Data Mining and Predictive Analytics individual project

Anna Gaplanyan

17, May, 2021

Contents

Overview

Criminality is a major danger to humanity. It is growing and spreading at a rapid and broad scale. The police face significant challenges in crime prediction and criminal detection. Hence, the **goal** of the project is to use various data mining tools and techniques in order to predict the crime rate and detect crime hotspots based on locations. This project will aid police forces in forecasting and identifying crime in a given location and diminish the crime rates.

Literature review

In order to do comprehensive analyzes, I have researched by reading various articles regarding this topic.

Some authors used Multi Linear regression for forecasting the percapita of Crime rate, and the k-nearest neighbors algorithm (k-NN) and Logistic Regression models were also tested, but the Multi Linear regression produced minimal error while training the model (Mahendra et al., 2020).

Other authors used Apriori Algorithm. They did this in order to come up with a list of all crime hotspots along with their related frequent time. They also applied Multinomial Naive Bayes, which is used for multinomial distributed data that conforms to the categorical features in their datasets. Additionally, they created a Decision Tree Classifier model, and for evaluating the quality of the split, they applied the entropy function for the information gain (Almanie, et al, 2015).

The authors used clustering methods to help investigators anticipate and eliminate criminal activity. They used the K-means clustering algorithm (Sangani,et al, 2019).

Other authors used K-Nearest Neighbor and Naive Bayes algorithms in order to detect the places that are inclined to offense (Reddy, et al, 2018). Other authors used Decision Trees and Random Forest Classification, Naive Bayes Classification, and Linear Regression in order to predict the topmost features that affect the high crime rate (Yerpude & Gudur, 2017).

Other authors did Time Series analysis in order to tackle the crime trends forecasting problem. They display how the number of crime incidents changed over time. They found some trends and seasonality in the data

and applied Triple Exponential Smoothing (Holt-Winters) to their data. They tested the Naive Bayes model by using cross-validation, and they used 60% of the data as training data and the rest for validation. They also run KNN classifier using rectangular kernel and k = 150 to train their data. After that, they found that larger k does not guarantee smaller errors. Larger k may lead to under-fitting. The results of KNN was better than Naive Bayes (Feng, et all, 2018). As the variables of the data of this project are numeric, I use **KNN regression**, **Random forest for regression**, and **Multiple linear regression** to make predictions of crime rates of local authority areas. The **Classification analysis** for this project will not work. I use **K -Means** algorithm to group local authority areas based on their features, as some of the mentioned authors did in their research.

Research Methodology

The data contains information about the crime rate during two emigration flows in the **United Kingdom**. From this data, I choose the part for only 2008 year. The data is obtained from the **Harvard Dataverse**. The variables of the data set are the names of police force areas, the names of local authority areas, id of local authority area, year of the observation, number of violent crimes reported by year and local authority, number of burglaries reported by year and local authority, number of robberies reported by year and local authority, the number of thefts of motor vehicles reported by year and local authority, number of thefts from motor vehicles reported by year and local authority, number of female asylum seekers is dispersal accommodation by year and local authority, number of male asylum seekers is dispersal accommodation by year and local authority, total number of asylum seekers is dispersal accommodation by year and local authority, number of female asylum seekers receiving subsistence support by year and local authority, number of male asylum seekers receiving subsistence support by year and local authority, total number of asylum seekers receiving subsistence support by year and local authority, total number asylum seekers by year and local authority, total estimated population, mid-year (Office for National Statistics), total estimated population Aged 15-24, mid-year (Office for National Statistics), total estimated population Aged 0-14, mid-year (Office for National Statistics), A8 registrations on the worker registration scheme by year and local authority, claimant count unemployment rate, mid-year (Office for National Statistics), total benefit claimants (Department for Work and Pensions), predicted inflow of A8 immigrants. In this project, in the **Visualizations** part, I investigate which variable might be useful for predicting the crime rate by creating a **correlation matrix**. I create a**bar plot** to demonstrate the top 10 local authority areas with the highest crime rate. This will help to detect crime hotspots. I create another **bar plot** to illustrate the top 10 local authority areas with the lowest crime rate. It will allow identifying the safest areas. I also create a map of England and Wales and paint the local authority areas according to their crime rate. As the data variables are numeric, I use **KNN regression**, **Random forest for regression**, and **Multiple linear regression** to make predictions of crime rates of local authority areas. As predictor variables, I choose the features that have high correlation coefficients with the crime rate. For **Clustering analysis**, I use the **K -Means** algorithm to group local authority areas based on their features.

Analysis

Visualizations

Correlation coefficient is a measure of the strength of the relationship between two variables.

