UK Parliamentary Data Analysis

For DE2 MEng Data Science Module

Maximilian Matthews[a], Chinene Chukwuma[a], Owain Pill[a], William Jiang[a]

[a]*Dyson School of Design Engineering, Imperial College London*

June, 2022

Abstract

The project we embarked on was to determine the political, social, and economic situation of constituencies by look at data that was collected from the constituency dashboard on the Parliament website and Census data. Each group member looked at a different aspect of these questions so that we would get an overall view on what features of a constituency influenced it the most and use this to decide where to live. The team all used at least one of the following methods: Ordinary Linear Regression, Logistic Regression, Decision Trees. After each member had conducted their analysis, the results were that it was possible to get a picture of the factors in question (accuracy of all models was above 0.8). Most factors could be well predicted with fewer than 5 features and the most features in a final model was 10.

Contents

Figure 1: UK Parliament Website

1. Introduction

With the rise of globalization and continuous technology advancement, the level of flexibility and freedom for general public to move from one region or locale to another to seek gainful employment in their field and better standard or quality of living have become increasingly higher.

However, the escalation of geographical mobility has proposed a new challenge to us – which place is better for living and working than the others?

1.1. Project Aim

In this project, our group attempts to create a model to predict the quality of living of each constituency in the UK by analysing its social, political and economic conditions. The results could be used as a reference for one that is seeking to transfer from one constituency to another.

2. Related Work

SAS Paradise Found project uses data science and machine learning to calculate the "best place on earth" to live. In comparison to city rankings that are often determined by editorial choices based on limited statistics and criteria, SAS uses a self-learning algorithm that uses a dataset with 5 million data points from thousands of sources that cover 193 countries and 148,233 cities. The sources are compromised largely of publicly available data, including city studies, social media sites, review sites like TripAdvisor, geodata and reports from statistical services and international agencies.

The first city ranking using AI determined that the inner suburb of West Perth in Australia is the algorithm's choice for best place to live. SAS said it is often too easy for unconscious bias to affect results when selecting the criteria to use when determining which data should be collected and analysed so they processed all the available data and allowed machine learning algorithms to decide which criteria were important.

Using a variable reduction technique, the key criteria was cut down to eight, namely: living expenses; safety and infrastructure; healthcare; restaurants and shopping; the environment; culture; attractiveness to families; and education and employment.

3. Methodology

Given the strain the cost-of-living crisis has put on households, many people currently living in big cities such as London are considering whether to move to another, more affordable, location. We aim to predict key indicators such as the number of businesses and house prices in an area to allow people to determine whether they want to live there.

We chose to use a group of datasets freely available on the UK Parliament website. They offer data on key metrics broken down by UK Parliamentary Constituency. This gives us the geographic variation needed for our investigation. The data is originally sources from either the UK Census or other ONS studies.

A combined dataset was created where data on election results, housing stock, economy, deprivation, social mobility, education, religion, prevalence of health conditions, age distribution and ethnic group breakdown were put alongside the Parliamentary Constituency name. This was ordered from A-Z.

Most of this data was in the form of absolute numbers which given the differences in constituency size, is not a good comparative metric. Thus, additional data was "engineered" including percentage results for political parties and voter turnout; the ratio of median house price to median salary; and number of businesses per capita: preventing dataset imbalance.

It was found that data for some metrics including social mobility index, median house price and school funding was only available for England: not the rest of the UK. Additionally, the major UK political parties do not operate in Northern Ireland. Thus, non-English constituencies were removed from the dataset.

In order to mitigate overfitting issues the dataset was randomly split into training, testing and validation sets. This was done to make our models usable on other future datasets including the 2021 Census data once it releases.

4. Local Politics - Maximilian Matthews

4.1. Aim

Using the UK Parliamentary Constituency dataset, I aim to predict whether a constituency could vote

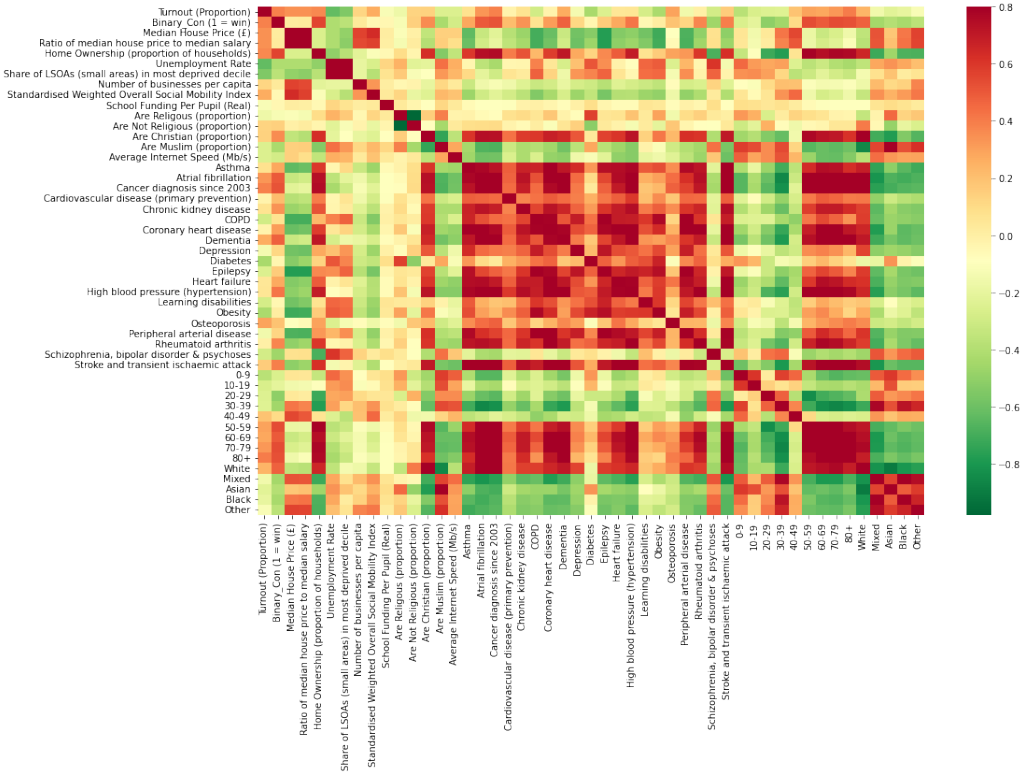


Figure 2: Heatmap showing the correlation of the data-set's features with Binary Con (whether the Conservative Party win the constituency, 1 = win)

for the Conservative Party or another political party.

4.2. Data Preparation

4.2.1. Feature Engineering

Initially the predictor, Conservative vote proportion was in the form of continuous data (0-1). Since this is a classification problem this was converted into a binary data (BIN_CON).

