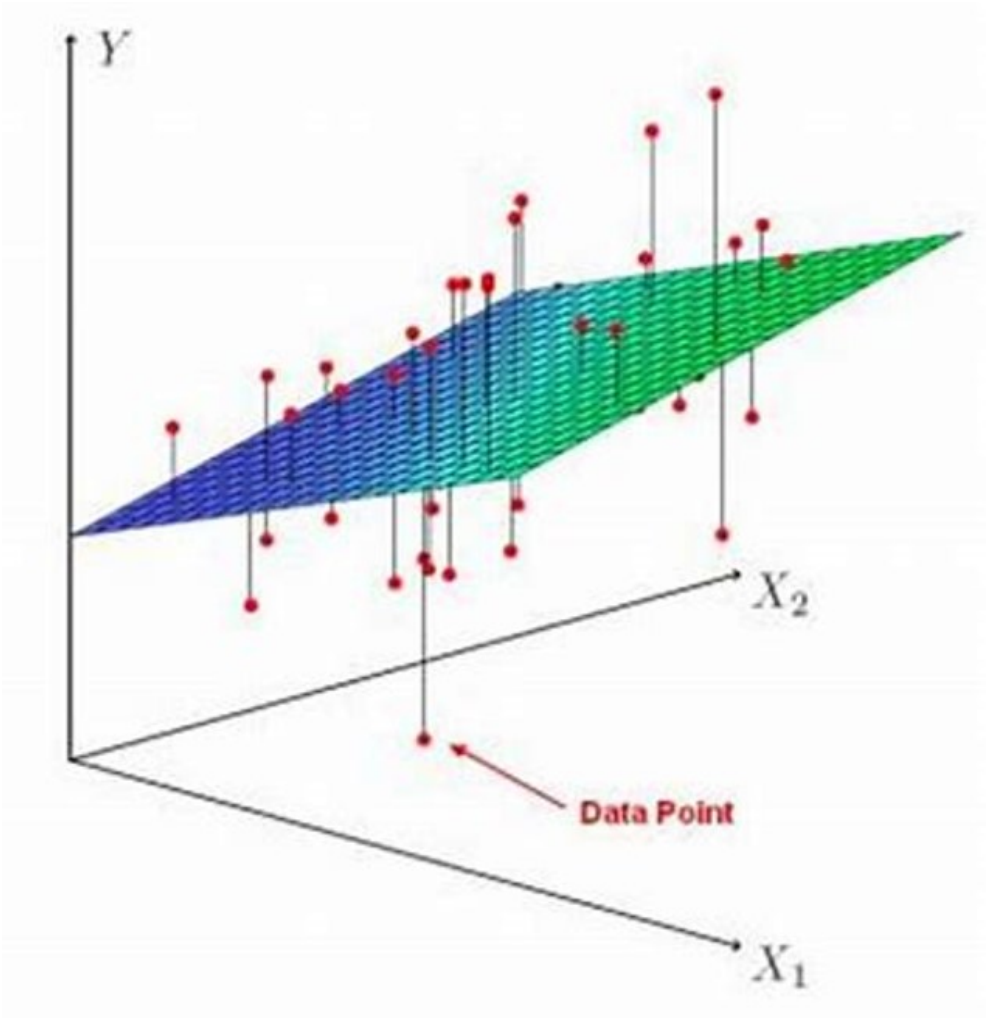

DIT Business Case Study

LexisNexis Insurance Claims Analysis

Jim Coen

May 8, 2018

Contents

Chapter 1

Introduction

Main Findings

- An alternate source of crime data lead to a model giving good predictions.
- Automatic variable selection is effective where there are a large number of predictors.
- The data preparation process can take an inordinate amount of time.

The primary research question is:
Is it possible to predict insurance claim score from a broad range of crime and demographic features?

Better predictions would allow insurance companies to offer more competitive prices. The analysis here follows a non-traditional approach to estimating insurance claim score. Instead of calculating ratings for individuals, this report looks at all persons within a postal district in an urban area. The available data covers postcode sectors in the Greater London area. The analysis objective is to draw a relationship between the crime and demographic characteristics of postcode sectors, on one hand, and the claim score for individuals within those sectors.

In answering the primary question, there were insights gained into the data itself and the analysis process. The data cleaning challenge was to combine data recorded at different interval levels. The difficulty for data analysis was to identify predictors for a linear model from 76 features in the combined dataset.

This report covers the mapping and transformation of variables from claims, crime and demographic observations. It was necessary to bring all observations to the same interval level, i.e. postcode sector. Correlation and variance accounted for by a model informed the set of predictors for linear models. These models had R-squared values ranging from 22% to 35%.

Automatic variable selection based on an "all relevant" algorithm identified a subset of features having higher importance that the main body of variables. However, the R-squared value was low and certain predictors were not significant. During this exploratory phase, discrepancies were noticed in the crime data. This prompted retrieval of data from an alternative source. Using the alternate data resulted in a model that satisfied the linearity assumption and produced an improved R-squared value.

Conclusions

- As the analysis work proceeds, it is important to check the data and to check the processes that produced that data.
- Automatic variable selection is a useful tool in helping to construct linear models.
- It is better to focus on accurate representations of the data. This helps set aside prior expectations and encourages acquiring domain knowledge.
- When the goal is to obtain a high R-squared value alone, it overlooks analysis assumptions and model quality.

Recommendations for future work

A modification to Principal Components Analysis that takes account of the response variable would be useful for data analysis with a large number of features. It could produce components that capture most of the variability in the data and form the predictors in linear models.

Limitations

The necessity to bring all observations to the same interval level resulted in loss of observations. There was a further reduction due to removal of sectors that were not representative.

Chapter 2

Linear Regression

Linear regression formulates the relationship between a response variable (y) and one or more explanatory variables (x_j) as a linear combination of the explanatory variables, given by equation 2.1. This relationship also encapsulates uncertainty about the response by including an error term (ϵ).

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 \cdots + \epsilon \tag{2.1}$$

The base value coefficient, β_0, is the point where the line crosses the y axis and is in units of the response variable. Often it will not have physical meaning. For this project, since the response variable (claim score) contains Z-distribution values, zero does have statistical meaning. Other coefficients give the rate of change in the response per unit change in the corresponding explanatory variable. For instance, in the analysis here, it could be change in claim score rating per unit change in crime rate.

The error term ϵ represents systematic errors in the measurement and recording process. The linear regression model makes the assumption that errors have a normal distribution with mean of zero and variance σ^2. This means that small errors are more likely than larger errors.

$$\epsilon \approx N(0, \sigma^2) \tag{2.2}$$

Equation 2.1 defines the underlying relationship for the population of postal districts in all urban areas. The coefficients in equation 2.1 represent the true, unmeasured values that pertain to a linear model of the whole population. In practice, linear regression models the underlying relationship by finding the line of best fit for a sample set of observations. This uses the least squares estimate, described in section 2.2.

2.1 Assumptions on Linear Regression

Linear regression applies specifically where the response is a continuous, numeric variable. Also, the following assumptions must be true for linear regression to be valid (Kabacoff, 2015).

1. **Linearity**
   It must be possible to equate the mean of the response to a linear combination of predictors (Wade & Koutoumanou, 2017). In addition, each predictor must have a linear relationship to the predictor. Otherwise, one range of values of a given predictor might overestimate the actual response while another set of values underestimate it (Lane et al., 2018).

2. **Independence**
   Predictors contribute independently to the response (no multicollinearity).

3. **Normality**
   For fixed values of a predictor variable, the observed response variable must have a normal distribution. This means that the residuals have a normal distribution.

4. **Homoscedasity**
   Residuals must have constant variance at all levels of each predictor.

2.2 Linear Regression Example

Simple Linear Regression involves one explanatory variable and the response variable. Figure 2.1 plots observations in the LexisNexis dataset. They are recordings of road traffic accidents (RTA) and claim score across all postcode (PC) sectors. Creating a linear model that represents adequately the distribution of these points involves 'fitting' a line to these points.

In this example, $\hat{y_i}$ represents the value of claim score estimated by the linear model. x_i is the i^{th} value of RTA incidents for postcode sector i. The linear model for these observations is the regression line given by equation 2.3.

$$\begin{aligned} \hat{y_i} &= \hat{\beta}_0 + \hat{\beta}_1.x_i \\ \hat{Claim}Score &= \hat{\beta}_0 + RTA.\hat{\beta}_1 \end{aligned} \tag{2.3}$$

The coefficients in the model are approximations of the population values.