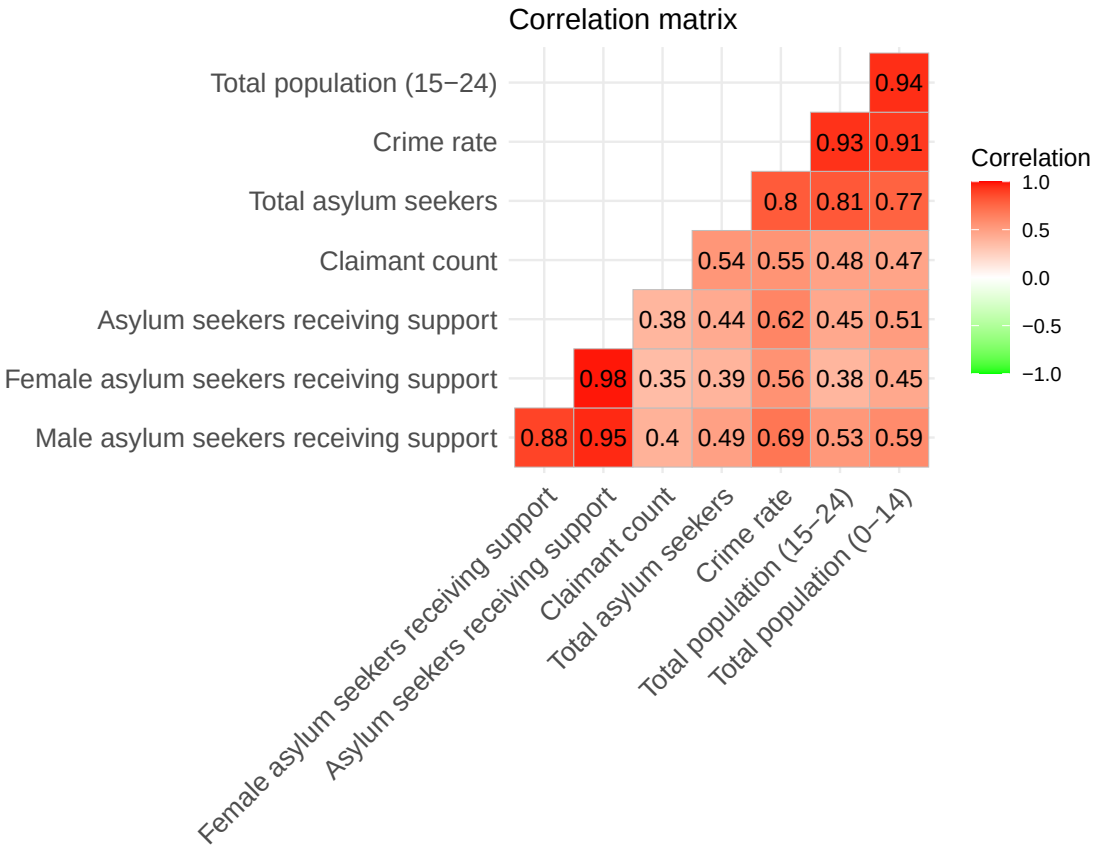

Figure 1: Correlation matrix

According to the **correlation matrix**, the relationships of crime rate and the total number of asylum seekers, number of male asylum seekers receiving support, number of asylum seekers receiving support, number of female asylum seekers receiving support, and claimant count rate are strong. This is because the correlation coefficients are greater than 0.5.

