Chanya Deshani Subasingha Arachchige

Abstract

[Draw your reader in with an engaging abstract. It is typically a short summary of the document.   
When you’re ready to add your content, just click here and start typing.]

**Assignment Report**

CT7202 Data Analysis and Visualisation Principles

Table of Contents

[1 Introduction 2](#_Toc101385222)

[2 Hypothesis 2](#_Toc101385223)

[2.1Number of national court cases in the United Kingdom for theft and handling, burglary and robbery are decreasing from 2015 June to 2018 March. 2](#_Toc101385224)

[3 Data preparation 2](#_Toc101385225)

[3.1Data Cleaning 3](#_Toc101385226)

[3.2Feature engineering 3](#_Toc101385227)

[3.3Missing value imputation 3](#_Toc101385228)

[4 Descriptive analysis 4](#_Toc101385229)

[5 linear regression 10](#_Toc101385230)

[6 Classification 12](#_Toc101385231)

[7 Clustering 14](#_Toc101385232)

# Introduction

The Crown Prosecution Service (CPS) publishes a monthly report on outcome of CPS proceedings in magistrates' courts and in the crown court by principal offence category. The data set contains twenty-seven monthly court outcome reports for the thirty-three months period from July 2015 to March 2018, the period after new the new data assurance regime was introduced [1]. Each report contains number and percentage of convictions and unsuccessful conviction by defendant basis.

One court case can represent several observations in the data set as there can be number of defendants in one court case [1]. Principal offence is deciding at the time of case finalization. They are categorized as homicide, offences against the person, sexual offences, burglary, robbery, theft and handling, fraud and forgery, criminal damage, drugs offences, public order, motoring, or other offences excluding motoring. The data set also contains column for administrative finalisations, the cases which could not proceed due to an administrative issue such as unexecuted warrant for the arrest of the defendant, or summons have not served by the police because they were unable to trace the defendant, or the defendant has died or is unfit to plead. These cases are summarized without categorising by into principal offences [1].

# Hypothesis and research questions

## Number of national court cases in the England and wales for theft and handling, burglary and robbery are decreasing from July 2015 to March 2018.

## Is there a seasonal variation in total drug offences court cases for the period from July 2015 to March 2018 in England and wales?

## Total court cases for motoring offences are dependent on the area and can be classified using court name for the period from July 2015 to March 2018 in England and wales.

# Data preparation

All the data files have same columns. Each data file has 51 variables holding number of court outcomes and percentage of court outcomes, and 43 observations representing each magistrates' courts, crown courts and national values.

First, I noticed that the column names had long names with spaces which make it hard to work with in R functions as it is not compatible with R variable names.

It is possible to remove spaces with below code.

colnames(dataframe)<-str\_replace\_all(colnames(dataframe), c(" " = "" ))

However, the data column names are too long, and the code lines tend to get longer with the long column names. Therefore, all the long column names renamed to their abbreviations. This also make the code more user-friendly and readable. I used capital letters for the column names to make it clear that they are abbreviations. Then a new data frame was created merging all the data sets. As a preparation to combine datasets, new variable *Date* was introduced to all the data sets using *lubridate* library. Even the data set only has the month and year, date was added as first day of the respective month for the simplicity. Merging data sets reduce code complexity and processing time as one data set can be processed with better performance than looping through all the data sets each time.

## Data Cleaning

The dataset examined for missing values but there are no missing values in the data file for a particular month. There are some observations recorded with “-“ in the percentage column in homicide percentages columns, these are not missing values but resulted by zero number of court cases for respective month. However, some months were missing when considering whole period as can be seen in Figure 3.1.

Chart, bar chart, histogram

Description automatically generated

Figure 3.1 Number of theft and handling convictions from July 2015 to March 2018

Then data types of the dataset were checked. Columns with number of court cases are numeric, the columns with percentages of court outcomes have character type values. These ppercentages columns were converted to numeric because numerical values can carry more information and more insight can be drawn from them. I wrote a code to first, remove percentage sign, and then convert values to numeric values.

Then after a round of data visualizing with box plots and bar graphs, noticed that the data set containing national values of the principal offences court outcomes have extremely lager values. As shown in Figure 4.1, national values which are outliers and much larger than all the other values, hinders the data visualisation and identifying patterns and prominent features. Hence, I separated national data from the data set and copied to another data frame. Which can analyse separately.