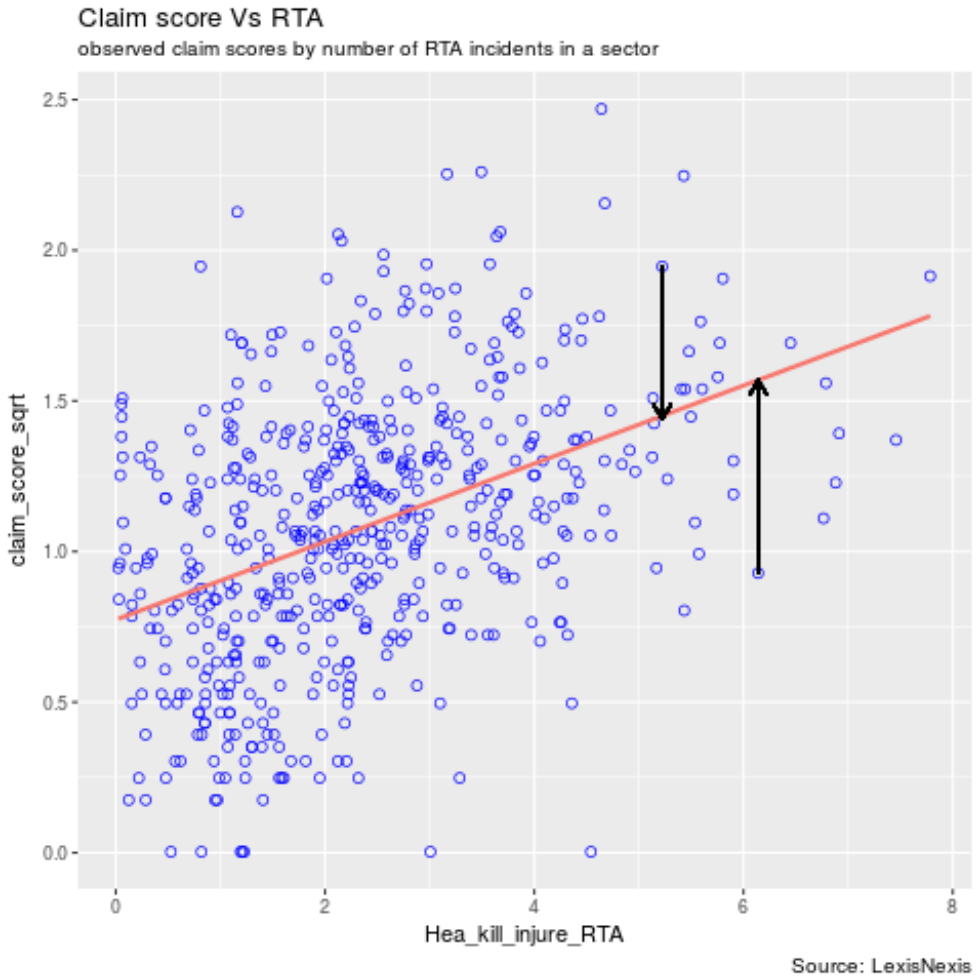

Figure 2.1: **Claim score by RTA incidents.** Linear regression line fitted to observed values of claim score and RTA incidents in a PC sector.

2.2.1 Residuals

The black lines in figure 2.1 are the vertical distances between the observed value of the response and the fitted line (red). These distances are the residuals and are given by:

$$e_i = y_i - \hat{y}_i \tag{2.4}$$

where $\hat{y}_i$ is the value on the fitted line for the particular value of the predictor (RTA incidents). It gives the values of the response, estimated by equation 2.3 at given levels of the predictor.

In addition, these residuals provide a way of approximating the variance in the observed values of the response:

$$\sigma^2 \approx \frac{1}{n-p} \sum_{i=1}^{n} e_i^2 \tag{2.5}$$

where there are n data points and p is the number of explanatory variables.

2.2.2 Ordinary Least Squares

Ordinary Least Squares (OLS) is the method for fitting the line. The aim is to minimise the sum of the squared residuals (SSe), given by:

$$\begin{aligned} SSe &= \sum_{i=1}^{n} \{y_i - \hat{y}_i\}^2 \\ &= \sum_{i=1}^{n} \{y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)\}^2 \end{aligned} \tag{2.6}$$

Differentiating the above equation wrt $\hat{\beta}_1$ and setting the result to zero gives the value of $\hat{\beta}_1$ that minimises SSe. Using vector notation, where X represents all values x_i, this gives:

$$\hat{\beta}_1 = Cor(Y, X) \frac{Sd(Y)}{Sd(X)}$$

Since the regression line always through the mean values $\bar{X}$ and $\bar{Y}$, the estimate $\hat{\beta}_0$ can be calculated from:

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$$

Note that β_0 has units of Y and β_1 has units of Y/X.

2.2.3 Variance Explained

The R-squared value (Rsq) is an important metric since it quantifies the amount of variance explained by the model. It is the variance in the response that is due to the relationship with the predictor. The total variance in the response is:

$$SSyy = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

And the variance due to error is σ^2 given by equations 2.2. Note that the predictors are the deterministic part of the model. Error is due to uncertainty about the observed value of the response. Equation 2.5 gives the estimate for the error as SSe. Therefore, the proportion of variance due to error is:

$$\frac{SSe}{SSyy}$$

Consequently, the variance due to the model, that is, due to the relationship of the response with the predictor, is:

$$Rsq = 1 - \frac{SSe}{SSyy}$$

2.3 Multiple Linear Regression

In this case, there is more that one explanatory variable as formulated in equation 2.7. Here, the Least Squares method fits the observed data points to a multi-dimensional plane rather than to a line. Figure 2.2 is an example with two explanatory variables, X_1 and X_2. The residuals are the vertical distances (wrt Y-axis) from the observed response to the plane delineated by the two explanatory variables. The values of the response estimated from equation 2.7 lie on this plane. As with simple linear regression, OLS produces estimates for the coefficients that minimise the sum of squared residuals.

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i,1} + \hat{\beta}_2 x_{i,2} + \hat{\beta}_3 x_{i,3} \tag{2.7}$$

2.3.1 Multiple Regression Interpretation

Each coefficient in the multiple regression equation 2.7 is the impact of that predictor on the response, while holding the other predictors constant. Thus, it is the unique impact of that predictor. That is, it is the expected change in the response per unit change in the predictor, while holding all other predictors constant.

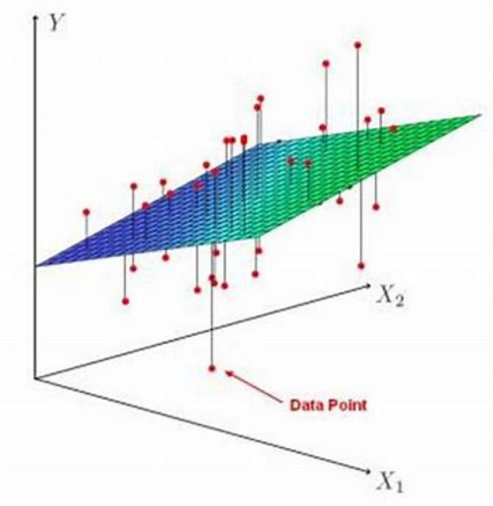

Figure 2.2: **Multiple Linear Regression**. Multiple Linear regression plane formed by two explanatory variables. OLS fits this plane to observed values of response and explanatory variables by minimising the sum of squares of residuals.

Where there are multiple predictors, it is necessary to check for correlation between the predictors themselves in order to avoid confounding. Adding another predictor to a linear model always increases the R-squared value. The adjusted R-squared value takes account of this.

Chapter 3

Data Analysis Cycle

Data analysis is a cyclical process proceeding from exploratory investivation and making improvements as greater understanding of the data arises (Zheng, 2015). Reflecting upon the primary research question leads to an initial set of assumptions about the sourcing of data and the required analysis as described in figure 3.1. However, when the results of analysis are inconsistent, unrea-

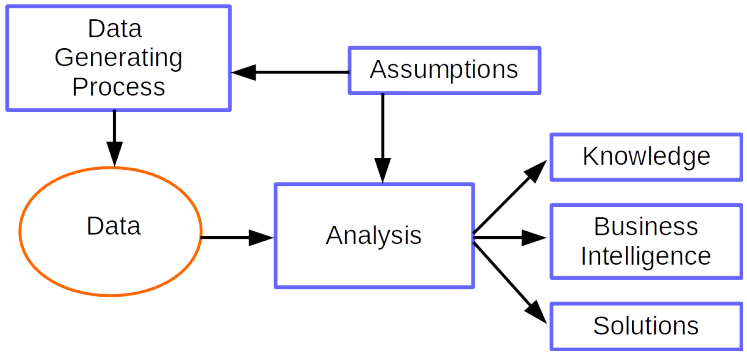

Figure 3.1: **Data Analysis Cycle.** Stages in the cyclical process of data analysis.

sonable or incorrect, it is necessary to revise assumptions. Cyclic adjustments produce better sourcing and preparation of data along with more appropriate statistical tests and modelling methods.