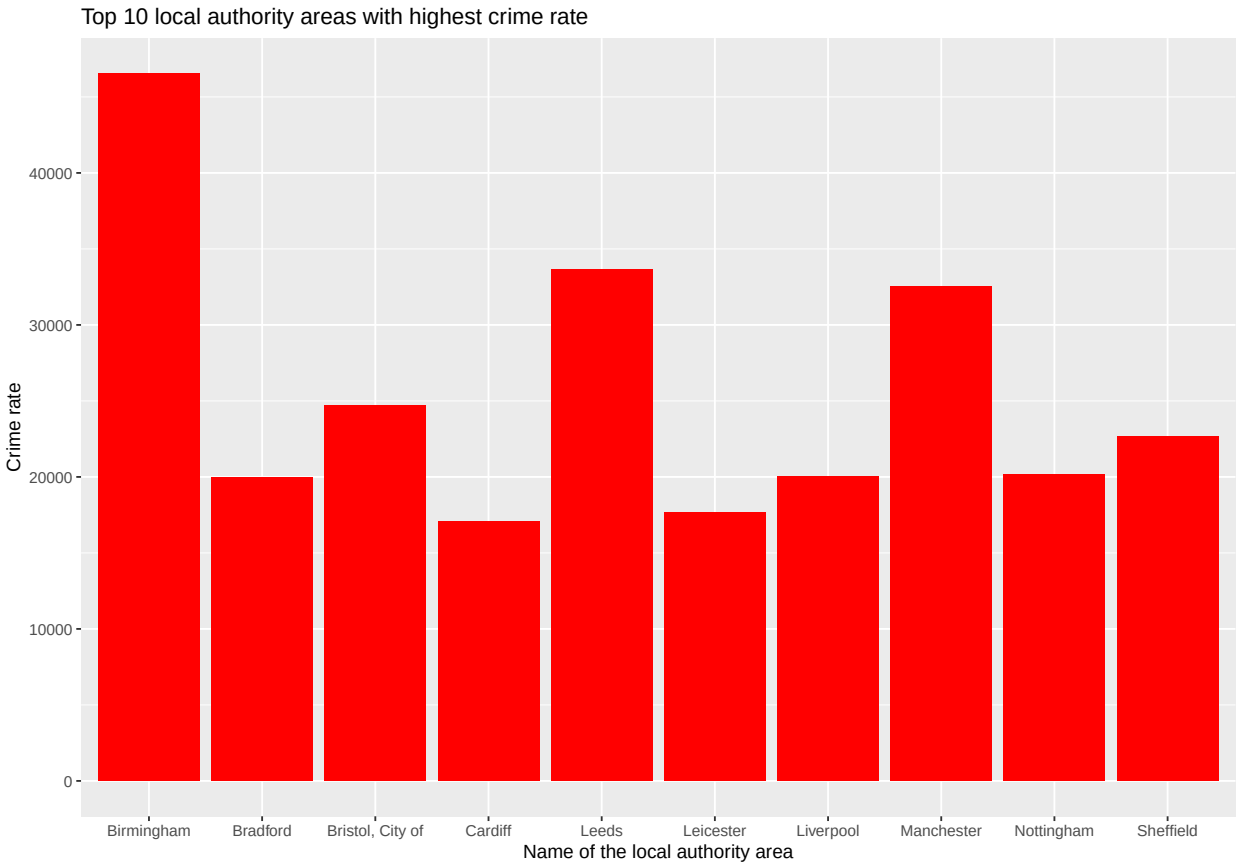

Figure 2: Bar plot

In this bar plot, I have demonstrated the ten local authority areas with the highest crime rate. As it is shown, the highest crime rate has the Birmingham local authority area.

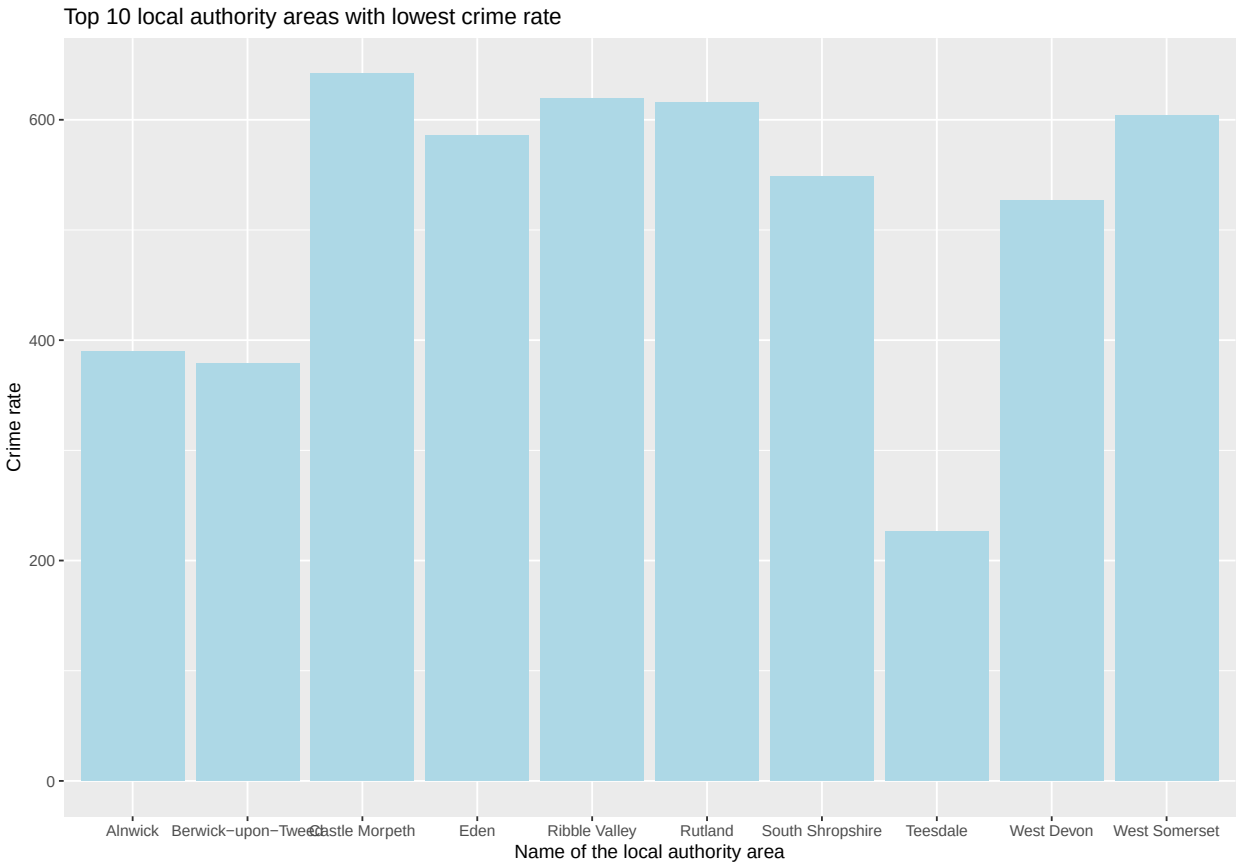

Figure 3: Bar plot

In this bar plot, I have demonstrated the ten local authority areas with the lowest crime rate. As it is shown, the lowest crime rate has the Teesdale local authority area.