If the Conservative Party on the constituency this was set to 1, else it is equal to 0.

4.2.2. Attribute Removal

In order to reduce the number of useless features the forward selection algorithm has to test, certain features were removed.

- *Constituency Name.* This is not relevant to creating a predictive model, and since it's a series of strings it interrupts with numeric analysis algorithms.
- *Electorate, Number of Businesses.* These two features had already been used to engineer per capita features and are thus no longer useful.
- *Conservative votes, Labour votes, Lib Dem votes and vote proportions.* Voting data is directly affected by the size of our predictor (if Conservatives do well, Labour and Lib Dem automatically do more poorly), thus they are too dependant.

4.2.3. Data Splitting

To prevent overfitting, the dataset was split at random into training, validation and testing subgroups. This was done at a ratio of 60:20:20. Initially an 80:10:10 ratio was tested however this resulted in an excessive gap (0.35) between the training and validation groups).

4.2.4. Data Balancing

The Conservative Party came first in English Constituencies 63% of the time causing an imbalance in the training data which would have caused poor results in the test data.

Thus the majority class (BIN_CON = 1) was undersampled at random and assigned to a new dataset variable. This was only done for the training data since validation and test should be an accurate depiction of reality.

4.3. Dataset Overview

In order to obtain a first understanding of the relationships between BIN_CON and other features, a correlation heatmap was generated (Figure 2. In

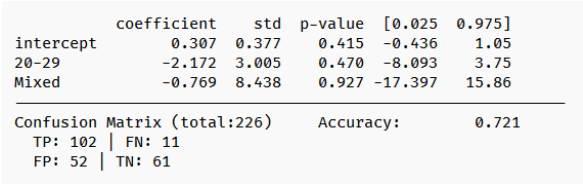


Figure 3: Confusion matrix using features selected by the automatic forwards selection algorithm (training data).

addition to this a correlation metric was obtained for each feature (with BIN_CON).

Table 1: Correlation between features and BIN_CON.

Feature	Correlation
$20-29$	-0.5095
Unemployment Rate	-0.4164
Are Christian (proportion)	0.3617
$50-59$	0.5415
Home Ownership (proportion of households)	0.5625

Correlation numbers in Table 1 shows both negative and positive correlation between features. These confirm expectations that area with older, white, more affluent residents tend to vote Conservative.

4.4. Logistic Regression

From the initial correlation analysis it was obvious that there were moderately strong correlations between multiple features and BIN_CON. Since this is present and this is a categorisation problem, logistic regression was identified as the most suitable model. This is where a sigmoid function is fitted between two binaries.

4.5. Feature Selection

4.5.1. Forward Selection Algorithm

In order to select the best features for the model, without adding too many and risking overfitting, a forward selection algorithm was used.

Initially a model with no features was created. In an iterative process the feature with the next strongest correlation to BIN_CON was identified and added to the model. If the addition of this feature improved the validation accuracy of the model by 0.005 (5%) then it would be included, else it would be rejected. This was repeated until no more features were added, which occurred after 2-3 variables typically.

Compared to selecting the features manually, this produces a more accurate model and the 0.005 inclusion criteria prevents overfitting.

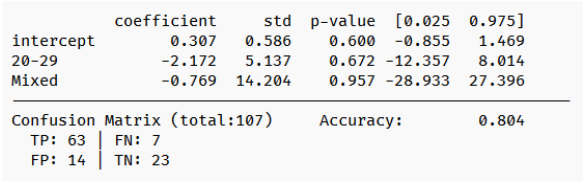


Figure 4: Final model on test data.

After iterating through all columns, the forwards selection algorithm selected features listed in Figure 3. Despite only two features being used, adding additional ones would have improved accuracy by ¡2% risking overfitting. Although the p-values are fairly high the features selected scored well in the prior correlation test: -0.5095 and -0.4458 respectively.

4.6. Testing

A final logistic regression model was created using the selected features and this was trialed using the test data. See Figure 4 and Table 2.

Table 2: Performance metrics from running test data on model

Performance Matrix	Value
Accuracy	0.804
Precision	0.818
Recall	0.9
Δ Accuracy (Val - Test)	0.065

4.7. Discussion

4.7.1. Successes

The performance metrics obtained via the testing dataset show that, with both a high precision and accuracy, my model has high predictive power in determining whether an area would vote for the Conservative party.

Additionally through data engineering, splitting, attribute removal and balancing, as well as a thorough feature selection process, the model does not suffer from overfitting. By obtaining a high accuracy with only two features, a Δ Accuracy of only 0.065 was obtained.

Both the age distribution as well as ethic group distribution, which the model is based on, are provided regularly via the census meaning this model is repeatable in the future.

4.7.2. Limitations

Since this model was only trained on English Constituencies geography is a key limitation. Due to differences in which political parties operate there and the political circumstances, this model could

not be extrapolated onto Scotland, Wales or Northern Ireland.

Furthermore, there have been multiple political realignments in the last decade. This includes more affluent, older traditionally Conservative areas becoming more open to voting for other parties. Thus a model based on demographics could lose its predict power very quickly.

5. Local Unemployment - William Jiang

5.1. Aim

Predicting whether a constituency has a low unemployment rate or not.

5.2. Initial Dataset Overview

The predictive model was built upon the dataset extracted from 2021 census dataset. The dataset consists of 533 constituencies with 55 features (1 ID, 1 categorical, 53 continuous).

5.3. Predictor Variable

The unemployment rate, a widely recognized key indicator of the economic performance of the labor market of a specific region, was chosen to be the predictor variable. The objective of the predictive model was to predict whether a given constituency would have a low unemployment rate at the present time. A low unemployment rate would be considered desirable as a result of more job opportunities available and higher economic output of the constituency. The unemployment rate provided within the database was measured on a scale of 0-100 (0: total employment; 100: total unemployment). See Figure 5.

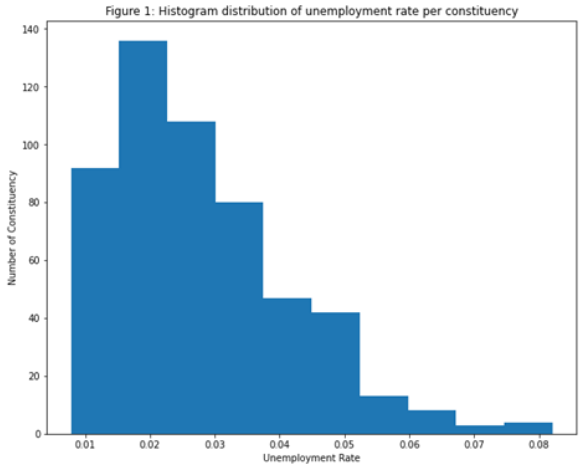


Figure 5: Histogram distribution of unemployment rate per constituency.

