset is that New Zealand (2) dataset has day and time values, which we can use them as feature variables.

# Acquire More Feature Variables

## Adding latitude and longitude based on Meshblock

Main code: New\_NZ\_Data\_modelling.py

Input: /PP\_NZ\_Data/data/Dataset/NZ datasets/ Nz\_victim\_timeplace.csv

Output:

1. Feature’s scores by different algorithms
2. Store updated data with latitude and longitude feature in new CSVsv file, New\_NZ\_Crime\_dataset.csv
3. Accuracy score
4. Classification report
5. Confusion matrix

Checklist:

* With latitude and longitude feature

feature\_variable = ['Week', 'Day', 'Month', 'Year', 'Latitude', 'Longitude']

target\_variable = ['Crime\_Type']

To do list:

1. Perform SelectKBest and extra-tree classifier feature selection technics. Please see [appendix 13.4](#_Feature_Selection_technics)
2. Replace ANZSOC Division, Occurrence Day of Week, Locn Type Division and Weapon based on classification report code. Please see the [appendix 13.5](#_New_Zealand_Classification) for NZ classification report file.
3. For NZ\_victim\_timeplace dataset we have Meshblock column and based on this Meshblock data by performing Left-join operation we merge latitude and longitude in [New\_NZ\_Crime\_dataset.csv](#_New_Zealand_crime).
4. Perform Machine learning algorithms, Logistic regression, XGBoost and Decision Tree.

Table 6 With latitude and longitude feature

|  |  |
| --- | --- |
| Algorithm | Accuracy score |
| Logistic Regression | 0.98 |
| XGBoost | 1.0 |
| Decision Tree | 0.99 |

* By adding two extra feature Latitude and Longitude we faced overfitting issue. We got 100% accuracy which relates to overfitting issue.

## Adding demographic features (Age group and Ethnicity)

* Based on New Zealand Crime and Safety Survey (NZCSS) report Age, Gender, Ethnicity and Relationship/ partnership status are more likely to be the victim of a violence offence.

A screenshot of a cell phone

Description generated with very high confidence

Figure 3 NZ Crime and Safety Survey report findings. Retrieved from <http://www.justice.govt.nz/assets/Documents/Publications/NZCASS-201602-Main-Findings-Report-Updated.pdf>

* Feature selection result by adding age and ethnicity on NZ crime dataset, please see [appendix 12.8](#_Feature_selection_score) for score

### Data Pre-Processing

**Main code:** prepare\_dataset.py, clean\_dataset.py

**Location:** <https://bitbucket.org/vimi_vaidya/predictive_policing-vimi/src/master/Predictive_Policing/src/data/>

**To do – Feature Engineering:**

1. Read the raw data from CSV file
2. Drop the unused columns
3. perform feature selection
4. Rename the columns
5. Rearrange columns
6. Export the CSV for interim process
7. load process data which has selected features after feature selection.
8. apply feature transformation on Crime\_Type, Weapon and Day\_of\_Week columns based on NZ classification structure code.
9. apply feature engineering techniques to converted age group , ethnicity, and area unit categorial data into numeric.
10. store clean data into new csv file as NZ\_Cleaned\_Data.csv for splitting and modelling operation

**Takeaway:**

1. We have a new and cleaned New Zealand crime dataset, ‘NZ\_Cleaned\_Data.csv’.
2. We have 3 crime types to predict:
   1. Theft and Related Offences
   2. Unlawful Entry with Intent/Burglary, Break and Enter
   3. Acts Intended to Cause Injury

### Perform Modelling

**Main code:** temp.py

**Input**:/Predictive\_Policing/data/processed/NZ\_Cleaned \_Data.csv

**Output:** Comparison of score with and without Age group and ethnicity variables

1. Accuracy score
2. Classification report
3. Confusion matrix

Table 7 Modelling result comparison

|  |  |  |
| --- | --- | --- |
| Algorithm | Without Age group and ethnicity | With Age group and ethnicity |
| Logistic Regression | 0.56 | 0.56 |
| XGBoost | 0.61 | 0.89 |
| Decision Tree | 0.59 | 0.88 |
|  |  |  |

# Task in progress

## Improve predictive modelling cycle 3

1. Split data into train, validation, and test (60:20:20) ratio.
2. Save separate CSV file of train, validation, and test data.
3. Save model as a .pkl file
4. run .pkl model file to predict result
5. Evaluate all of our saved models on the validation set
6. Select the best model based on performance on the validation set
7. Evaluate that model on the holdout test set

## Deep learning : PyTorch

After we reach 0.9 accuracy score with machine learning algorithm, we work on deep learning approach.

Deep learning code is located at, <https://bitbucket.org/vimi_vaidya/predictive_policing-vimi/src/master/Predictive_Policing(Deep%20Learning)/>

# Conclusion

1. We can improve the scores by feature engineering, but we are facing the overfitting problem in tree-based algorithms.
2. When the train and test parts are 70% and 30% respectively, we have about 90% accuracy.
3. Comparison of improve predictive modelling cycle 1 and 2 with baseline score is in [appendix 13.9](#_Comparison_of_Improve).
4. We expect to improve the quality of models by parameter tuning and acquiring more data.
5. Next task, we will work on Deep Learning.