Column *percentage of L motoring offences unsuccessful* has removed from the data set as this column has values only 100 and NA. The column has no data variation resulting it not valuable to analysis for pattern recognition or prediction.

## Feature engineering

Total number of finalized court cases for each offence category was calculated using the summation of number of convictions and number of unsuccessful cases. The new variables *total\_theft\_handling, total\_burglary, total\_robbery, total\_cases* (for drug offence) and *total\_motoring* was created for theft and handling, burglary, robbery, drug offence and motoring offence respectively.

Date column was used to extract month and created another column, then the month column was used to create season column using conditional statement. Month and season columns are used in the analysis to find patterns that are seasonal. The variable *motor\_cat* was introduced to store levels of success rates for motoring offences. Percentage of successful conviction separated in to three categories, average, high and very high depending on the value. The court names are string values they are not able to use in classification. Therefore, they are converted to numeric values and stored in the new column *court\_num*.

## Missing value imputation

A column no\_months (number of months) created using the date column of the data set. Then using a vector containing the complete value set for number of months, filled the missing value of the data set. Linear interpolation [3] [4] [5]is used to fill missing values for columns 'N\_BC', 'N\_BU', 'N\_RC', ' N\_RU', 'N\_THC', 'N\_THU' as the column values has a downward trend. Using simple imputations such as mean, or median will produce values that go against the trend. Hence, the linear imputation, which uses values closer to missing data point to impute missing data was a more appropriate choice to preserve the trend.

# Descriptive analysis

Analysis caried out plotting bar graphs against court name and time separately to identify patterns such as trends, or seasonality in the data. Once such pattern is identified those columns are analysed further. Histograms and box plots are used to visualise distribution on the data and spot outliers.

Initial descriptive analysis using box plot shows that there are outliers present in the dataset, the box plot is used here as it enables clearly identify outliers than other visualisation method such as bar plots. National data is identified as an extreme outlier as shown in

figure 4.1. It is separated from the dataset as described in section 3.1(Data cleaning) because having an outlier which is far apart from all the other values hinders the analysis by suppressing the features in the rest of the data. National data being an outlier is an obvious result as the national column contains the summation over all the other observation in the same column for the respective month. Metropolitan and city court area is also an outlier in all the categories. As can be seen from the Figure 4.7 these outliers are far apart from the other courts’ number of cases. This area has highest court cases among all courts in England and wales. There are other outliers identified but they vary with the type of offence.

Chart, scatter chart

Description automatically generated

Figure 4.1 Number of robbery conviction in July 2015

## Analysis of theft and handling, burglary, and robbery national court cases

National total number of finalized theft and handling court cases have minimum 5559 maximum 10114. National total number of finalized burglary and robbery has maximum of 1775 and 608 respectively. Number of theft and handling cases are much higher when comparing to robbery cases and burglary cases as shown in table 4.1.

The distribution of theft and handling cases is a multimodal and does not show skewness according to the Figure 4.2. The figure also shows that the distribution of the total national robbery court cases is skewed to left indicating that there are a smaller number of months with higher number of robbery cases finalised in courts. Burglary cases follows a distribution close to a normal distribution. Both distributions for burglary and robbery are unimodal distributions.

Table 4.1 Summary of the basic statistics for Theft and handling, robbery and burglary total national court cases form July 2015 to March 2018

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Variable | Min | 1st Quartile | Median | Mean | 3rd Quartile | Max | Range |
| Theft and handling court cases | 5559 | 6981 | 7793 | 7911 | 8783 | 10114 | 4555 |
| Burglary court cases | 982 | 1317 | 1395 | 1414 | 1516 | 1775 | 793 |
| Robbery court cases | 373 | 443 | 464 | 464 | 474 | 608 | 235 |

Chart

Description automatically generated

Figure 4.2 Number of robbery conviction in June 2015

It is clear from the figure 4.2 that the national data for theft and handling outcomes are following a clear downward trend. There is a spike in the first three month of 2017 for national number of theft and handling convictions, other than that only slight deviations from the downward trend are observed. The number of convictions seem to be higher in the first four month each year as seen in Figure 3.1. Even though, the unsuccessful convictions numbers are much smaller than convictions it also follows a downward trend.