5.4. Dataset Preparation

5.4.1. Feature Engineering

As the predictor variable was in the form of continuous data ranging from 0-100, a threshold value was necessary to binarize the predictor attribute.

The mean unemployment rate was calculated to equal 0.028256. Constituencies with mean unemployment rate or below ($<= 0.0028256$) were assigned a value of 1 and those above (> 0.0028256) were assigned a value of 0. A new column. "UR_B", was used to record the binary values. By using mean as a threshold value, imbalance of dataset with respect to y classes could be avoided.

5.4.2. Attribute Removal

The following attributes were removed for the models:

- *Constituency Name.* Served as id and was a non-predictive attribute.
- *Conservative votes, Labour votes, Lib Dem votes.* Proportion of voters for each party in the dataset was considered to be a better representation of political preference of a constituency.
- *Religion.* This attribute was not considered to have causal relationship with the predictor variable.
- *Unemployment rate.* This column was no longer useful after binarization, and was replaced by "UR_B".

5.4.3. Final Pre-processing

The updated dataset was split into training, validation and test (60%, 20%, 20%). The purpose of training set was to train the model, with the validation set for accuracy, recall, and precision measurement. Finally, the test set would be used as a final measure to check the accuracy was obtained to prevent overfitting of data.

All dataset was standardized and rescaled to have the means of 0 and the standard deviation of 1 for requirement of model construction.

5.4.4. Context for Performance Metrics

Three metrics were used to compare the performance of the predictive models:

- *Accuracy.* The proportion of constituencies that have been correctly predicted with high or low unemployment rate.
- *Precision.* The proportion of constituencies that have been predicted with low unemployment rate were correct.

- *Recall*. The proportion of actual constituencies with low unemployment rate were correctly predicted by the model to be with low unemployment rate.

After analyzing the different cost of mistakes, predictive models with high recall rate were prioritized. This was due to False Positive predictions were considered to be the most undesirable outcome among all. Imagine a person chose to move to a constituency to work based on our prediction of the constituency is low in unemployment rate, but in fact the constituency is high in unemployment and is very hard to find jobs.

5.5. Predictive Models

5.5.1. Logistic Regression with LASSO

Even after attribute removal, there are still more than 40 available features. And hence, it was vital to first identify which features have more significant correlation with the predictor variable. The LASSO method was then implemented to select the "best" variables to use for a logistic regression model using the principled way. See Figures 6, 7, 8.

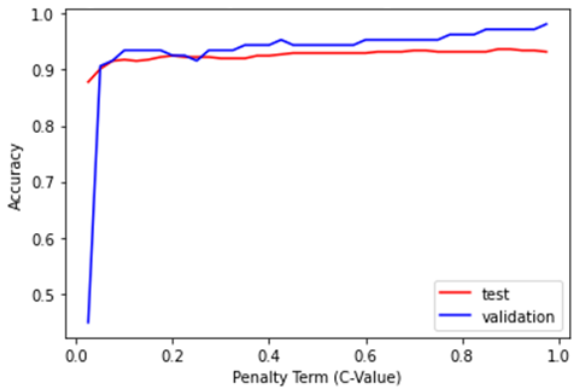


Figure 6: Accuracy Score of training and validation set.

According to these Figures, the penalty term (C-Value) peaked at 0.1 for both accuracy and recall. And as mentioned before, the performance metrics of recall to minimize false positive was prioritize, so the value of 0.1 of C was identified to be the optimum value.

Table 3: Summary of metrics of performance for logistic regression with LASSO at C=0.1.

Type of dataset	Accuracy	Precision	Recall
Accuracy	0.918	0.922	0.941
Precision	0.916	0.931	0.915
Recall	0.935	0.933	0.949

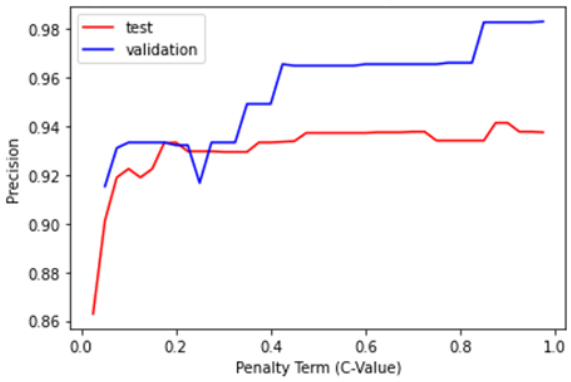


Figure 7: Precision Score of training and validation set.

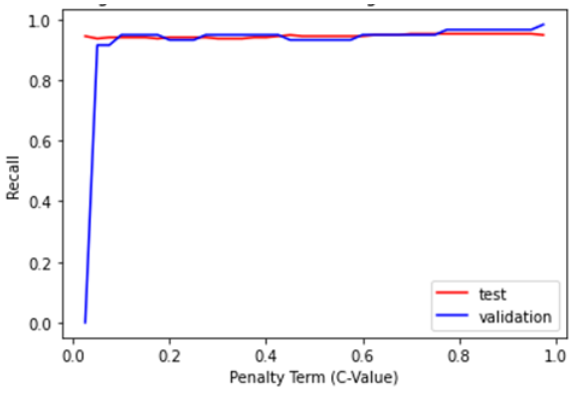


Figure 8: Recall Score of training and validation set.

5 features were selected to have relatively significant correlationship with predictor variable with C = 0.1, as seen in table 2. The 5 features would be used for further decision tree predictive model construction.

5.5.2. Decision Tree

Decision Tree was employed to construct a predictive model built on Gini Impurity, an useful measurement of the probability of a randomly assigned datapoint is incorrect. Graphs with Recall plotted against the Maximum Depth and Minimum impurity was constructed for better identification of the optimal value for both depth and impurity, and to

Table 4: Summary of chosen features with LASSO at C=0.1.

Features	Coefficients
Turnout (proportion)	0.303
Share of LSOAs (small areas)	-1.092
Obesity	-0.026
Schizophrenia, bipolar, psychoses	-0.336
Age group (0-9)	-0.048

prevent overfitting of the model. See Figure 9.

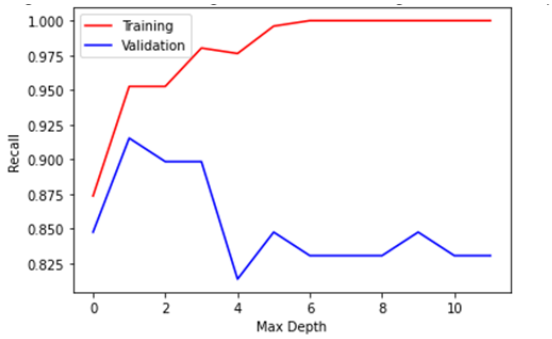


Figure 9: Recall of Training and Validation Data against Max Tree Depth.