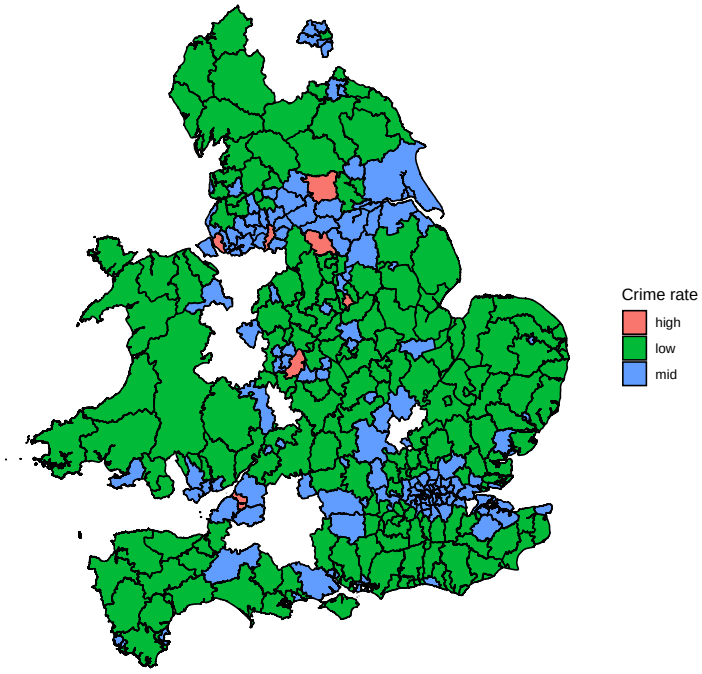

Figure 4: Map of of England and Wales

The **red** points show the **LAs** that have a high crime rate, and the **green** points show **LAs** that have a low crime rate.

KNN regression

The prediction that **KNN regression** made provides an average for the nearest neighbors. For **KNN regression**, I select the variables with a high correlation coefficient with the crime rate. I divide the data set into training and testing sets by using an 80/20 ratio.

```
##  Total asylum seekers Male asylum seekers receiving support
##  Min.   :   0.00      Min.   :  0.00
##  1st Qu.:   0.00      1st Qu.:  0.00
##  Median :   1.00      Median :  0.00
##  Mean   :  76.22      Mean   :  6.67
##  3rd Qu.:  20.75      3rd Qu.:  5.00
##  Max.   :1450.00      Max.   :105.00
##  Asylum seekers receiving support Female asylum seekers receiving support
##  Min.   :  0.00                   Min.   :  0.000
##  1st Qu.:  0.00                   1st Qu.:  0.000
##  Median :  1.00                   Median :  0.000
##  Mean   : 16.66                   Mean   :  9.989
##  3rd Qu.:  6.00                   3rd Qu.:  5.000
##  Max.   :420.00                   Max.   :325.000
##  Claimant count   Crime rate
##  Min.   :0.50   Min.   :  226
##  1st Qu.:1.10   1st Qu.: 1892
##  Median :1.65   Median : 2920
##  Mean   :1.88   Mean   : 4841
```

```
##  3rd Qu.:2.40   3rd Qu.: 5869
##  Max.   :5.30   Max.   :46541
```

I normalize the range of the variables using **Z-score normalization** as the range of the variables varies on a large scale. I normalize the data in order to decrease the influence of the arbitrary variable on the model.

As the square root of the number of observations is equal to 19.23538., I calculate the **RMSE** for different values of **k** starting from one to the square root of the number of observations.