Chart, bar chart, histogram

Description automatically generated

Figure 4.3 National number of theft and handling court cases from July 2015 to March 2018

Figure 4.4 shows that the national data for burglary outcomes are following a downward trend with slight variations. Unsuccessful convictions are also following a downward trend over the period. Minimum number of both convictions and unsuccessful conviction is reported in December 2017 which is towards the end of the period under analysis. Maximum number of burglary convictions are reported in October 2015 with over 1500 cases.

Chart, bar chart, histogram

Description automatically generated

Figure 4.4 National number of burglary court cases from July 2015 to March 2018

Total national robbery cases finalized have highest values which are around 600, at the beginning of the period in July 2015 and September 2015. Then the values slowly decreased with fluctuations to 375 in December 2016. Then shows a slight increase until March 2017 then decreased again and reach its lowest value 373 in December 2017. Overall, the values decrease over the period with fluctuations. Robbery convictions follows the same pattern as total court outcomes. However, the unsuccessful cases are mostly stays between 75 and 100. Relatively higher values are observed at the beginning of the period in July and September 2015 and January 2017, exceeding 125 cases.

Chart, bar chart

Description automatically generated

Figure 4.5 National number of robbery court cases from July 2015 to March 2018

Total numbers of case in robbery, burglary and theft and handling have a positive correlation with each other as shown in the scatter plot matrix in the Figure 4.6. Theft and handling and burglary cases seems to have a good linear relationship. Even though the number of robbery case does not follow a clear linear relationship with theft and handling total cases, it shows a positive relationship.

When compared with number of month (representing the time), all three variables shows negative correlation. Here also, theft and handling and burglary court cases have a strong negative correlation with number of month variable, while robbery court cases do not have such a strong relationship with number of month variable

Chart, scatter chart

Description automatically generated

Figure 4.6 Scatter plot matrix for national number of theft and handling court cases from July 2015 to March 2018

Correlation calculation of the national number of convictions and unsuccessful cases of burglary, robbery and theft and handling with date was done using Pearson correlation to measure linear relationship [6]. Burglary case outcomes and theft and handling case outcomes have a strong negative correlation with the date column as shown in Table 4.2 which suggest that both those offence categories have a good linear relationship with time. Robbery convictions also have a negative corelation, however, that is not strong as other two type of case outcomes.

There are outliers present in Robbery Convictions, Robbery unsuccessful and Burglary Convictions. Therefore, Spearman Correlation analysis is also carried out to identify monotonic relationship [6] as Pearson method is highly sensitive to outliers [6]. Both correlation analysis shows closer values. Spearman correlation results shows that there is a decline in the number court cases for burglary, theft and handling and robbery court cases over the time as both conviction and unsuccessful convictions are decreasing. Specially, burglary and theft and handling court cases show clear decline over the time.

Table 4.2 Correlation of the number of case outcome and date

|  |  |  |
| --- | --- | --- |
| Number of national case outcome | Pearson correlation with date | Spearman correlation with date |
| Theft And Handling Convictions | -0.883 | -0.861 |
| Theft And Handling Unsuccessful | -0.913 | -0.919 |
| Burglary Convictions | -0.761 | -0.761 |
| Burglary Unsuccessful | -0.815 | -0.851 |
| Robbery Convictions | -0.494 | -0.478 |
| Robbery Unsuccessful | -0.582 | -0.479 |

Chart, bar chart, histogram

Description automatically generated

## Analysis of drug offence court cases

Total drug offence court cases are visualised against court in the figure 4.7. It is clear that the Metropolitan and city court has the highest number of drug offences in England and wales for the period June 2015 to March 2018. West Midlands and Merseyside has next higher values however, they are around one fifth of the highest value. Warwickshire and Durham have the lowest drug related court cases. There is a variation of number of drug offence court cases finalized between courts.

Timeline

Description automatically generated with medium confidence

Figure 4.7 Total drug offence cases vs court they finalized from July 2015 to March 2018

Court names are represented by y axis to make the graph more readable as the court names are longer. If x axis were used for court names they would overlap. One step is to avoid over lapping is angle the x axis label, however switching axis resulted in more readable plot.

Table 4.3 Summary of the basic statistics for drug offence total court cases form July 2015 to March 2018

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Minimum | 1st Quartile | Median | Mean | 3rd Quartile | Maximum | Standard Deviation |
| 4.00 | 38.00 | 63.00 | 98.36 | 97.00 | 1281 | 164.5618 |

Table 4.3 shows the summary of the drug offence court cases in England and wales. The distribution has high standard deviation. First quartile, median, mean and third quartile lies closer when compared to the maximum. The values clustered towards higher end of the scale is resulted from the Metropolitan and city court when cross reference table 4.3 and Figure 4.7.