# Appendices

## Data Visualisation on result for NZ (1) Dataset

|  |  |
| --- | --- |
| Label | Result |
| Number of crimes by year |  |
| Number of crimes by Month |  |
| Top Crimes |  |
| Top 10 Ethnicity percentage of crime |  |

## Data Visualisationon result for NZ(2) dataset

|  |  |
| --- | --- |
| Crime based on year |  |
| Crime based on month |  |
| Crime based on day of week |  |
| Crime based on hour of day |  |
| Crime based on number of records |  |
| Crime based on Locn type division |  |
| Crime based on weapon |  |

## Pandas profile report for NZ(1) dataset, <https://bitbucket.org/vimi_vaidya/predictive_policing-vimi/src/master/PP_NZ_Data/Output/profile_report.html>

## Pandas profile report for NZ(2) dataset, <https://bitbucket.org/vimi_vaidya/predictive_policing-vimi/src/master/PP_NZ_Data/Output/profile_report_new_NZ_Data.html>

## Feature Selection technics result

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Table 8 Extra Tree classifier result   |  |  | | --- | --- | | **ExtraTree classifier** | **score** | | ANZSOC Group | 0.45 | | ANZSOC Subdivision | 0.37 | | Location Type | 0.12 | | Locn Type Division | 0.04 | | Occurrence Hour Of Day | 0.01 | | Occurrence Day Of Week | 0.01 | | Area Unit | 0.00 | | Meshblock | 0.00 | | Territorial Authority | 0.00 | | Weapon | 0.00 | | Victimisations | 0.00 | | Month Year | 0.00 | | Year Month | 0.00 | | Year Month (copy 2) | 0.00 | | Table 1 | 0.00 | | Number of Records | 0.00 | | Table 9 SelectKBest result   |  |  | | --- | --- | | **SelectKBest** | **score** | | Meshblock | 22672178.59 | | Location Type | 11775086.99 | | ANZSOC Group | 1145990.54 | | ANZSOC Subdivision | 776311.76 | | Area Unit | 427709.02 | | Occurrence Hour Of Day | 242718.00 | | Locn Type Division | 202103.12 | | Occurrence Day Of Week | 62543.47 | | Territorial Authority | 55941.36 | | Victimisations | 16218.72 | | Weapon | 2723.59 | | Year Month | 328.31 | | Year Month (copy 2) | 328.31 | | Month Year | 328.31 | |

## New Zealand Classification report

<https://teams.microsoft.com/l/file/DB1B9B95-1494-4DB7-9BD1-83203814EAA6?tenantId=450e6824-88ab-4ad2-914d-b0f385da600c&fileType=pdf&objectUrl=https%3A%2F%2Fotagopoly.sharepoint.com%2Fsites%2FPredictivePolicing%2FShared%20Documents%2FGeneral%2FDatasets%2FNZ%20datasets%2FNZ_Crime_with_new_feature%2Fnz_recorded_crime_victims_manual_v1.2.pdf&baseUrl=https%3A%2F%2Fotagopoly.sharepoint.com%2Fsites%2FPredictivePolicing&serviceName=teams&threadId=19:3342246000ff4fb79193257873a76452@thread.tacv2&groupId=5f39a75a-e978-4e1b-89d4-b73aa2fe9cd3>

## New Zealand crime dataset with latitude and longitude feature

File is located at, <https://otagopoly.sharepoint.com/sites/PredictivePolicing/Shared%20Documents/General/Datasets/NZ%20datasets/NZ%20New%20Dataset%20with%20Lat&Long%20feature/New_NZ_Crime_dataset.csv>