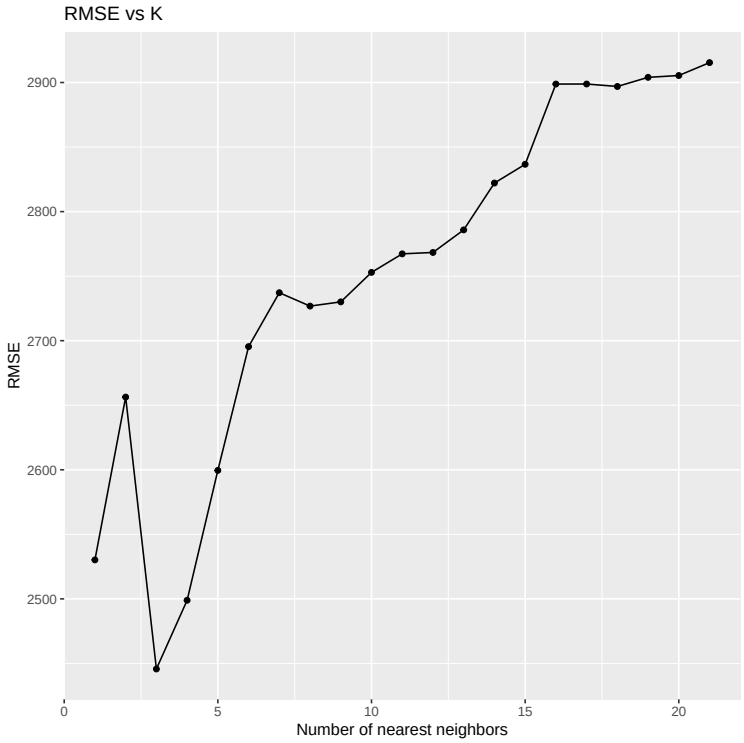

Figure 5: RMSE vs K

The **optimal value** of **k** is equal to **three**.

The root mean squared error for the test dataset is equal to 2400.341. The closer the root mean squared error is to zero, the more accurate the model is. Hence, it can be stated that the model is inaccurate.

K-Means Algorithm

In this project, I use the **K-Means algorithm** in order to divide the **LAs** into groups based on their characteristics. I normalize the range of the variables by using **Z-score normalization**. In order to find the optimal value of **k**, I use the **Elbow method**.

I use **fviz_nbclust()** function from **factoextra** package in order to create **Elbow curve**.

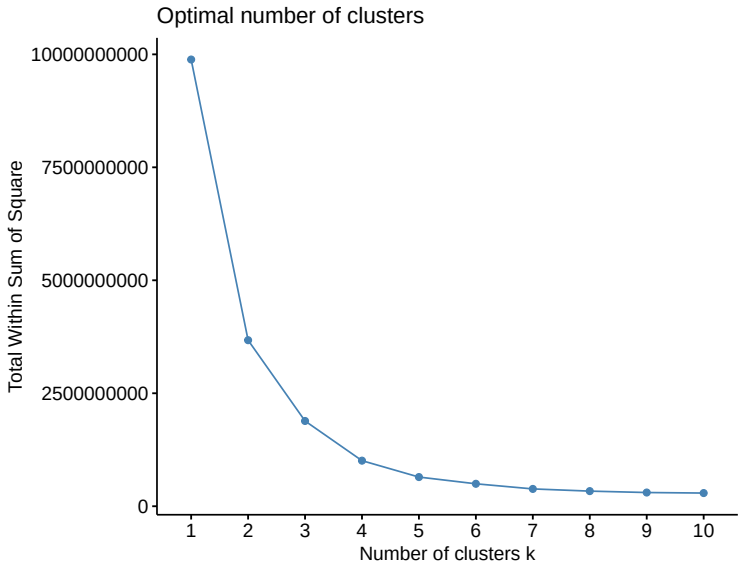

Figure 6: Elbow curve

According to the graph, it can be stated that by increasing the number of k, the total within sum of squares decreases. The optimal number of k for the **k-means algorithm** is the number for which the decrease in **WSS** will be small as the number of k increases. In this case, the optimal value for **k** is **three**.

I evaluate the model performance using **internal measures** such as **Silhouette coefficient**, **Dunn index**, **Connectivity**.

Connectivity is **13.1921**, which means that not all the nearest neighbors are in the same cluster. The **Silhouette coefficient** is equal to **0.7290**, and it can be concluded that the clustering is good as it is close to one. The **Dunn index** is **0.0066**. This means that **min.separation** is lower and **max.diameter** is higher. As the **min.separation** is lower, between cluster distance is lower. As the **max. diameter** is higher, the within-cluster distances are higher.

80.91 percent of the total variance in the data can be explained by the clusters dividing Between Groups Sum of Squares by Total Sum of Squares.

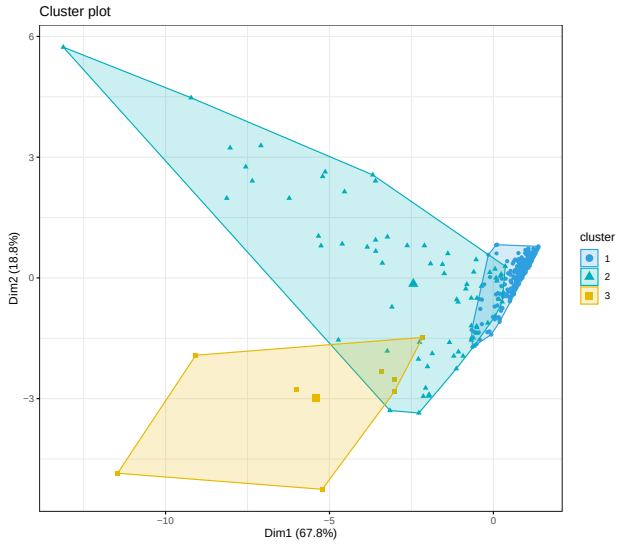

Figure 7: Cluster plot

According to the **Cluster plot**, the spread of the points in cluster number 1 is less than the spread of cluster number 2 and 3. This means that within group sum of squares for cluster number one is less than for cluster number 2 and 3.