Primary Research Question
Is it possible to predict insurance claim score from a broad range of crime and demographic features

3.1 Assumptions on Data

3.1.1 Claims

Insurers have their own formulas for calculating claim scores. However, in general, a score of 770 or higher is considered good (unlikely to claim). Conversely, a score of 500 or lower is poor (Insurancescored.com, 2016). The dataset provided specifies claim score standardised to the Z distribution. This enumerates each score in relation to the mean score. So potential customers from a particular sector, with a claim score of 2, for instance, are two standard deviations above mean. The probability of getting a claim score of at least 2 is 0.977.

3.1.2 Crime

The sponsor, LexixNexis, has stated that the crime count values need to be converted to population ratios. Since demographic records are at the per 1000 persons ratio, this is the transformation to apply to the crime counts.

This data does not have the attributes of tidy data (Wickham, 2014). It exhibits one of the problems detailed by Hadley Wickham:
Column headers are values, not variable names.
This because column headers are categories of crime.

3.1.3 PopProfile

Demographic data is at the ward interval level. This presents a significant data preparation challenge. It is necessary to transform from ward to PC district to PC sector in order to bring all data to the same observational level. It is preferable to map data to the lowest observational level since finer granularity in data leads to greater statistical power (Shintani, 2014).

In addition, another of Wickham's tidy data problems occurs here:
Multiple types of observational units are stored in the same table.
Each observation specifies a ward and a PC district. The two are incongruent and refer to different area ordinances. In fact, observations duplicate the ward level data for each PC district within its bounds.

Four of the features in the PopProfile data are count values. These too need transforming to per 1000 persons ratio.

3.2 Assumptions on Analysis

Company law constrains insurance companies to certain types of analysis, such as Linear Regression. Not having a decision boundary, as with logistic regression, the analyst must denote meaning to the estimates and numerical outcomes of Linear Regression.

3.2.1 Assessing Quality of a Model

- **Analysis Objective**
  Linear regression is the standard method for predicting a numerical outcome and also for understanding the relationship between predictors and response. When the priority is prediction accuracy, the criterion for model selection is the R-squared value When an interpretable model is important, the metric for identifying significant predictors is the p-value.
- **Parsimony**
  Parsimony is a recommended, if not necessary, property of a linear model for it to be useful. However, arriving at such a model may be difficult due to the large number of features and the many crime categories. Besides from the complexity of the task itself, a large number of variables is problematic for linear regression (Zumel & Mount, 2014).
- **Application**
  The quote "All models are wrong, but some are useful" (Box, 1979) reminds that a model needs to be applicable in a broad range of use cases.

The sponsor has recommended using composite features. Again, the feasibility of this approach needs investigation since composite features may lead to overfitting. In addition, there is the loss of meaning.

Because this is an observational study, regression analysis needs to deal with potential confounders (Shintani, 2014).

Chapter 4

Data Preparation

4.1 Data Sources

This is an observational study. The sponsor, LexisNexis, has provided three datasets:

1. **Claims scores.**
   In general, an insurer calculates an insurance claims score for individual persons. Formulation of accident and insurance claim history along with credit reports determine this score (Progressive, 2018). The claim score number represents the likelihood of making an insurance claim. The lower the score, the less likely a person is to make an insurance claim.

   For the analysis here, the sub-population of each postcode (PC) sector in the Greater London Area (GLA) has an insurance claims score. This is the average score for individuals in that sector. So the observational units are PC sectors rather that individuals. The abreviation for this dataset in this report is "Claims". Note that a PC sector consists of several postcodes in a geographic area. A distinct postcode refers to a street or road.

2. **Crime Data.**
   The crimes dataset comes from the UK Crime Statistics website (UK-Crime-Statistics, 2018). They collate raw crime data from the UK Police which is recorded at the individual postcode level. The data supplied by the sponsor has been mapped to postcode sector level. Unfortunately, the time period, at which the data was recorded, is not specified. The dataset tabulates types of crimes as counts of incidents for GLA PC sectors. The dataset abreviation is "Crime".

So the interval level, at which observations are made, is the same for both Claims and Crime.

3. **Population profiles.**
   The source of the population profiles is the London Datastore (Greater_London_Authority, 2015). This data has already been modified to tabular form and records demographic features with ward as observational unit. Each ward has a range of features pertaining to PopProfile services, such as health, housing and education. The type of numeric value recorded differs across these features. The value given can be counts of persons, average counts or per 1000 persons ratios. Names of features in this dataset identify the data collection period as 2015. The abreviation used here for this data is "PopProfile".

4.2 Data Wrangling - Transformation

Data Wrangling refers the the transformation and mapping of "raw" data to a form suitable for analysis (Tomar, 2017). As for the LexisNexis datasets, this stage requires significant effort in order to combine the three datasets such that they are all at the same interval level.

4.2.1 Response Variable Transformation

The variable of interest is claim-score in the claims data. The values given have been scaled to z-scores. Traditionally, insurers use both credit report information and accident and claims history (Progressive, 2018). They aim to predict the likelihood that an individual policy holder will make a claim. Prediction analsis assigns a value to each credit and accident feature. This assignment uses information from policy holders with similar credit and accident characteristics. It then sums these values to give a claim score.

Data from the sponsor differs from the traditional approach:

- It excludes credit history.
- It considers prior claims, types of crimes and demographic features.
- Rather than characterising individual policy holders, it aggregates features for all persons within a postcode sector.
- The claim score value assigned to a postcode sector has been scaled to the Z distribution.

The goal is to predict a potential customers claim score from the crime and demographic PopProfile of the area where they live. This is a reasonable assumption given that areas of PopProfile deprivation, for instance, tend to have higher crime rates and, consequently, more insurance claims.

A plot of distribution of the response variable (claim-score) displayed right skew 4.1. Linear regression provides better results when response has a normal distribution (Wade & Koutoumanou, 2017). The Box-Cox power transform (Kabacoff, 2015) recommended an exponent value of 0.42. This is close to 0.5 so the square root transformation was applied to the claim-score variable. This necessitates shifting values by a constant to avoid a square root of zero error. Figure 4.1 shows the effect of the square root transform on the response variable. The square root transform has the advantage of being easy to perform the inverse transform to obtain claim-score as a Z value.

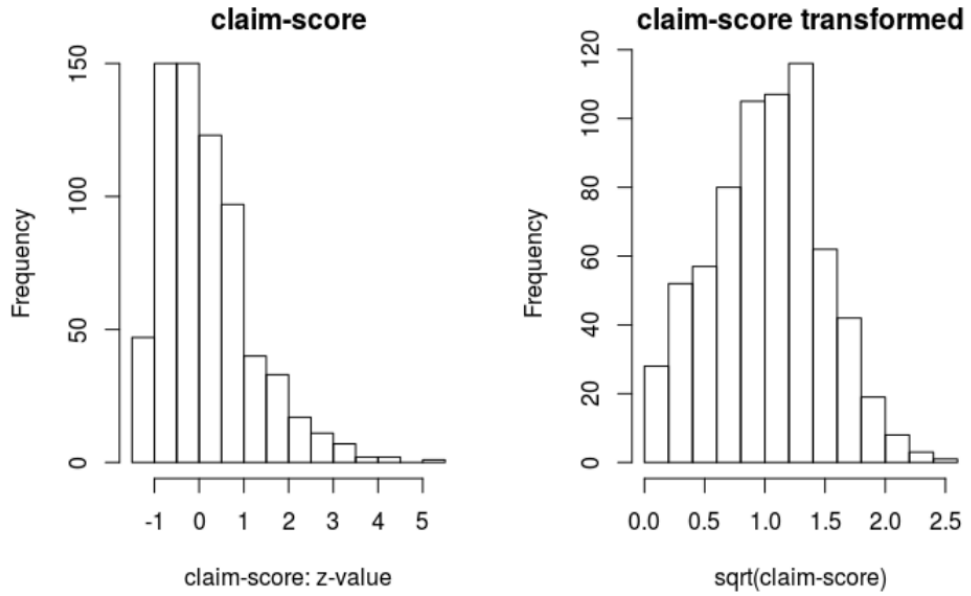

Figure 4.1: **Square root transform on response.** Effect of adding a constant and transforming claim-score by the square root.

4.2.2 Transforming Features

- **Complete Cases**
  Having performed conformance across the three datasets as detailed in appendix A, PopProfile still contained 151 observations with NA entries across all features. These were removed. Of the remaining, only two features contained NA values. Nearest neighbour imputation (k = 5) ensured that PopProfile contained only complete cases (Hastie, Tibshirani, Narasimhan, & Chu, 2017).