## Feature selection score by adding Age group and ethnicity in NZ Crime dataset

**Main code:** prepare\_dataset.py

**Input**:/Predictive\_Policing/data/raw/New\_Dataset\_with\_additional\_feature/NZTimeplace\_AEG\_Dataset.csv

* Extra Tree classifier feature selection technic

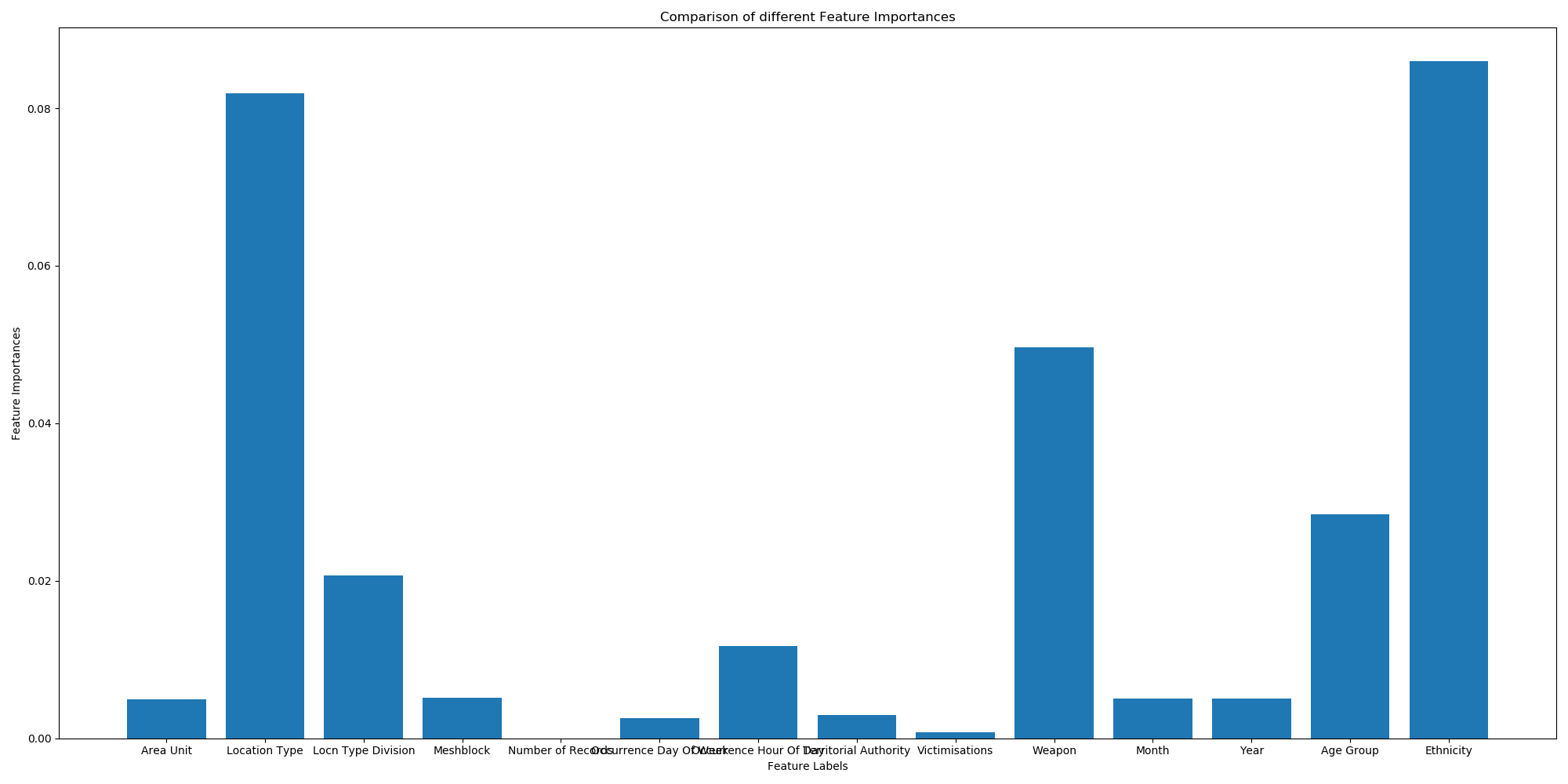


Figure 4 extra tree classifier result with age group and ethnicity feature

* SelectKBest feature selection technic

Table 10 SelectKBest result with age group and ethnicity feature

|  |  |  |
| --- | --- | --- |
| Score | Specs |  |
| 5 Meshblock 1.806401e+07 | |  |
| 3 Location Type 7.560930e+06 | |  |
| 0 ANZSOC Group 5.331636e+05 | |  |
| 1 ANZSOC Subdivision 4.701660e+05 | |  |
| 2 Area Unit 1.050337e+05 | |  |
| 4 Locn Type Division 6.148412e+04 | |  |
| 15 Ethnicity 4.834579e+04 | |  |
| 9 Territorial Authority 2.935014e+04 | |  |
| 8 Occurrence Hour Of Day 2.189761e+04 | | |
| 10 Victimisations 1.389316e+04 | |  |
| 14 Age Group 5.438754e+03 | |  |
| 11 Weapon 2.553997e+03 | |  |
| 13 Year 2.806161e+02 | |  |
| 12 Month 2.630459e+02 | |  |
| 7 Occurrence Day Of Week 6.804215e+01 | | |

## Comparison of Improve cycle 1, 2 with baseline score

|  |  |  |  |
| --- | --- | --- | --- |
| **Algorithms** | **Baseline score** | **Improve cycle 1 (Acquire new feature latitude and longitude)** | **Improve cycle 2**  **(Acquire new feature Age group and Ethnicity)** |
| LogisticRegression | 0.57 | 0.98 | 0.56 |
| DecisionTreeClassifier | 0.54 | 1.0 | 0.88 |
| XGBClassifier | 0.63 | 0.99 | 0.89 |

1. <http://archive.stats.govt.nz/methods/classifications-and-standards/classification-related-stats-standards/offence.aspx> [↑](#footnote-ref-1)