Random forest for regression

In random forest for regression algorithm, I choose the variables that have high correlation coefficients with the crime rate. I divide the data into **training** and **testing** set by using 80/20 proportion.

I define the value of **mtry** to be equal to the square root of the number of the predictor.

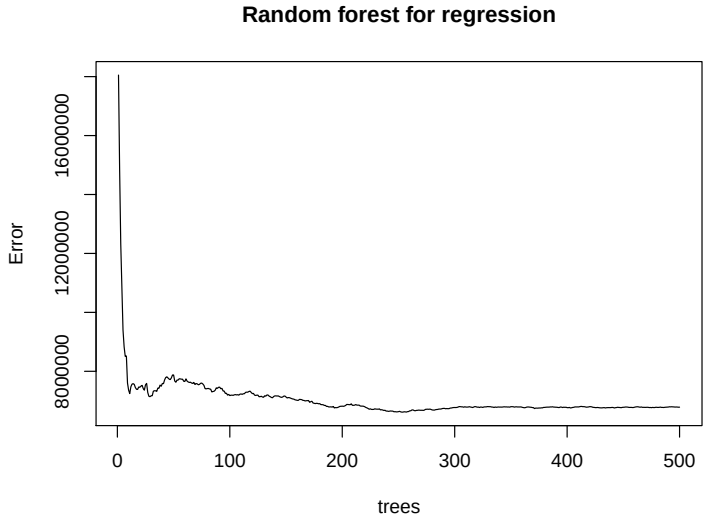

Figure 8: Error vs trees

According to the graph, the error is decreasing by adding more and more trees and average them.

The optimal number of trees is equal to 253.

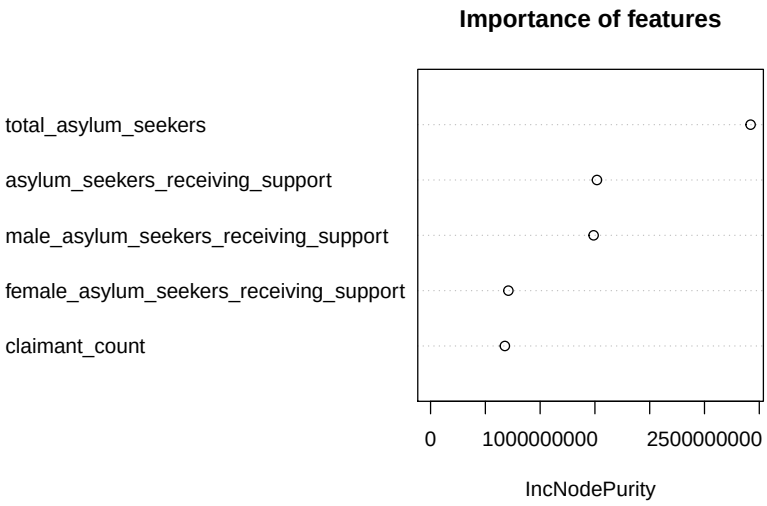

Figure 9: Importance of features

For the **Mean Decrease Gini (IncNodePurity)**, the most important variable is the **total number of asylum seekers**.

The root mean squared error is equal to 2649.095. The closer the root mean squared error is to zero, the more accurate the model is. Therefore, it can be stated that the model is inaccurate.

Multiple linear regression

I use **Multiple linear regression** to make predictions of crime rates of local authority areas.