Figure 10 shows the validation recall peaked at the range of 1-3 and drops significantly after the depth of 3. This notable drop in recall on validation data demonstrates the overfitting of the training data.

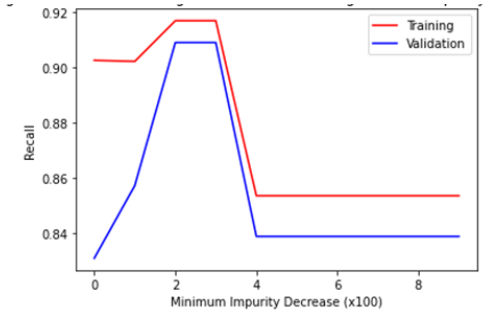


Figure 10: Recall of Training and Validation Data against Min Impurity Decrease.

Figure 11 shows the recall peaked at a minimum impurity decrease of 0.02. To improve the predictive nature of our model, the minimum impurity decrease of 0.02 was chosen to maximise recall rate while a max tree depth of 3 was used to prevent overfitting.

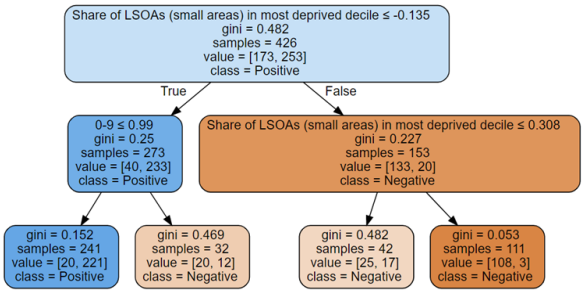


Figure 11: Decision Tree of the Final Model (max depth = 3, minimum impurity increase = 0.02

5.6. Discussion & Conclusion

5.6.1. Results

The two chosen machine learning models were tested against the test dataset as shown in Table 5.

Table 5: Summary of metrics of performance for logistic regression with LASSO at C=0.1.

Performance Metric	Logistic Regression (LASSO)	Decision Tree
Accuracy	0.935	0.925
Precision	0.933	0.947
Recall	0.949	0.915
Δ Recall	+0.034	+0.068

The results showed both chosen models achieve considerably good predictive performance, with accuracy of 92.5% or higher during final testing. The results also suggested both models were not overfitting owing to the considerably insignificant difference between validation and test sets (<=0.068).

The Logistic Regression with LASSO was selected as the final predictive model to identify whether a constituency is with low unemployment or not due to its exceptional recall rate of 94.9%. Despite its lower precision than the decision tree model, the ultimate aim of this model is to minimize False Positive predictive outcome as explained in section 3.4. And hence, with higher accuracy and recall, the logistic regression with LASSO was considered to be a better and more suitable predictive model.

5.6.2. Evaluation

Limitation of the predictive models of this study include not using the most up-to-date data. The models were constructed on the 2021 census dataset of the UK Parliament[2], and hence, the results may differ due to the time accuracy of the data. Other limitations include not considering factors such as number of business, economic output, cost of living that may also have a sizeable impact on unemployment rate of a constituency. The predictive models in this study are solely built upon the available datasets from the common library of UK Parliament.

6. Brexit Voting Patterns - Owain Pill

6.1. Aim

Predicting whether a constituency voted for Brexit based on data about it.

6.2. Data Preparation

6.2.1. Dataset Observations

From Figure 12, the data looks to be slightly negatively distributed with a spike in values around 0.55.

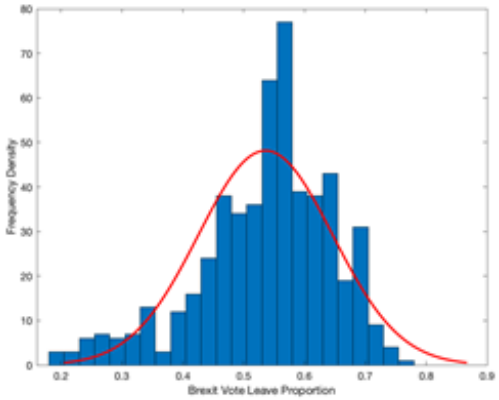


Figure 12: Histogram of Brexit voting data with normal distribution fitted.

6.2.2. Linear Regression

The R^2 value was used to see correlation between Brexit votes and individual features. The 5 features with the highest correlation are shown in Table A.9.

Table 6: R^2 between features and Brexit voting ratios.

Feature	Correlation
Brexit	1.000
Brexit Predicted	0.9919
CON%Lev_4_Qual	0.7681
CON%PROF	0.7655
CON%Lev_1_Qual	0.719

The most closely correlated feature that gives an informative result is *CON%Lev_4_qual*, a graph of which is shown in Figure 13.

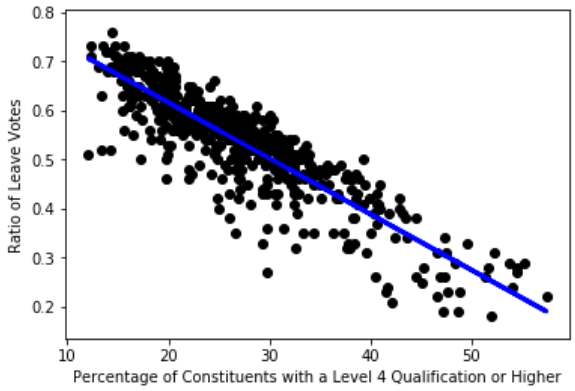


Figure 13: Relationship between Leave vote ratio and Level 4 Qualifications

6.2.3. Feature Engineering

The dependent variable (Brexit) was converted from a ratio to a binary value (1 or 0).

The threshold for this conversion was decided to be 0.5 as this would decide whether a constituency as a whole contributed to voting to leave the EU. These values were assigned to a new column 'Brexit binary'.

As the median of Brexit voting ratios was 0.55, there would be fewer constituencies that voted to remain. This number was found to be 172 vs 394 constituencies that voted to leave. The larger group was randomly undersampled so that both groups would match thereby avoiding dataset imbalance. Random selection with no variable should lead to 0.50 accuracy.

6.2.4. Attribute Removal

After binarization, a number of columns were removed from the dataset to make predictions better.

- *Constituency Name.* This was just the id for each row.
- *Conservative votes, Labour votes, Lib Dem votes.* Proportional figures for this data was included.
- *Brexit, Brexit Predicted.* This data had been binarised so needed to be removed.

6.2.5. Data Splitting

This updated dataset was then split into training, validation, and test. The split was 50%, 25%, 25%. This larger validation set helped minimize overfitting to the validation set in forward selection.