- **Unnecessary Features**
  In Crime, features pertaining to postcode statistics, such as mean lifetime of a postcode, were removed.
  Analysis at PC sector level uses the population values in Crime. Thus, features pertaining to population count in PopProfiles were removed since they are at the ward level.

- **Crime Ratios**
  Crime features record are counts of incidents of types of crimes. In order to compare PC sectors with different popupation numbers, the crime counts need transforming to count ratio per 1000 persons.

4.3 Data Wrangling - Mapping

4.3.1 Mapping of Features

Assumptions and Justification

The available data is not complete. For instance, Claims and Crime data list only four postcode sectors for district N14, when in fact, there are five (?, ?). Therefore, it is reasonable to use approximate evaluations on such data. So the mapping of population PopProfile measures from ward to PC district level is based on the number of districts within a ward. More precise techniques require more accurate data.

4.3.2 Matching Claims, Crime and PopProfile Datasets

It was necessary to remove observations from Crime and PopProfile so that each entry has a corresponding record in Claims, which contains the response variable. Appendix A gives details of this selection process.

4.3.3 Mapping from Ward to PC Sector

Observations in PopProfile are at the Ward level while Claims and Crime record at the PC sector level. So it is necessary to map PopProfile wards to PC sectors as shown in figure 4.3.

This is a two-stage process:

1. Map from ward to PC district.

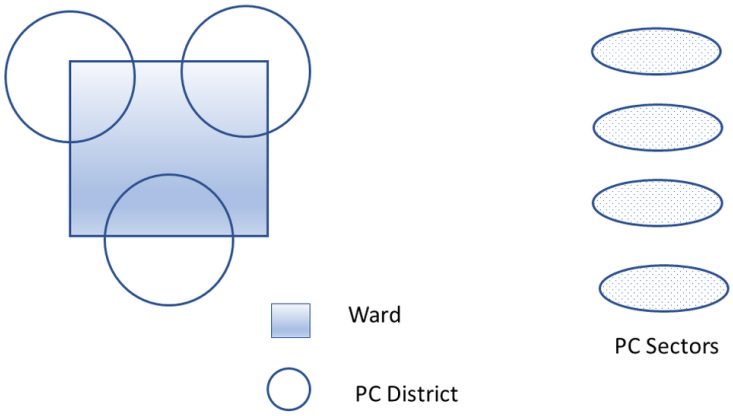

Figure 4.2: **Mapping from Wards to PC Sectors.** Each ward contains a number of PC districts. These need to map to PC sectors for compatibility with Claims and Crime

2. Aggregate to PC district.

3. Apportion PC district values to constituent PC sectors.

The split-apply-combine can implement these operations (Rocks, 2011).

Ward to PC District

Figure 4.3 gives the example of ward 043. This ward has within its boundaries three PC districts. Since all recordings in PopProfile are at the ward level, this dataset duplicates these values for each of the constituent PC districts. In this respect, the data is misleading. It proports to give feature values specific to district N11 within ward 043, for instance. Instead, the value for the whole of ward 043 is duplicated in districts N11, N20 and N14. This is acceptable when the value recorded is a ratio (percent, average, per 1000 persons). However, certain features contain count numbers and attributing the total count for the ward to each of the constituent PC districts is not correct.

The mapping employed here distributes count values at ward level equaly among PC districts as in figure 4.4. Where the number is a ratio, the duplicated value is appropriate.

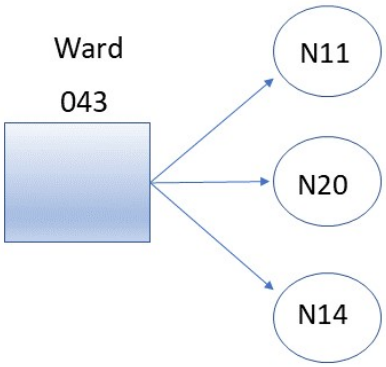

Figure 4.3: **Districts in ward 043.** Each district has a duplicate of the value for ward 043

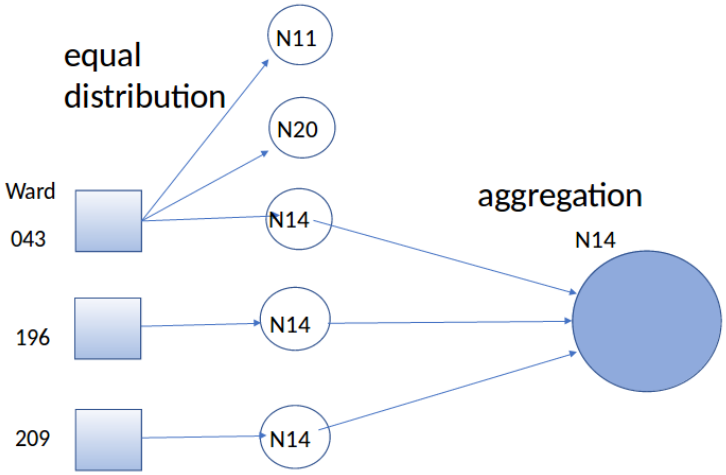

Figure 4.4: **Distributing counts from ward to PC district.** Equal distribution among constituent districts followed by aggregation at district level.

PC District aggregation

Figure 4.4 also shows that PC district N14, for instance, has entries for multiple wards (043, 196 and 209). To arrive at an estimation of the total value for N14, it is necessary to aggregate the contributions from each ward/PC district listing. Count and ratio values are treated differently. Count values are summed while the average is used to obtain the aggregated value for ratios. The latter involves taking the mean of the mean in some instances.

The R code below is an implementation of the split-apply-combine algorithm using data tables.

```
profiles_agg <- profiles[ , map(.SD, sum) , .SDcols = (cols),
                                    by = postcode_district]
```

This demonstrates the efficacy of data tables. `.SDcols` specifies features with count values and the `by` operator groups all entries for each district.

Apportion PC District to Sector

The final stage in mapping of feature values is from PC district to constituent PC sectors. Count values are apportioned by the population ratio for each sector to the total for the district. Mean, rate and per 1000 persons numbers are duplicated across the sectors, as in figure 4.5.

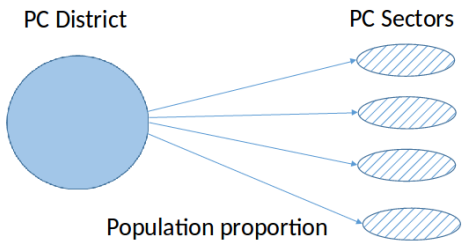

Figure 4.5: **Distributing counts from ward to PC district.** Equal distribution among constituent districts followed by aggregation at district level.

4.4 Discrepancies in Crime Data

- **Total Population**
  The sum of population values across all PC sectors in Crimes is 9,680

Mn. This does not agree with the 2011 census (8,174 Mn) nor the 2016 projection (8,633 Mn) (Demography@london.gov.uk, 2018).

- **Compare to top 100**
  For those PC sectors listed in the top 100 UK sectors by crime, the population values in Crime were exactly twice that given in the top 100 (Economic Policy Centre, 2015).

- **Zero Values**
  Sixteen PC sectors had a zero value for population and were removed. These were all in the central area.

- **Sector out of use**
  One PC sector, E6 7, is no longer in use and was removed from crime.

Central Area Bias

The central PC sectors of London have low population. All are less than 700 persons and some are as low as 10. However, these PC sector still experience crimes due the transient groups of tourists and commuters. Therefore, crime ratios computed for these sectors would give inflated figures. As mentioned above, the UK Crime Statistics website curate this data and source it from the UK Police.

There is an alternate source of crime data for London available from the Economic Policy Centre (Economic Policy Centre, 2015). This data excludes the central PC sectors, which cover 87 PC sectors.

Table 4.1 describes a single Crime feature (Anti-PopProfile Behaviour) and compares the spread of values for all PC sectors with outer sectors. The first row shows the inflated values in the fourth quartile. When the central sectors are removed, there is an order of magnitude reduction in extreme values. However, outliers are still an issue.

Table 4.1: **Effect of excluding central PC sectors** Quartiles for Anti-PopProfile Behaviour variable with and without central sectors. Note that values listed are per 1000 population ratios.

	Min	Q1	Mean	Q3	Max
all_sectors	0.000	2.418	3.694	7.095	4571.429
outer_sectors	0.000	2.440	3.515	5.825	129.139

Having removed the central PC sectors, the number of data points reduces from 680 to 591. Matching all three datasets in order to merge them results in a further reduction to 553 entries. Reducing sample size decreases the

statistical power to detect true difference (Shintani, 2014). However, this is an acceptable loss in order to have confidence that the sample contains valid observations.

The abbreviation for the combined dataset if ClaimsFull.