I create a correlation matrix for independent variables. As the correlation coefficient between Total population (15-24) and Total population (0-14) is 0.9418742, I eliminate the Total population (15-24) variable (multicollinearity).

```
##                                        Total asylum seekers
## Total asylum seekers                              1.0000000
## Male asylum seekers receiving support             0.4886139
## Female asylum seekers receiving support           0.3906785
## Claimant count                                    0.5396670
## Total population (0-14)                           0.7663966
## Total population (15-24)                          0.8062696
##                                        Male asylum seekers receiving support
## Total asylum seekers                                                0.4886139
## Male asylum seekers receiving support                               1.0000000
## Female asylum seekers receiving support                             0.8800271
## Claimant count                                                      0.4013208
## Total population (0-14)                                             0.5904653
## Total population (15-24)                                            0.5281479
##                                        Female asylum seekers receiving support
## Total asylum seekers                                                 0.3906785
## Male asylum seekers receiving support                                0.8800271
## Female asylum seekers receiving support                              1.0000000
## Claimant count                                                       0.3517004
## Total population (0-14)                                              0.4457313
## Total population (15-24)                                             0.3811708
##                                        Claimant count Total population (0-14)
## Total asylum seekers                        0.5396670               0.7663966
## Male asylum seekers receiving support       0.4013208               0.5904653
## Female asylum seekers receiving support     0.3517004               0.4457313
## Claimant count                              1.0000000               0.4702233
## Total population (0-14)                     0.4702233               1.0000000
## Total population (15-24)                    0.4761352               0.9418742
##                                        Total population (15-24)
## Total asylum seekers                                  0.8062696
## Male asylum seekers receiving support                 0.5281479
## Female asylum seekers receiving support               0.3811708
## Claimant count                                        0.4761352
## Total population (0-14)                               0.9418742
## Total population (15-24)                              1.0000000
```

```
##                                       Crime rate Total asylum seekers
## Crime rate                                  1.00                  0.80
## Total asylum seekers                        0.80                  1.00
## Male asylum seekers receiving support       0.69                  0.49
## Female asylum seekers receiving support     0.56                  0.39
## Claimant count                              0.55                  0.54
## Total population (0-14)                     0.91                  0.77
## Total population (15-24)                    0.93                  0.81
##                                       Male asylum seekers receiving support
## Crime rate                                                             0.69
## Total asylum seekers                                                   0.49
## Male asylum seekers receiving support                                  1.00
## Female asylum seekers receiving support                                0.88
## Claimant count                                                         0.40
## Total population (0-14)                                                0.59
## Total population (15-24)                                               0.53
##                                       Female asylum seekers receiving support
## Crime rate                                                               0.56
## Total asylum seekers                                                     0.39
## Male asylum seekers receiving support                                    0.88
## Female asylum seekers receiving support                                  1.00
## Claimant count                                                           0.35
## Total population (0-14)                                                  0.45
## Total population (15-24)                                                 0.38
##                                       Claimant count Total population (0-14)
## Crime rate                                      0.55                    0.91
## Total asylum seekers                            0.54                    0.77
## Male asylum seekers receiving support           0.40                    0.59
## Female asylum seekers receiving support         0.35                    0.45
## Claimant count                                  1.00                    0.47
## Total population (0-14)                         0.47                    1.00
## Total population (15-24)                        0.48                    0.94
##                                       Total population (15-24)
## Crime rate                                                0.93
## Total asylum seekers                                      0.81
## Male asylum seekers receiving support                     0.53
## Female asylum seekers receiving support                   0.38
## Claimant count                                            0.48
## Total population (0-14)                                   0.94
## Total population (15-24)                                  1.00
##
## n= 370
##
##
## P
##                                       Crime rate Total asylum seekers
## Crime rate                                     0
## Total asylum seekers                           0                    0
## Male asylum seekers receiving support          0                    0
## Female asylum seekers receiving support        0                    0
## Claimant count                                 0                    0
## Total population (0-14)                        0                    0
## Total population (15-24)                       0                    0
##                                       Male asylum seekers receiving support
```

```
## Crime rate                                 0
## Total asylum seekers                       0
## Male asylum seekers receiving support
## Female asylum seekers receiving support    0
## Claimant count                             0
## Total population (0-14)                    0
## Total population (15-24)                   0
##                                           Female asylum seekers receiving support
## Crime rate                                 0
## Total asylum seekers                       0
## Male asylum seekers receiving support      0
## Female asylum seekers receiving support
## Claimant count                             0
## Total population (0-14)                    0
## Total population (15-24)                   0
##                                           Claimant count Total population (0-14)
## Crime rate                                 0              0
## Total asylum seekers                       0              0
## Male asylum seekers receiving support      0              0
## Female asylum seekers receiving support    0              0
## Claimant count                                            0
## Total population (0-14)                    0
## Total population (15-24)                   0              0
##                                           Total population (15-24)
## Crime rate                                 0
## Total asylum seekers                       0
## Male asylum seekers receiving support      0
## Female asylum seekers receiving support    0
## Claimant count                             0
## Total population (0-14)                    0
## Total population (15-24)
```

I use **vif** function in order to measure the amount of multicollinearity in a set of multiple regression variables. It helps to understand whether one of the independent variables is highly correlated with others or not.

```
##                     `Total asylum seekers`
##                                   2.696263
##   `Male asylum seekers receiving support`
##                                   5.675734
## `Female asylum seekers receiving support`
##                                   4.627173
##                           `Claimant count`
##                                   1.467616
##                 `Total population (0-14)`
##                                   2.942381
```