6.2.6. Performance Metrics

To measure the performance of the datasets, the cost of different sorts of mistakes were analyzed.

Considering that the data was nearly normally distributed and centered around 0.55 which is close to the threshold value and that the penalty for either false negatives or false positives is not fatal or costly, accuracy was chosen as the performance metric. This prioritizes getting the greatest proportion of correct predictions on a scale from 0 to 1.

6.3. Logistic Regression

Logistic regression was chosen as the method because it is useful for finding the correlation between binary dependent variables and many independent variables. To evaluate the effectiveness of any model that would be created, baseline values for accuracy were needed given different conditions.

A graph of accuracy vs number of features included in the model was created. Each iteration had completely different variables so there was a lot of noise in the signal. The training data accuracy (0.90 overall average) and validation data accuracy (0.84 overall average) were well above randomly picking (0.50). The average accuracy of training and validation data is achieved after less than 10 features.

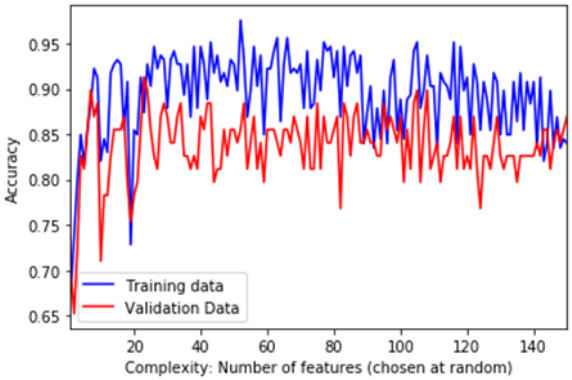


Figure 14: Accuracy vs random features

The same graphing method was used with an ordered list of the R^2 values from highest to lowest obtained from the linear regression. The results showed that there was an initial performance increase when there were fewer variables as these were highly correlated and that adding more variables did not improve either the training or validation accuracies beyond 30 variables as shown in Figures 15 and 16.

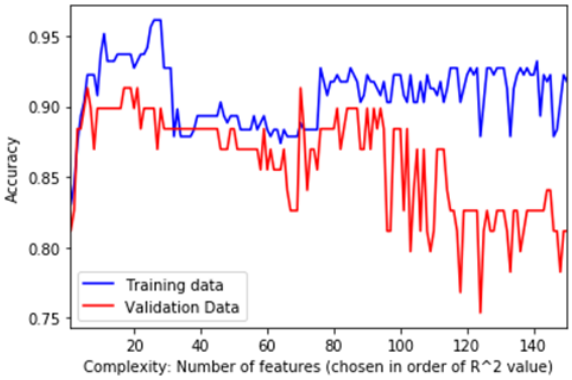


Figure 15: Accuracy vs features by order of R^2

A 'Forward Selection' algorithm was created to automatically pick a specified amount of the best features based on validation accuracy. Every cycle, the variable with the highest accuracy was added. This meant that accuracy could stay the same or drop after a cycle. The accuracy was for train and validation was higher (0.93 average and 0.92 average respectively, Figure 17) than for the previous two models so was taken forward.

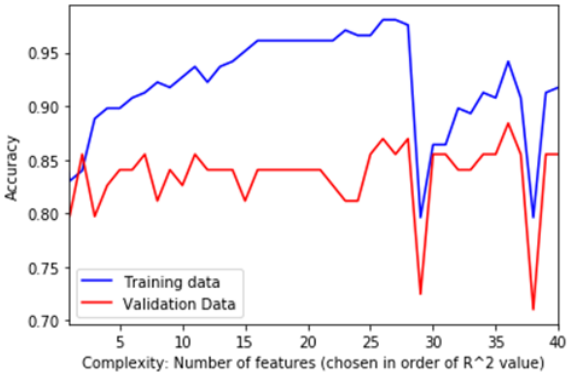


Figure 16: Accuracy vs features by order of R^2

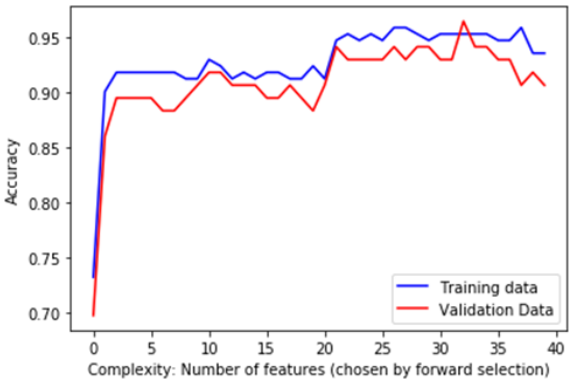


Figure 17: Accuracy vs features by forward selection

The model had many variables with high p-values which meant there was a lot of prediction overlap between variables. To mitigate this, the data was normalized and the Lasso method (penalty = 'l1', C = 0.1) was used. This removed 33 variables causing a training accuracy drop from 0.948 to 0.89 and a rise in validation accuracy from 0.90 to 0.93.

A final measure was to manually remove features that were clearly related to each other. An example of this was removing 'Labour Vote (proportion)' as 'Conservative Vote (proportion)' was already included in the model. After each removal, the models' accuracies were checked to prevent underfitting. The final model is shown below in Figure 18 and Table A.9.

This model was then run without the data being undersampled and achieved a test accuracy of 0.926 which confirmed its robustness.

6.4. Discussion

The model selected provides very good predictive power of whether a constituency will vote for Brexit

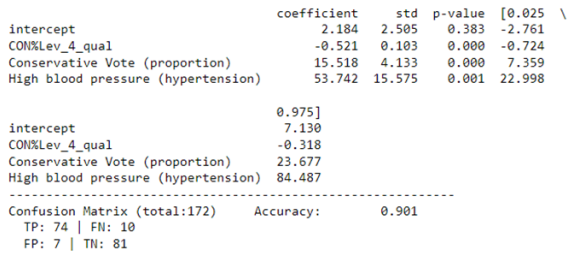


Figure 18: Features, pvals and Confusion Matrix for final model

Table 7: Accuracy of final model

Performance Metric	Value
Training Accuracy	0.901
Validation Accuracy	0.907
Test Accuracy	0.907

or not if there were to be a second referendum (accuracy > 0.9). The fact that it is only made up of three features means that it is less prone to overfitting. This is reinforced by the fact that train, validation, and test accuracies are all very similar. Blood pressure might seem an unlikely feature but probably has to do with providing a good age range as older people are far more likely to suffer from high blood pressure (see Seaborne plot) – other age ranges are present in the data but only in increments of 10 years.

A limitation is that while the results from the final model seem very high, the dataset lends itself to high accuracies as demonstrated by the fact that randomly picking more than 10 variables leads to accuracies of over 80% so the bar for performance is not 0.5.