Chapter 5

Exploratory Data Analysis

5.1 Bivariate Relationships

Given such a large number of features, correlation analysis across all variables becomes unweildy. Nevertheless, it is worth inspecting a subset of variables. Figure 5.1 displays correlation, distribution and scatterplots of crime features involving theft. Immediately apparent is the problem of outliers. Outliers that were three standard deviations above mean were identified. However, there were no outliners in common for these six features. Nevertheless, it is still possible to deal with outliers by checking for high positive and negative residuals subsequent to linear regression. Also, inspecting leverage needs to part of regression diagnostics.

The correlation values in figure 5.1 are for Pearsons coefficient. However, due to the presence of outliers, Spearmans correlation is recommended. As a result, figure 5.2 shows that the strong correlation between the response variable and robbery disappears. There is still strong positive correlation between `other-theft` and `theft-from-the.person`. Therefore, `other-theft` was chosen as representative of both. Selecting representative features has the added advantage of reducing potential confounding, as highlighted in section 3.2.

Similar analysis was carried for other subsets of features with common characteristics and appendix B.1 contains the results.

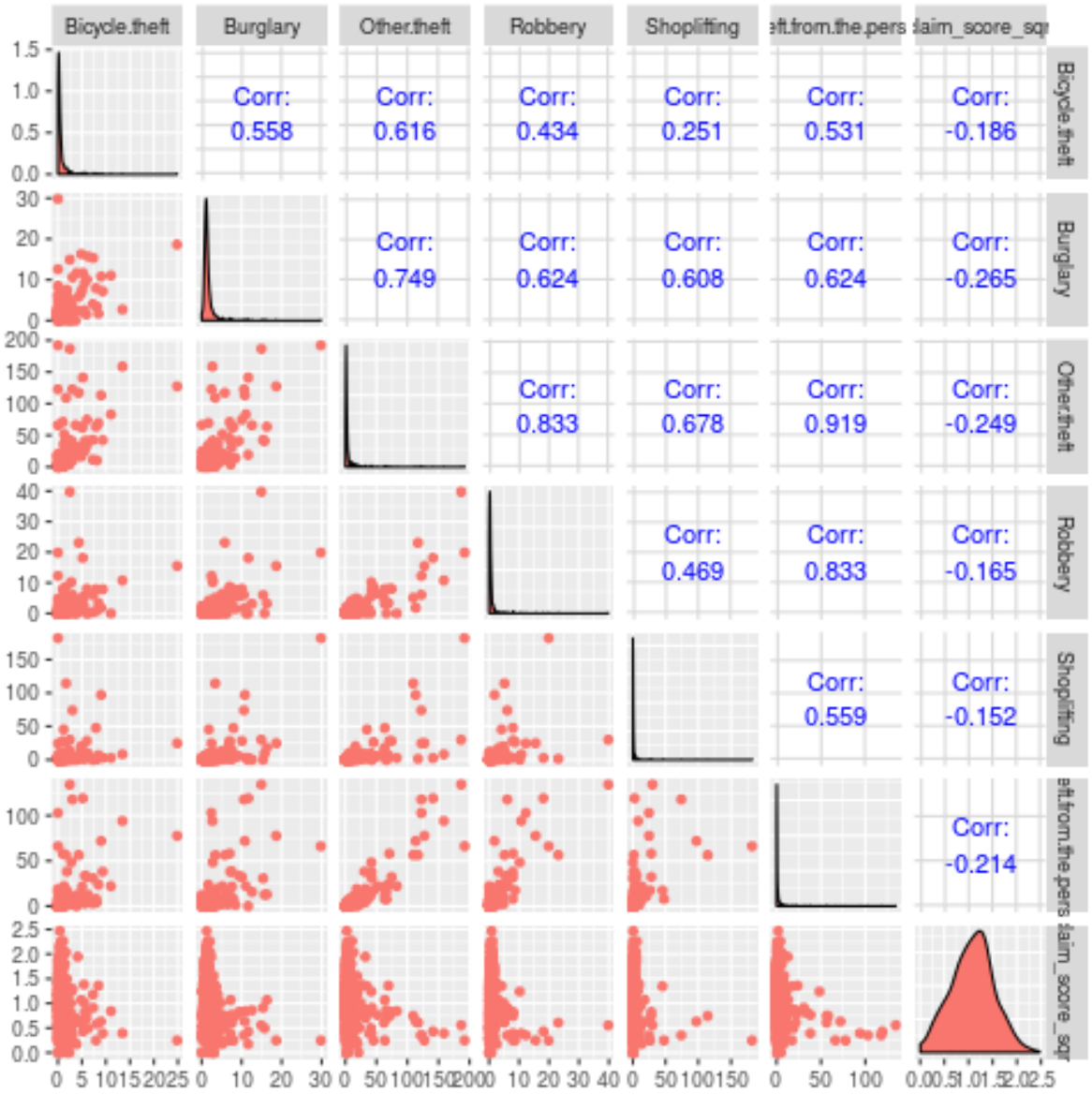

Figure 5.1: **Correlation matrix plot of theft variables.** Correlation value, univariate distribution and scatterplots of crime features pertaining to theft.

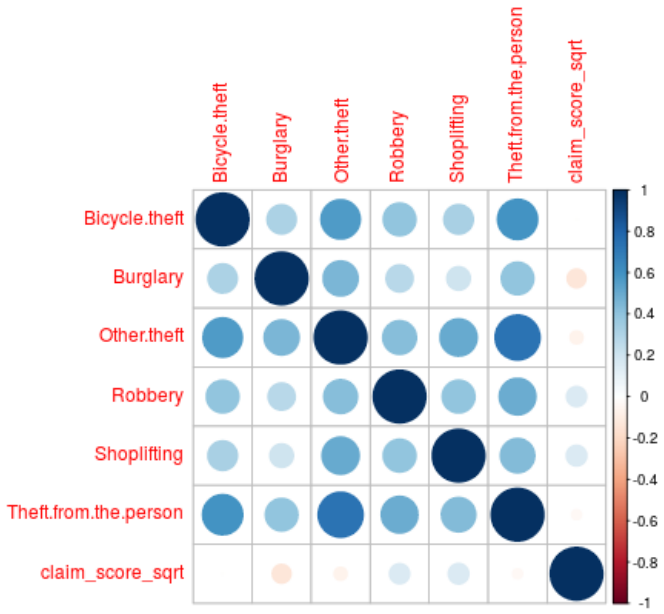

Figure 5.2: **Correlation matrix using Spearmans coefficient.** Spearmans correlation giving bivariate relationships among crime features pertaining to theft.

5.2 Feature Selection

A dataset with 53 variables is relatively large and calls for some type of dimension reduction. One form is principal components analysis. This is an unsupervised method and does not take into account the relationship between the features and the target or response variable. Also, it transforms features into high-variance components with the resultant loss of meaning. The analysis here considers another form of dimension reduction, feature selection, which looks at the impact individual features in relation to a response variable.

Feature selection helps to identify a subset of potential predictors for a regression model. Traditional methods of feature selection are stepwise selection and best subsets selection. The former adds or removes variables in turn and checks the effect on the overall model. The latter determines the 'best' subset of a specific number of predictors for a given criterion, such as the R squared value.

Another kind of feature selection is the 'all-variables' approach of the Boruta algorithm (Kursa & Rudnicki, 2010). Application areas include high-dimension genomics data. It employs Random Forest classification. As such, it adds randomness to a system and then analyses the ensemble of features with randomisation. It attempts to reduce the effect of random fluctuations and correlations and to identify those features which are considered important.

The steps in the algorithm are:

1. For all the features, it creates a corresponding set of shadow features.
2. It randomises the shadow features by shuffling them.
3. It calculates Z-scores for the ensemble system and then calculates the mean Z-score among the shadow attributes (MZSA).
4. By comparing Z-scores for the original features with the MZSA, it determines importance measures. So the importance metric is the Mean Decrease Accuracy (Dutta, 2016).
5. It rejects attributes that have a low importance metric and retains those with high values.

5.3 Exploratory Analysis with Boruta

The first pass of Boruta identified 40 variables that were confirmed important. Figure 5.3 classifies features according to importance. The box-plots

show maximum, average and minimum Z-score. There is one box-plot in between yellow and green. This blue box-plot represents the Z-scores for the shadow features. Thus, all features above this baseline, randomised stage are considered tentative or important. Tentative features are those that the algorithm is unable to affirm as significantly greater than the baseline for the default confidence level.