When looking at the histogram in Figure 4.8, it is evident that the distribution of drug offence court cases is highly skewed to left. Furthermore, there are no observations recorded after 400 until 800. In the histogram, it is clearly visible that how far apart the values for Metropolitan and city court lies.

Chart, histogram

Description automatically generated

Figure 4.8 Histogram for total drug offence cases

Total drug offence cases show decline over the period from July 2015 to March 2018 as visualised in Figure 4.9. The highest values recorded in Jan 2016 and lowest values recorded in December 2017. There is a spike in March 2017 which go against the downward trend beginning for the June 2016.

Chart, bar chart

Description automatically generated

Figure 4.9 Number of drug offence court cases from July 2015 to March 2018

Mean drug Offence court cases are highest in April according to the Figure 4.10 and lowest in December. There is slight variation in each month. Here mean number of cases per month is used to visualise the number of court cases each month rather than total number of cases as the number of data point for each month is different.

Chart, bar chart

Description automatically generated

Figure 4.10 Mean drug offence court cases per month from July 2015 to March 2018

There are two clear clusters in the month and total\_cases scatter plot, months with a low number of drug offence court cases and high number of drug offences court cases as shown in Figure 4.11. Small cluster in with high number of cases can be identified as data point representing Metropolitan and city when cross reference with Figure 4.7. Similarly, in the court\_num and tatal\_cases scatter plot, the cluster of data points separated from the other data can be identified as Metropolitan and city.

Diagram

Description automatically generated with medium confidence

Figure 4.11 Scatter plot matrix for total drug offence cases finalized from July 2015 to March 2018

## Analysis of motoring cases

Metropolitan city and Bedfordshire have reported more average success of convictions for motoring offences. Wiltshire, Suffolk, Norfolk and Durham have very high success rate for motoring court offences. There is no identifiable seasonality in the success rate of convictions for motoring offences as shown in Figure 4.12.

Table

Description automatically generated with medium confidence

Figure 4.12 Motoring court case success rate from July 2015 to March 2018

Table 4.4 Summary of the basic statistics for success rate of motoring offences court cases form July 2015 to March 2018

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Minimum | 1st Quartile | Median | Mean | 3rd Quartile | Maximum | Standard Deviation |
| 61.5 | 84.20 | 87.9 | 87.28 | 91.1 | 100.00 | 5.3 |

Motoring case success rate are generally high, most values lie over 70%. There are no observations recorded below 61.5% for this period as mentioned in Table 4.4. The distribution in Figure 4.13 shows that frequency of the values less than 70 are very low and highest frequency is recorded for the range 85% to 90%.

Chart, histogram

Description automatically generated

Figure 4.13 Histogram of motoring court case success rate from July 2015 to March 2018

# linear regression

Three separate simple linear regression models were fitted to total national number of cases for theft and handling, burglary, and robbery offences. Residuals are summarized in the table 5.1. Residuals of the total theft and handling cases and burglary cases are negative skewed. However, robbery cases have almost normal distributed residual distribution.

Table 5.1 Regression model summary of residuals

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Min | 1Q | Median | 3Q | Max |
| Total theft and handling cases | -1037.29 | -427.36 | 35.03 | 315.00 | 1276.55 |
| Total burglary cases | -250.66 | -45.87 | 22.92 | 60.26 | 180.39 |
| Total robbery cases | -86.191 | -18.020 | -1.118 | 24.613 | 102.176 |

Table 5.2 Regression model summary of coefficients

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | | Estimate | Std. Error | t value | Pr(>|t|) |
| Total theft and handling cases | Intercept | 9529.039 | 176.578 | 53.97 | < 2e-16 \*\*\* |
| no\_months | -101.129 | 9.484 | -10.66 | 6.76e-12 \*\*\* |
| Total burglary cases | Intercept | 1636.370 | 35.241 | 46.434 | < 2e-16 \*\*\* |
| no\_months | -13.921 | 1.893 | -7.355 | 2.8e-08 \*\*\* |
| Total robbery cases | Intercept | 511.7745 | 13.6157 | 37.587 | < 2e-16 \*\*\* |
| no\_months | -2.9755 | 0.7313 | -4.069 | 0.000302 \*\*\* |
| Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1 | | | | | |