The data is from different years – some from the 2011 census while other data is from a 2019 constituency review. This will lead to discrepancies with newer data as society changes over time. The dataset is also small which led to there being differences in final results after every time running the code. This was offset by the low number of features in the model and after running the entire code 10 times, the average final undersampled test accuracy was 0.91 and for the whole dataset, the accuracy was 0.93.

In terms of deciding where to live, we can only predict constituency level support. However most people will live in much smaller communities so it difficult to generalise from our results to a granular level. Social policy is also largely made at the federal level in the UK so political leanings of an area matter less than the economic situation when it comes to quality of life. When compared to somewhere like the USA where states have a lot more control over an individual's life (for example trigger bans on abortion), it can be seen that knowing the political leanings of an area can be vital when deciding where to move.

7. Number of Businesses - Chinene Chukwuma

7.1. The Dependent Variable

The number of businesses per capita (NBC) is an important factor to consider when looking for an ideal place to live. A higher number of businesses per capita mean that, proportionally, there are more job opportunities per person.

7.2. Data Processing

7.2.1. Dataset Choice

There are many factors to consider when selecting a dataset as the characteristics of the dataset can determine whether the model is overfitted. Overfitting is displayed when the accuracy of the training set is too high. This means the model will not work for other concepts and unseen datasets so scores poorly on the test data. The 2011 census was chosen with these considerations in mind.

7.2.2. Clean Up

The following variables present in the dataset were removed as, though there may be a correlation with the number of businesses per capita, they were deemed to not have a causal relationship:

1. Election Results
2. Health Conditions
3. Religion
4. Ethnic Groups

A brief exploratory analysis of the dataset was conducted with respect to the dependent variable. Analysis revealed there to be an anomaly in the data, the 'Cities of London and Westminster', which had a value of 1.4 while other constituencies had NBC values below 0.4. See Figures 19 and 20. This was removed to not skew the data and negatively affect the accuracy of the prediction. Other studies on the number of businesses per capita were noted to have taken the same course of action [1].

7.2.3. Binarisation

Originally, the NBC was in continuous form but was changed as a linear regression model to predict future values was not the purpose of this report. Consequently, the mean number of businesses per capita, 0.0570 (3 s.f.), was used to binarise the predictor attribute.

If the number of businesses per capita for a constituency was within the range of 0 ¡ x ¡= 0.0570

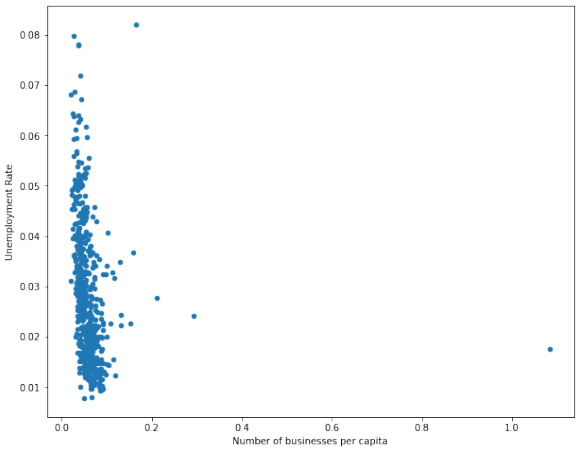


Figure 19: NBC (before data cleanup)

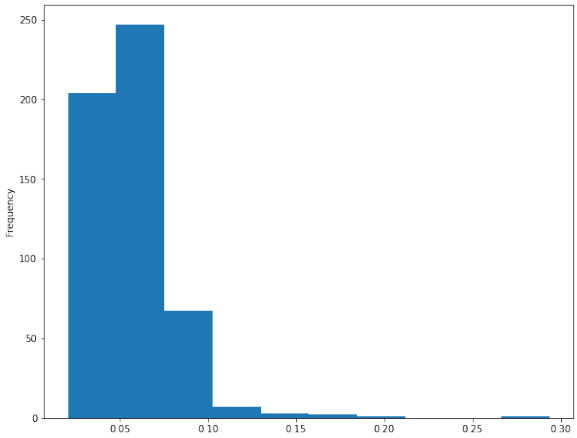


Figure 20: NBC histogram (before data cleanup)

(3 s.f.), it was classified as 0 (low NBC), otherwise they were classified as 1 (high NBC).

In addition to being a balanced dataset, using the mean to binarise the NBC mitigates instances of model bias as everything is balanced. This means accuracy was a suitable and effective metric to assess the competency of a model. See Figure 21

7.2.4. Predictive Model

Due to the chosen predictive model and its coding method, changes were made to the dataset including the removal of null values. The dataset was also subject to label encoding. This ensured that no errors would occur when running vital code.

The dataset was also split up into 80% training data, 10% validation data, and 10% test data.

7.2.5. Performance Metrics

The three metrics used the asses the chosen predictive model included:

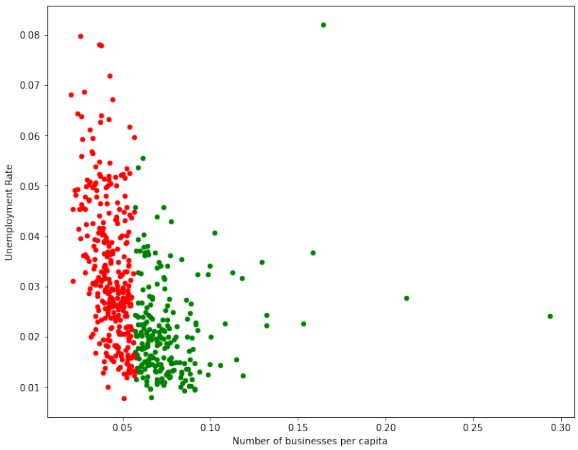


Figure 21: NBC to unemployment post cleanup and binarised

1. Accuracy: The proportion of constituencies correctly predicted to have a high NBC.
2. Precision: Out of all the constituencies predicted to have a high NBC, what percentage was correct.
3. Recall: The proportion of high NBC constituencies that were predicted to have a high NBC.

7.3. Method Selection

7.3.1. Comparison

Classification methods such as the SVM classifier were not considered as the parameters for what counts as a high NBC and low NBC were clearly defined.

Decision trees and random forests find the variable with the minimum weighted impurity level. Random Forests use many randomly generated decision trees to produce the same variable. They have greater accuracy than a single decision tree that is prone to overfitting. The problem with this method is that the final model only considers one variable and is most effective when used with datasets with multiclass classifications.

Forward and backward selection, though effective at finding a combination of features with high predictive power, are generally incapable of finding the global maximum. On the other hand, they are more realistic as they take into account the effect of different combinations of variables.