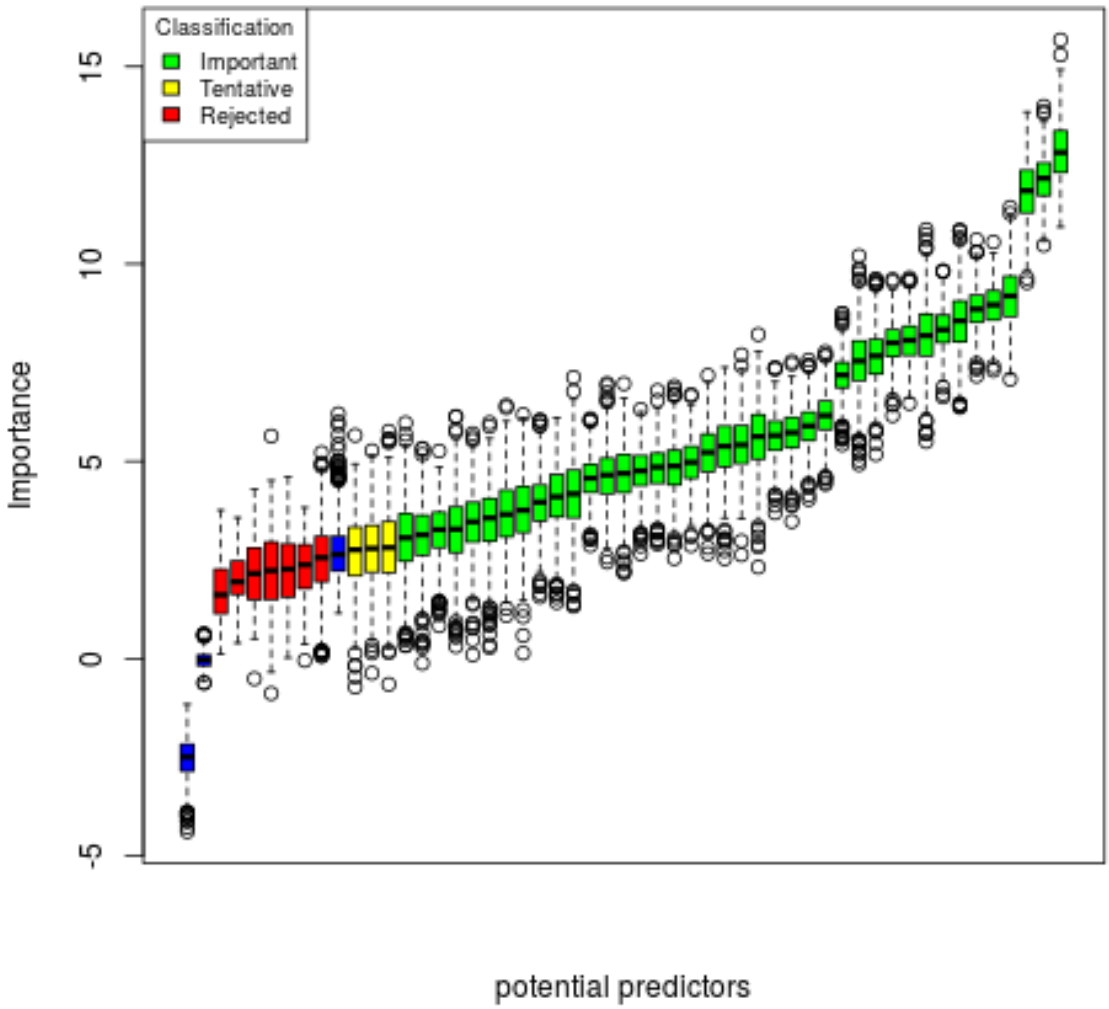

Figure 5.3: **Feature selection with Boruta.** The Boruta package in R classifies importance of variables in relation to shadow, randomised variables.

Due to the large number of features, it is not possible to display them on the x-axis. Table 5.1 lists the top ten variables by importance. Both this table

Table 5.1: **Feature importance metrics** The top ten features and mean importance calculated from Mean Decrease Accuracy.

	meanImp	medianImp
Hea_kill_injure_RTA	12.854	12.822
Dem_children_pc	12.147	12.171
Violence.and.sexual.offences	11.819	11.856
Robbery	9.186	9.188
Dem_working_pc	8.953	8.961
L_num_jobs_area	8.857	8.853
Other.theft	8.561	8.556
H_flat_mais_apt_pc	8.355	8.329
Shoplifting	8.189	8.185
T_av_public_acc_score	8.063	8.064

and figure 5.3 show that the first three variables are higher that the main body of features. However, Violence.and.sexual.offences and Dem_children_pc have strong positive skew. This means that most PC sectors are at the lower value range but a few sectors have very high numbers.

Fitting a linear model with these three features results in an adjusted R-squared value of 22.9%. Also, the P-value for the violence feature is not significant. Rather than attempting corrective measures, re-evaluation may be more appropriate.

```
lm(claim_score_sqrt ~ Hea_kill_injure_RTA + Dem_children_pc +
                            Violence.and.sexual.offences,
                      Data = claims)
```

Chapter 6

Second Pass Analysis

Issues with discrepancies in crime features, detailed in section 4.4 along with the strong positive skew, shown in figure 5.1, call for a reappraisal of this dataset. Such a re-evaluation is part of the cyclic nature of Data Analysis, described in section 3.

Section 4.4 referred to an alternate source of crime data (Economic Policy Centre, 2015). This is contains all the features in the LexisNexis crime data and is available at the PC sector level. For comparison purposes, figure 6.1 displays the same features involving theft as the LexisNexis data given in figure 5.1. There is an improvement in the distribution of some of the features even thugh, for others, outliers are still a problem. The broader spread of values has the potential for improving linearity with the response variable. Most striking, however, is the reduction in multicollinearity.

The EPC crime data was merged with Claims and PopProfile and analysis was performed on the new combined dataset.

6.1 Feature Selection

Feature selection with Boruta on the combined data classified 36 features as important. At least this is a reduction compared the LexisNexis data, if only by 4 features. Figure 6.2 displays the classification for each feature. It is noticeable that one feature (RTA incidents) lies above all others

Table 6.1 lists the values for the mean importance measures. Interestingly, the same three features as in the exploratory analysis in section 5.3 are greater in importance than the main body of variables.