As the variance inflation factor of the **Male asylum seekers receiving support** variable is greater than five, it should be removed. After removing it I run linear model.

```
##
## Call:
## lm(formula = `Crime rate` ~ `Total asylum seekers` + `Female asylum seekers receiving support` +
##     `Claimant count` + `Total population (0-14)`, data = mult_reg)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5378.8  -675.3  -173.9   416.1 11334.5
##
## Coefficients:
##                                          Estimate   Std. Error t value
## (Intercept)                           -1472.695728   259.492060  -5.675
## `Total asylum seekers`                    5.027457     0.790525   6.360
## `Female asylum seekers receiving support`   27.492662     3.383510   8.125
## `Claimant count`                          460.805754   114.049190   4.040
## `Total population (0-14)`                   0.187073     0.008219  22.760
##                                                 Pr(>|t|)
## (Intercept)                          0.00000002824636941 ***
## `Total asylum seekers`               0.00000000060506028 ***
## `Female asylum seekers receiving support`  0.00000000000000698 ***
## `Claimant count`                     0.00006508558842395 ***
## `Total population (0-14)`                < 0.0000000000000002 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1773 on 365 degrees of freedom
## Multiple R-squared:  0.8839, Adjusted R-squared:  0.8826
## F-statistic: 694.8 on 4 and 365 DF,  p-value: < 0.00000000000000022
```

I measure the amount of multicollinearity in a set of multiple regression variables one more time.

```
##                     `Total asylum seekers`
##                                   2.695264
## `Female asylum seekers receiving support`
##                                   1.290863
##                           `Claimant count`
##                                   1.464881
##                 `Total population (0-14)`
##                                   2.591686
```

After removing the **Male asylum seekers receiving support** variable, there is no variance inflation factor greater than five. In case when the variance for all observations is not the same, **heteroskedasticity** occurs. I check the heteroscedasticity by using **bptest** function from **lmtest** package. This function does the **Breusch-Pagan** test.

```
##
##  studentized Breusch-Pagan test
##
## data:  model_2
## BP = 50.147, df = 4, p-value = 0.0000000003364
```

The **p-value** is equal to 0.0000000003364, which is less than alpha. Hence, the null hypothesis should be rejected. This means that there is heteroscedasticity. As there is heteroscedasticity, the ordinary least squares no longer produce the best linear unbiased estimators **BLUE**, and standard errors estimated using least squares can be incorrect. Therefore, the useage of multiple linear regression is not valid.

Conclusion and Recommendations

During the project, I have done visualizations to represent the top 10 crime hotspots and the top 10 safe areas. I visualize the map of England and Wales and paint the local authority areas according to their crime rate. By creating **correlation matrix**, I figure out that the relationships of crime rate and the total number of asylum seekers, number of male asylum seekers receiving support, number of asylum seekers receiving support, number of female asylum seekers receiving support, and claimant count rate are strong. I use **KNN regression** and get that the **root mean squared error** for the test dataset is equal to **2400.341**. I use the **K-Means algorithm** to divide the **LAs** into clusters based on their features. The connectivity is equal to 13.1921, Dunn index is equal to 0.0066, and the Silhouette is equal to 0.7290. **80.91** percent of the total variance in the data can be explained by the clusters dividing Between Groups Sum of Squares by Total Sum of Squares. In the **Random forest for regression**, I find out that the optimal number of trees is equal to 253, and the most important variable for the **Mean Decrease Gini (IncNodePurity)** is the **total number of asylum seekers**. **The root mean squared error** for the test set is equal to **2649.095**. In **Multiple linear regression** I observe that there is a statistically significant relationship between **crime rate** and **Total asylum seekers**, **Female asylum seekers receiving support**, **Claimant count**, and **Total population (0-14)** As heteroscedasticity is detected in the **multiple linear regression**, it makes the method invalid. As the **RMSE** of the **KNN regression** is less than the **RMSE** of the **Random forest for regression**, I recommend using **KNN regression** to make predictions of crime rates of local authority areas.

References

1. A. Sangani, C. Sampat, V. Pinjarkar, Crime prediction and analysis, in Proceedings of 2nd International Conference on Advances in Science&Technology, SSRN: Elsevier, India (2019), pp. 1?5. 2.Ch. Mahendra, G. Nani Babu, G. Balu Nitin Chandra , A. Avinash , Y. Aditya. (2020, May 5). CRIME RATE PREDICTION. 3.Ginger Saltos, Ella Haig. (2017, May). An Exploration of Crime Prediction Using Data Mining on Open Data.
2. H. Toppi Reddy, B. Saini, G. Mahajan. (2018). Crime prediction & monitoring framework based on spatial analysis. Proc. Comput. Sci. 132, 696?705. 5.Mingchen Feng, Jinchang Ren, Qiaoyuan Liu. (2018, July). Big Data Analytics and Mining for Crime Data Analysis, Visualization and Prediction: 9th International Conference, BICS 2018, Xi'an, China, July 7-8, 2018, Proceedings. 6.Nahid Jabeen, Parul Agarwal. (2021, January). Data Mining in Crime Analysis. 7.Prajakta Yerpude, Vaishnavi Gudur. (2017, July). PREDICTIVE MODELLING OF CRIME DATASET. 8.Tahani Almanie, Rsha Mirza, Elizabeth Lor. (2015, July). CRIME PREDICTION BASED ON CRIME TYPES.