7.3.2. Justification

Choosing whether to live somewhere is binary in nature so logistic regression was deemed most suitable. The logistic distribution restricts the estimated probabilities (y) between 0 (low NBC) and 1 (high NBC).

The chosen predictive model was backward selection. This method was chosen instead of forward selection as there aren't many features to consider. Considering as many features as possible provides a holistic view of the causes and effects of different variables on the NBC.

7.4. Logistic Regression Backward Selection

7.4.1. Description

Logistic Regression Backward Selection first starts by considering all the features in a dataset, then calculating the accuracy before removing the feature with the least predictive power. This process is repeated until there are no more variables to remove or the validation accuracy has decreased by more than a certain amount.

7.4.2. Model Tuning

At first, the accuracy of the model was 0.96 (3 s.f.) so to avoid overfitting, the 'Number of Businesses' variable was removed as it was too closely related to the NBC, increasing the accuracy. The accuracy was reduced to a maximum of 0.814 (3 s.f.). See Figures 22 and 23

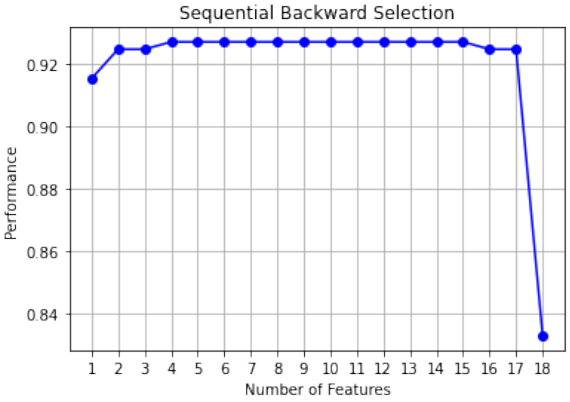


Figure 22: Backwards selection with number of businesses

7.4.3. Adjusting Hyperparameters

The hyperparameter called *k_feature*, which, in this case, is the minimum number of features your model will select, was defined as 'best' to find the best combination of features. Changing this hyperparameter only resulted in a loss of information.

The greatest change was visible when the cross-validation hyperparameter (cv) was increased from 0 to 5. CV determines the splitting strategy of the model [2] so results in more averaged values, reducing the risk of overfitting. See Figure 24.

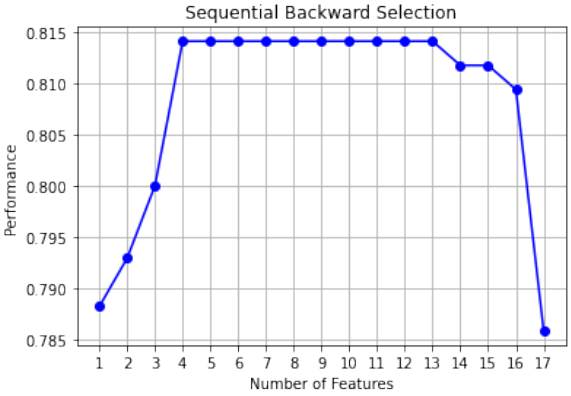


Figure 23: Backwards selection without number of businesses

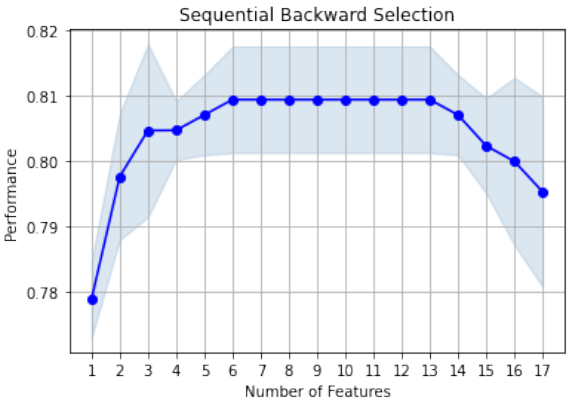


Figure 24: Backwards selection with increased cross-validation

7.5. Results Summary

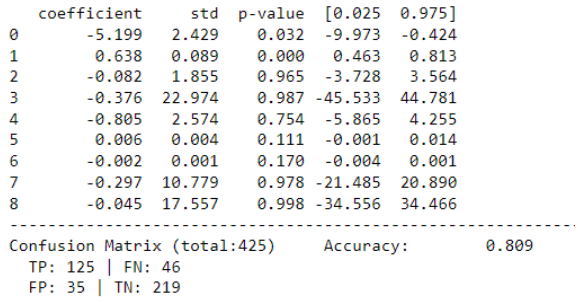
7.6. Discussion

Nine out of seventeen features were determined to be the 'best' feature combination with the most predictive power. These features included: Unemployment Rate, Social Mobility Index, and Home Ownership. Potential observations one could make from this include that areas with a high number of businesses per capita can be closely associated with a high social mobility index, making them ideal places to live long-term.

In terms of the evaluation metrics, though the recall values are much lower in comparison to the

Table 8: Final model results summary

Performance Metric	Training	Validation	Test
Accuracy	80.941	75.472	83.333
Validation Accuracy	80.941	75.472	88.889
Test Accuracy	40.235	50.943	46.296



	coefficient	std	p-value	[0.025	0.975]
0	-5.199	2.429	0.032	-9.973	-0.424
1	0.638	0.089	0.000	0.463	0.813
2	-0.082	1.855	0.965	-3.728	3.564
3	-0.376	22.974	0.987	-45.533	44.781
4	-0.805	2.574	0.754	-5.865	4.255
5	0.006	0.004	0.111	-0.001	0.014
6	-0.002	0.001	0.170	-0.004	0.001
7	-0.297	10.779	0.978	-21.485	20.890
8	-0.045	17.557	0.998	-34.556	34.466

Figure 25: Final model confusion matrix

accuracy and precision values, there is still an increase after training data results for all values. This proves that the selected feature subset improves the performance of the model in all aspects. Lower recall values imply that proportionally, it is a model with lower false positives (35). For this study, a lower number of false positives is better so one can be sure there is a higher chance of a constituency identified as ideal, actually being ideal.

Overall, this is a good model that is not overfitted and has a high chance of correctly predicting whether an area has a high or low number of businesses per capita.

8. Conclusion

Combined, the team has created a series of models that accurately predict the major characteristics of a constituency: political leanings, unemployment, and the number of businesses per capita. Logistic regression was the primary base used throughout as choosing whether or not to live in a constituency is a binary decision.

The study prioritised having a lower number of false positives so constituencies identified by the models as ideal, had a higher likelihood of being ideal in reality.

A logistic regression model was built to predict whether the Conservative Party would win the next election in an area. By using an optimised automatic forward selection method, a model obtained where predictions could be made with 80.4% accuracy while only using two features. This avoids the risk of overfitting meaning the model is widely applicable.