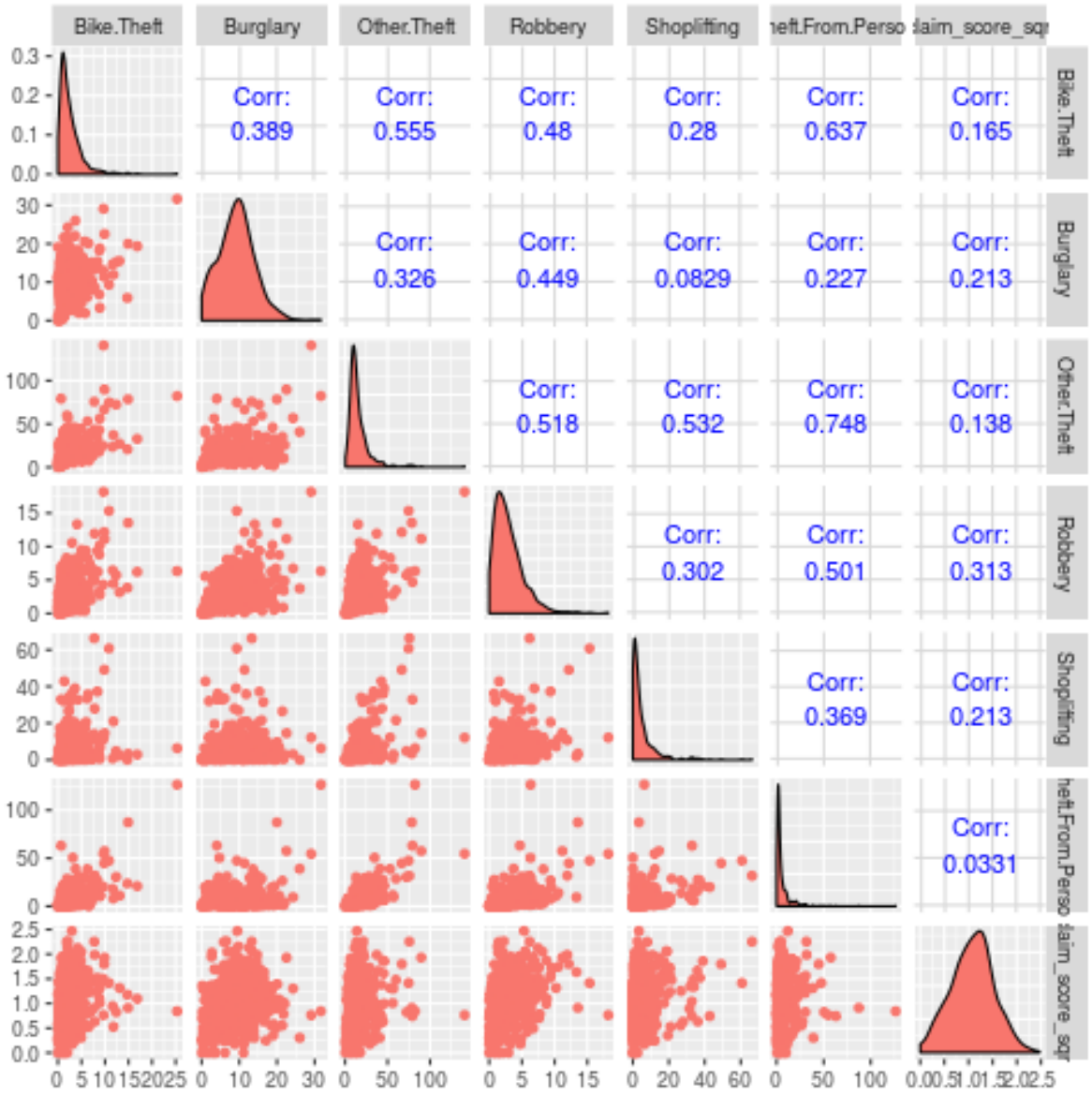

Figure 6.1: **Crime features dealing with theft.** Scatterplots, univariate distributions and correlations for features involving theft in the EPC data.

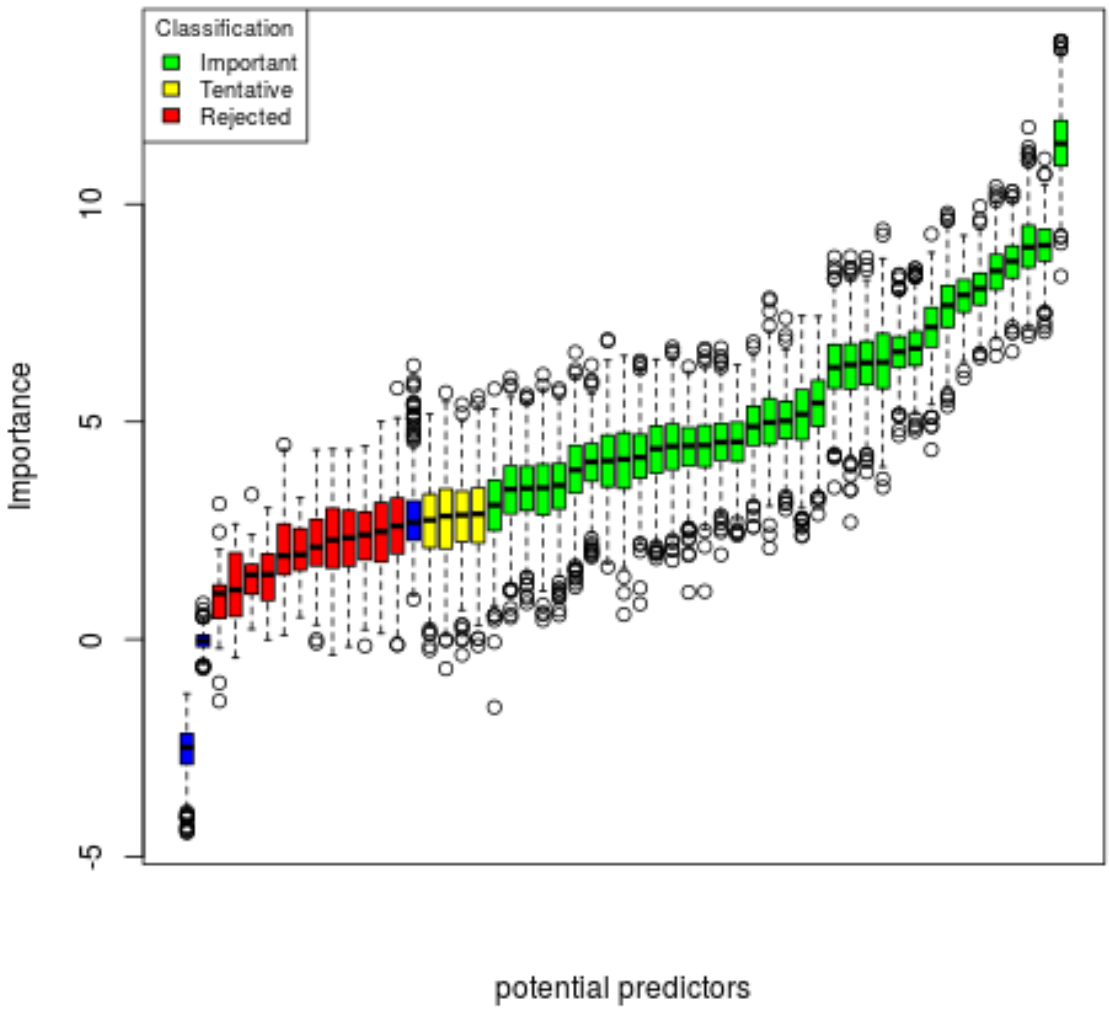

Figure 6.2: **Variable selection with Boruta.** Classification of features according to Boruta importance measures. Box-plots show the spread of importance values.

Table 6.1: **Boruta importance measures** Mean importance measures for the top 10 features identified by the Boruta algorithm.

	meanImp
Hea_kill_injure_RTA	11.399
Dem_children_pc	9.049
Violent.Crimes	9.027
Dem_working_pc	8.676
H_flat_mais_apt_pc	8.455
H_terr_houses_pc	8.053
T_av_public_acc_score	7.887
Total.Crime...ASB	7.638
Dem_age_av	7.155
H_house_spaces	6.676

6.2 Multiple Linear Regression

It is worthwhile inspecting the three features with highest importance measures. Figure 6.3 shows that each feature has a near normal distribution. In addition, there is a linear relationship with the response, as displayed in the lower three boxes. This suggests that these three would have significance as predictors in a linear model.

A multiple linear model was fit using the same formula as for the `lm` command in section 5.3 using the EPC data. This produces an adjusted R-squared value of 35.26%. Adding the next predictor in level of importance (Dem_working_pc) resulted in a negligible increase in the adjusted R-squared.

A full multiple linear model was fit with all 36 predictors recommended by Boruta. One would expect such a model to explain most of the variance in the response. However, the adjusted R-squared value was 43.92%.

This exercise was worthwhile in that the full model identified three features as significant. As noted by Wade (Wade & Koutoumanou, 2017), some predictors are significant only after adjusting for the other predictors. Two of these are the same as the three variable model (RTA incidents and Violent.Crimes). In place of Dem_children_pc, the Shoplifting variable is the one with significance. Unfortunately, this model gave lower adjusted R-squared value of 33.46%.

6.3 Prediction

By reserving test or 'hold-out' data, it is possible to make predictions by applying this data to the fitted model. It is also a means of measuring the

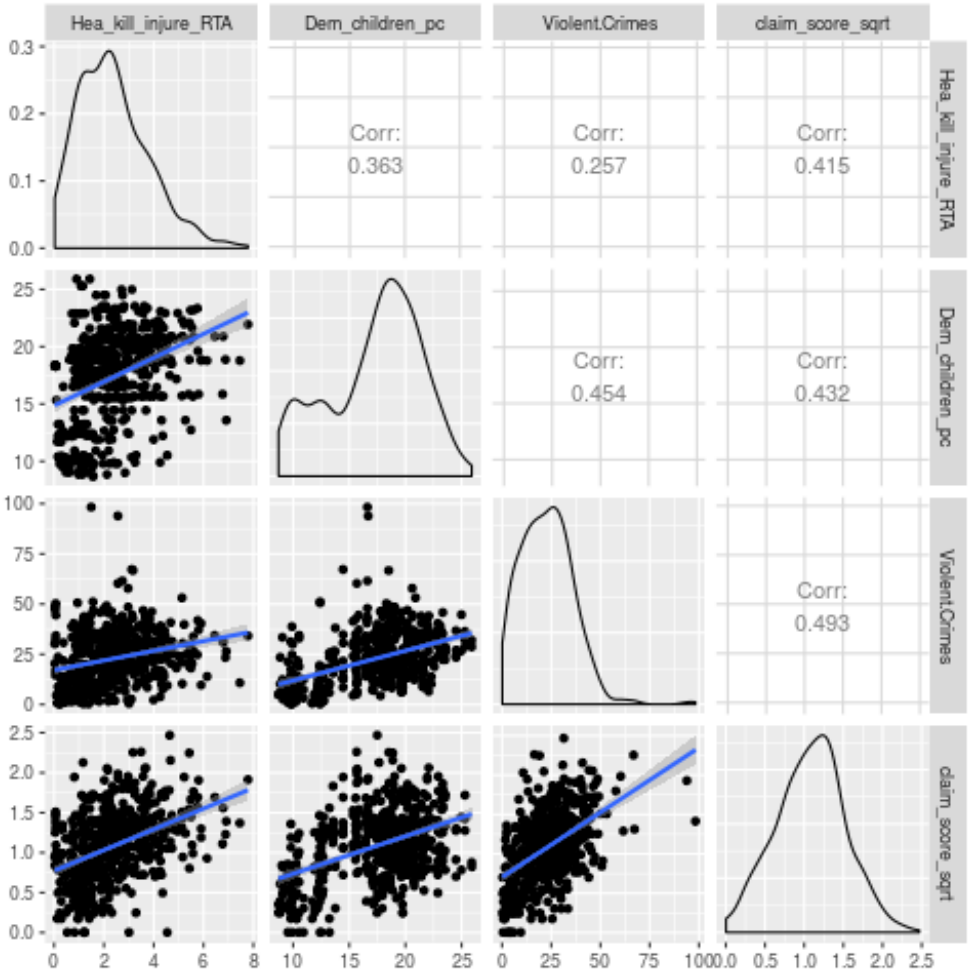

Figure 6.3: **Candidate features.** Analysis of the potential predictors for a linear model.