# TODO obtain t qt(p=0.05/2, 2.048407 value for a two-sided df=28, lower.tail= test FALSE) at a 0.05 significance level  
*Data Science and Big Data Analytics : Discovering, Analyzing, Visualizing and Presenting Data*, edited by Education Services EMC, and Education Services, EMC, Wiley, 2015.*ProQuest Ebook Central*, http://ebookcentral.proquest.com/lib/uniofglos/detail.action?docID=4548030.  
Created from uniofglos on 2022-04-19 14:30:34. Page 130-131

According to the table 5.2, linear model shows that each month finalized theft and handling court cases are reduced by 101.129. Linear model for total burglary cases shows that total cases finalized for burglary offence is reducing by 13.921 each month. Similarly, the model for robbery cases shows that there is 2.9755 finalized case reduction each month.

Standard deviation of the *total theft and handling* column is 1103.147. Residual standard error is 518.8 as mentioned in Table 5.3. As the residual standard error is much smaller than the standard deviation of the *total theft and handling* variable, the fit is considered as a good fit [6]. Additionally, R squared is close to 0.8. Which means explanatory variable no of month can explain 77% of the variation in total theft and handling variable. R squared have a value closer to one is also supports that the regression line is a good fit for the variable total theft and handling [6].

Similarly, regression line for total burglary cases can be accepted as a good fit as it has 0.63 R squared value. On the other hand, the regression line for the total robbery cases does not have a good fit when considering the 0.34 R squared value. Standard deviation of the total robbery cases is 48.763. When compared the residual error 40 for total robbery cases from the table 5.3, it is clear that the regression line is not a good fit for total robbery cases. Because the standard deviation of the variable and residual standard error is much closer suggesting that using a linear regression is not much better than using mean to predict the total robbery cases.

Table 5.3 Regression model summary of coefficients

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Variable | Residual standard error  on 31 degrees of freedom | Multiple R-squared: | F-statistic: on 1 and 31 DF | p-value |
| Total theft and handling cases | 518.8 | 0.7858 | 113.7 | 6.761e-12 |
| Total burglary cases | 103.5 | 0.6357 | 54.1 | 2.796e-08 |
| Total robbery cases | 40 | 0.3481 | 16.56 | 0.0003015 |

All three models have p values are less than 0.001. Therefore, I can conclude that, there is very strong evidence against the null hypothesis [6] [7]. Linear model for the robbery and time has higher error rate than theft and handling and burglary. Therefore, a different prediction method should be used for robbery cases to obtain a better prediction such as nonlinear model.

Incase robbery explained with court variable.

This can lead to “fishing expeditions,” where you keep adding variables to an equation, some of which have no conceptual relationship to the dependent variable, just to inflate the R 2 value. To avoid adding extra variables that do not really belong, an adjusted R 2 value is typically listed in regression outputs. This adjusted value appears in cell D10 of Figure 10.25. Although it has no direct interpretation as “percentage of variation explained,” it can decrease when unnecessary explanatory variables are added to an equation. Therefore, it serves as an index that you can monitor. If you add variables and the adjusted R 2 decreases , the extra variables are essentially not pulling their weight and should probably be omitted. We will say much more about this issue  
Albright, S. Christian, and Wayne L. Winston. *Business Analytics : Data Analysis and Decision Making*, Cengage, 2019.*Page 440*

*ProQuest Ebook Central*, http://ebookcentral.proquest.com/lib/uniofglos/detail.action?docID=6135939.  
Created from uniofglos on 2022-04-30 17:43:17.

# Clustering

## Clustering drug offence cases finalized

Drug offence cases finalized are clustered with k-means algorithm. The variables total drug offence cases, month, and numeric transformation of the court variable is used for clustering. The scaled values of the variables are used here as k-means algorithm use Euclidean distance to create clusters, and the total court cases which has values much large than other attributes with maximum of 308 (refer Table 4.3), which can dominate the distance calculation with month and court variable with maximum of 12 and 42 respectively.

Chart

Description automatically generated with medium confidence

Figure 6.1 Within-cluster sum of squares for total Drug offence cases finalized data

Within-cluster sum of squares (Withinss) was calculated for different number of clusters, and it is used to identify the best value for number of clusters. The elbow of the curve of the withinss curve was identified as four [8] as the plot in Figure 6.1 shows. Therefore, four clusters were created using k-means algorithm. Those clusters are visualized in the Figure 6.2. The function fviz\_cluster used to draw clusters, transforms the initial set of variables into a new set of variables through principal component analysis (PCA) [9]. There are overlapping cluster boarders when dimension reduction was used to plot the clusters. However, in k-means clustering there are no over lapping points.