While the model to predict voting tendencies in a constituency used forward selection, the Brexit votes predictive model applied forward selection. The model selected provides very good predictive power of whether a constituency will vote for Brexit or not if there were to be a second referendum (accuracy ¿ 0.9). The fact that it is only made up of three features means that it is less prone to overfitting. After many iterations, due to a high infection in data, the average final test accuracy was 91%.

LASSO was used to predict unemployment. The selection model carried out effective features selection with a large penalty term applied until 5 features remained. The selected features were considered to have a significant correlation with the predictor variable (unemployment rate) and were used for Decision Tree model for further predictive and comparison purposes. The model has a high training accuracy of 93.5%.

On the other hand, the model for the number of businesses per capita utilised Backward Selection so all features were considered before the feature combination with the most predictive power was selected. Using backward selection means you are more likely to end up with a local maximum that contains a longer list of features. This provides a more holistic view of the causes and effects of different variables on the NBC, resulting in a model with 83.3% accuracy.

In conclusion, the predictive models classified constituents as ideal and unideal effectively. In practice, they would be combined to create an algorithm where individuals have the power to define whether more or less of a feature is ideal for them and returns a list of constituents.

Appendix A. Number of Businesses - Chinene Chukwuma

Appendix A.1. Predictive Methods Comparison

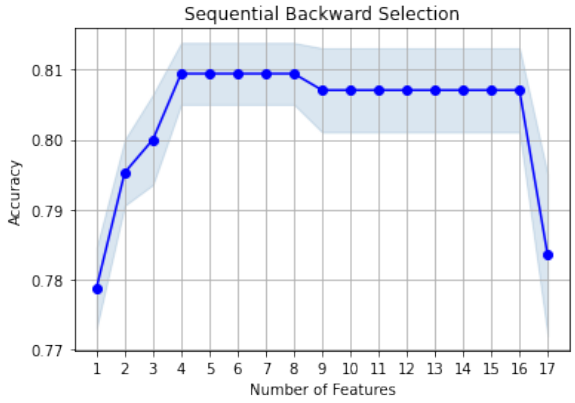


Figure A.26: Training accuracy of backwards selection

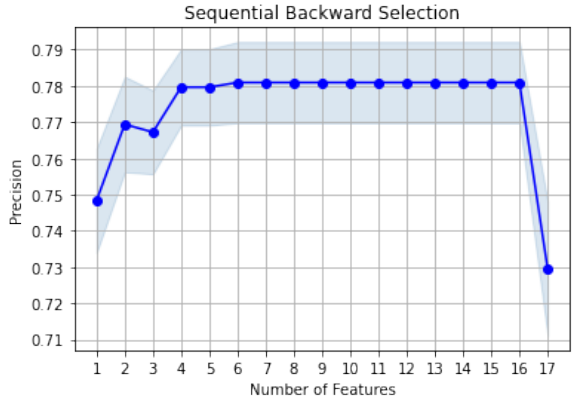


Figure A.27: Training precision of backwards selection

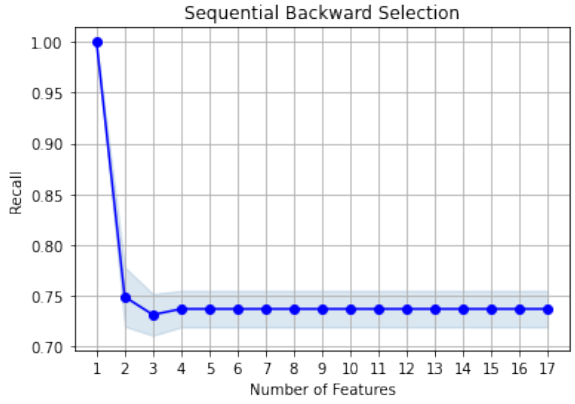


Figure A.28: Training recall of backwards selection

	feature_idx	avg_score
17	(0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13,...	0.783529
16	(1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14...	0.807059
15	(1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14...	0.807059
14	(1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14)	0.807059
13	(1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13)	0.807059
12	(1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12)	0.807059
11	(1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11)	0.807059
10	(1, 2, 3, 4, 5, 6, 7, 8, 9, 10)	0.807059
9	(1, 2, 3, 4, 5, 6, 7, 8, 9)	0.807059
8	(1, 2, 3, 4, 5, 6, 8, 9)	0.809412
7	(1, 2, 3, 4, 5, 6, 8)	0.809412
6	(1, 2, 3, 4, 5, 6)	0.809412
5	(1, 2, 4, 5, 6)	0.809412
4	(1, 4, 5, 6)	0.809412
3	(1, 4, 5)	0.8
2	(1, 5)	0.795294
1	(1,)	0.778824

Figure A.29: Backward selection process

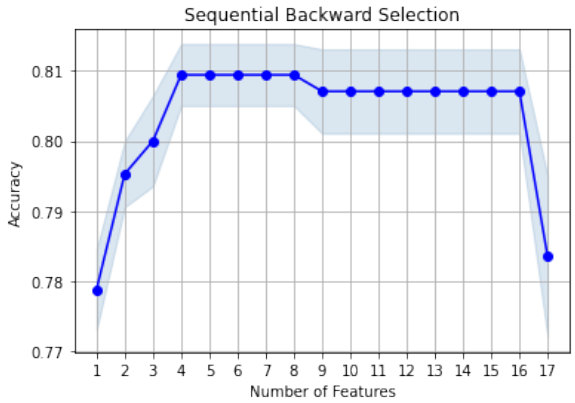


Figure A.30: Features with the most predictive power

Table A.9: Accuracy of final model

Method	Outcome	Advantages	Disadvantages
LR Forward Selection	Finds the combination of features with the most predictive power working forwards.	More efficient than backward selection when there is a large number of data variables [3].	Does not result in the global maximum.
LR Backward Selection	Finds the combination of features with the most predictive power working backward.	More effective than forward selection for a small number of variables [3].	Finds the local optimum combination, not the best combination.
LASSO	It helps selects the best variables for predictive analysis.	Useful if there are several insignificant variables.	Does not show how variables behave when combined.
Decision Tree	The variable with the minimum weighted impurity level.	The resulting variable is likely to give accurate results.	Prone to overfitting[4]. Cannot guarantee decision tree is optimum.
Random Forest	Out of many decision trees, the variable with the minimum weighted impurity level.	Takes many randomly generated decision trees into account. Greater accuracy than a single decision tree	Harder to interpret than decision trees and cannot view the impurity level of individual variables[4].
SVM Classifier	Finds the best separation between classes.	Works well when there is a clear distinction between classes.	Will be harder to execute if there is a lot of overlapping between classes [5].

Appendix B. Code