efficacy of a model. The dataset was split 80 / 20 into train and test data. The train data, containing 440 data points, was used to fit the linear model. The number of observations in the test data was 111.

The `predict` functions calculates the predicted response (`predict_values` for the given set of new predictor values. The amount of error in these predictions is the difference between the actual value for the claim score and the predicted value. The measure for the error is the root mean squared error and is given by:

$$RMSE = \sqrt{\frac{(Y_i - \hat{Y}_i)^2}{n}}$$

where Y_i are the actual values of the response (claim score) and $\hat{Y}_i$ are the predicted values. This is computed as:

```
RMSE <- sqrt(mean((test_data$claim_score_sqrt -
                              predict_values)^2))
```

This results in an RMSE value of 0.373.

Appendix A

Conformity between Claims, Crime and PopProfile

Observations in Crime and PopProfile must conform with Claims since the latter contains the response variable. In effect it means that analysis is possible only on instances of PC sectors that are in all three datasets. To this end, the following PC districts were removed:

- Claims data does not contain PC sectors WC2A1, WC2A2 and WC2A3. These sectors were removed from Crime. Also, district WC2A was removed from PopProfile.
- Claims does not contain 30 PC districts that are listed in PopProfile. So observations pertaining to these districts were removed from ProProfile. This reduces PopProfile from 1093 to 979 observations.
- Claims does not contain thirty-five PC sectors and the following were removed from Crime:
  "EC1M6" "EC1N6" "EC1N7" "EC1V3" "EC2M2" "EC2N2" "EC2Y5" "EC2Y9" "EC3A1" "EC3A3" "EC3A5" "EC3A6" "EC3A7" "EC3M5" "EC3M7" "EC3M8" "EC3N2" "EC3N3" "EC3R7" "EC3R8" "EC4M5" "EC4N5" "EC4R0" "EC4R2" "EC4V2" "EC4Y7" "EC4Y8" "SW138" "W1D4" "W1F0" "W1H1" "W1K3" "W1K5" "W1U7" "WC2E8"
- Claims does contain PC sectors EC1Y2 and EC3A2. However, since they are not in Crime, they were removed from Claims.

Central PC sectors were removed from Crime and subseqently from Claim, as described in section 4.4. This reduces further the number of valid observations to 591.

There were 8 PC districts in the joint claim-crime table that are not in PopProfiles. When these are removed, the number of rows in the combined Claims-Crime-PopProfile is 553.

Appendix B

Dimension Reduction

B.1 Representative Features

Where there were two variables with strong positive correlation (greater than 0.7), one was chosen as representative of both. In addition, the `corrplot` function identified clusters of variables that were correlelated with each other, as detailed in table B.1. For both bivariate and clustered relationships, one feature was chosen as representative of the others. The outcome of this stage is a further reduction to 50 features in addition to the response variable in the combined dataset.

While a composite feature with the average of correlated variables would have higher signal to noise ratio, selecting a representative feature preserves meaning. It is interesting that two are from the demographics data and two are crime variables. This suggests that a linear model using these four as predictors would be have the quality of being generally applicable.

Remove	**Represent with**
Theft.from.the.person	Other.theft
Public.Order	Violence.and.sexual.offences
Dem_age_med Dem_older_pc	Dem_age_avg
Dem_not_eng_lang_pc	Dem-born-uk-pc
Hea_ambul_call_alcohol_ill_rt	Hea_all_ambul_incidents_rt
H_prop_sold	H_house_spaces
H_in_band_CD_pc H_in_band_DEF_pc	H_in_tax_band_AB_pc
H_detach_houses_pc H_semiD_houses_pc	H_house_own_pc
L_employ.head_resid	L_num_jobs_area
B_out_of_work_rt B_employ_support_rt B_JSA_rt	B_housing_rt
E_GCSE_av E_A_score_entry_av	E_A_score_student_av
D_dep_child_HH_outofwork_pc	D_in_worst50_Nat_pc
CS_crime_rt	CS_viol_person_rt

Table B.1: **Representative Features.** Where there were two or more features with strong positive correlation, one was chosen as representative.

References

Box, G. E. P. (1979). Robustness in the strategy of scientific model building. In *Robustness in statistics* (pp. 201–236). Academic Press.

Demography@london.gov.uk. (2018). *Greater London Authority, Population Change 1939-2015.* Retrieved 2018-04-18, from `https://data.london.gov.uk/dataset/population-change-1939-2015`

Dutta, D. (2016). *How to perform feature selection.* Retrieved 2018-05-05, from `https://www.analyticsvidhya.com/blog/2016/03/select-important-variables-boruta-p`

Economic Policy Centre. (2015). *Crime by Postcode Sector.* Retrieved 2018-04-18, from `http://ukcrimestats.com/Postcode_Sectors/`

Greater_London_Authority. (2015). *Ward Profiles and Atlas.* Retrieved 2018-04-26, from `https://data.london.gov.uk/dataset/ward-profiles-and-atlas`

Hastie, T., Tibshirani, R., Narasimhan, B., & Chu, G. (2017). *impute: impute: Imputation for microarray data.*

Insurancescored.com. (2016). *What Is Your Insurance Score.* Retrieved 2018-05-03, from `https://www.insurancescored.com/`

Kabacoff, R. (2015). *R in action : data analysis and graphics with R.*

Kursa, M. B., & Rudnicki, W. R. (2010, sep). Feature Selection with the Boruta Package. *Journal of Statistical Software*, *36*(11), 1–13. Retrieved from `http://www.jstatsoft.org/v36/i11/` doi: 10.18637/jss.v036.i11

Lane, D. M., Scott, D., Hebl, M., Guerra, R., Osherson, D., & Zimmer, H. (2018). *Introduction to Statistics.* Rice University. Retrieved from `http://onlinestatbook.com/Online_Statistics_Education.pdf`

Progressive. (2018). *Insurance Scores: What You Should Know.* Retrieved 2018-04-02, from `https://www.progressive.com/shop/car-insurance-credit-scores/`

Rocks, R. (2011). *A quick primer on split-apply-combine problems.* Retrieved 2018-05-01, from `https://www.r-bloggers.com/a-quick-primer-on-split-apply-combine-problems/`

Shintani, A. (2014). *Comparing 2 Propotions, MED101x, Introduction to Applied Biostatistics.* Retrieved 2018-03-30, from https://courses.edx.org/courses/course-v1:OsakaUx+MED101x+1T2016/courseware/8404

Tomar, S. (2017). *A comprehensive introduction to data wrangling.* Retrieved 2018-04-27, from https://www.springboard.com/blog/data-wrangling/

UK-Crime-Statistics. (2018). *UK Crime Statistics.* Retrieved 2018-04-28, from https://crime-statistics.co.uk/

Wade, A., & Koutoumanou, E. (2017). *Introduction to Regression.* London: Centre for Applied Statistical Courses.

Wickham, H. (2014, sep). Tidy Data. *Journal of Statistical Software*, *59*(10), 1–23. Retrieved from http://www.jstatsoft.org/v59/i10/ doi: 10.18637/jss.v059.i10

Zheng, T. (2015). *Statistical Thinking for Data Science.* Retrieved 2018-03-29, from https://courses.edx.org/courses/course-v1:ColumbiaX+DS101X+1T2016/courseware/5c7

Zumel, N., & Mount, J. (2014). *Practical data science with R.* Shelter Island: Manning. Retrieved from https://www.manning.com/books/practical-data-science-with-r