Chart, radar chart

Description automatically generated

Figure 6.2 Cluster plot for total drug offence cases finalized data with four clusters

Cluster centers are summarized in the table 6.1. The cluster number one can be identified as the cluster containing Metropolitan and city data as it has extreme z value for total number of cases. This can be verified by de-normalizing the z value for court number using mean and standard deviation of the court number column [11]. The de-normalized values are used to identify the court name of the cluster centres.

Table 6.1 Resulted cluster centers of k-means clustering for total drug offence cases finalized

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Cluster number | Total cases | De-normalized total cases | Month | De-normalized month(rounded) | Court number | Court name |
| 1 | 6.0459 | 1093.3 | 1.6447e-17 | 7 | 0.2061 | Metropolitan & City |
| 2 | -0.0531 | 89.6 | 0.5771 | 9 | 0.9655 | Staffordshire |
| 3 | -0.1399 | 75.3 | -1.2836 | 2 | -0.0117 | Leicestershire |
| 4 | -0.2365 | 59.4 | 0. 5938 | 9 | -0.85 | Dyfed Powys |

K-means clustering with three clusters and five clusters are also analyzed as the number of attributes are relatively small [8]. When created five clusters, the withinss has decreased

Table 6.2 Results different number of k-means clustering for total drug offence cases finalized

|  |  |  |
| --- | --- | --- |
| Number of clusters | Total withinss | Betweenss |
| 3 clusters | 1517.015 | 1881.985 |
| 4 clusters | 923.9763 | 2475.024 |
| 5 clusters | 653.266 | 2745.734 |

# Classification

Percentage of convicted court cases for motoring offences were categorized as average, high and very high. Then three dependent variables court, month, total motoring court cases were used to build a decision tree classifier.

Timeline

Description automatically generated

Figure 7.1 Decision tree classifier for percentage of motoring offence convictions

When looking at the right side of the tree, it is evident that when total court cases are low, court is playing main role of classifying the success rate of the motoring offence convictions. Therefore, it is possible to say that there are areas with lower success rate than other areas. Namely, when looking the tree in figure 7.1, Bedfordshire and Avon and Somerset have more average success rate. It possible to investigate causes and factors affecting the success rate of these court areas to get better understanding.

Table 6.2 Confusion matrix of the decision tree classifier

|  |  |  |  |
| --- | --- | --- | --- |
|  | average | high | very high |
| average | 8 | 18 | 0 |
| high | 0 | 144 | 15 |
| very high | 1 | 31 | 40 |

Accuracy of the model is 71.36 as calculated using equation below. The function calculate accuracy based on the confusion matrix.

accuracy <- function(x){sum(diag(x)/(sum(rowSums(x)))) \* 100}

Decision tree was further tuned using minimum split. Minimum split is set to 10% of the train data set to avoid overfitting as the data set can be considered as large with 907 observations in the training set [8].

Timeline

Description automatically generated

Figure 7.2 Decision tree classifier for percentage of motoring offence convictions with minimum split 90

The decision tree uses the total motoring offence and court to classify success rate of the motoring offences. When the total court cases for motoring offence is larger than or equal 136, it does not result very high success rate. Furthermore, when the total motoring court cases go above 604, the success rate become average. I can suggest looking into this factor further to improve success rate of convictions.

Table 6.2 Confusion matrix of the decision tree classifier

|  |  |  |  |
| --- | --- | --- | --- |
|  | average | high | very high |
| average | 4 | 15 | 7 |
| high | 1 | 109 | 19 |
| very high | 0 | 35 | 37 |

Accuracy of the new model is 66.08. The accuracy is reduced after changing the minimum split parameter.

According to the resulted decision trees, it is clear that the month is not a good branching factor for percentage of successful convictions. Therefore, I can conclude that there is no seasonality in the success rate of the convictions for motoring offences for the period July 2015 to March 2018 in England and Wales. Total number of court cases and court is good parameters to predict successful conviction rate for motoring offences for this period. Since the accuracy of the model is not very high, it is good to look for more independent variables to classify success rate. Once the factors contributing to success rate are identified, it is good to use random forest for predictions.