Brexit means brexit!

Ahmed Ghafouri

20 Mar 2021

knitr::opts_chunk$set(echo = TRUE, fig.align="center")

knitr::include_graphics() ## Introduction

In 2016, the United Kingdom voted to leave the EU in one the of the greatest upsets in 21st century politics. With vote remain being the heavy favorite, many people have questioned the reasons for the great upset. This report will look at data obtained from the guardian website with the 5 following variables across all the electoral wards within the Uk:

abc1 - The ratio of individuals who were upper to middle class in that electoral

medianIncome - the median income of all the residents within that electoral ward

medianAge - the median of age of residents in that electoral ward

WithHigherEd - The proportion of residents with university level education in that electoral ward

notBornUk - The proportion of residents not born in the Uk in that electoral ward

All the variables have been normalized so that they can be assessed against each other fairly without scaling bias in any model to be built.

The effect each of these variables had on the way an electoral ward would vote will be analyzed using logistic regression and assessed against the results found on the Guardian page.

The variables will also be assessed against each other with the aim of understanding which variables had the strongest effect and which had the least. Logistic regression coefficients, 'greedy' logistic regression building and decision tree models will all be built and compared to understand the effect of each variable.

```
#importing the gurdain data
brexit=read.csv2('C:/Users/Ahmed/Desktop/Statistical Learning/Practical assessment 2/brexit.csv',
                 header = TRUE,sep = ",", quote = "\"",dec = ".")
```

Logitic regression - all 5 inputs

The first model built will be the logistic regression with all 5 variables as inputs.

The logistic regression takes normal linear regression equation as shown bellow where there are 5 variables labeled x1 to x5 and beta 1 to beta 5 representing the variables in this data set (abc1, median income etc)

$$Z = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 \quad (eq\ 1)$$

This linear regression input labeled z is fed into a sigmoid function shown in equation 2 bellow

$$p = \frac{1}{1 + \exp(-z)} \quad (eq\ 2)$$

$$p = \frac{1}{1 + \exp(-(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5))} \quad (eq\ 3)$$

In order to build the logistic regression, a glm function with a family input equal to binomial will be used. The coefficients will be solved numerically by the glm function used bellow

```
myglm = glm(voteBrexit ~ ., family=binomial, data=brexit)

summary(myglm)
```

```
##
## Call:
## glm(formula = voteBrexit ~ ., family = binomial, data = brexit)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.9793  -0.2296   0.3073   0.6032   2.0177
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)   -0.1386     0.8477  -0.164 0.870122
## abc1          17.5780     2.9114   6.038 1.56e-09 ***
## notBornUK      5.6861     1.8033   3.153 0.001615 **
## medianIncome  -6.3857     1.9217  -3.323 0.000891 ***
## medianAge      5.9209     1.4066   4.209 2.56e-05 ***
## withHigherEd -26.7443     3.5762  -7.478 7.52e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 426.52  on 343  degrees of freedom
## Residual deviance: 247.39  on 338  degrees of freedom
## AIC: 259.39
##
## Number of Fisher Scoring iterations: 6
```

From the logistic regression model built by the glm, the following coefficient (beta) is found.

	Estimate
(Intercept)	-0.1386
abc1	17.5780
notBornUK	5.6861
medianIncome	-6.3857
medianAge	5.9209
withHigherEd	-26.7443

The coefficients show how strongly a variable effects an electoral ward to vote remain or leave in the brexit vote in 2016. The more negative the coefficient, the more this variable pushes to suggest this electoral ward would vote remain and, the more positive a coefficient, the more this variable pushes to suggest this electoral ward is in favor of leave. The closer a coefficient is to 0, the weaker the variables impact on the way a constituency would vote.

The coefficient for percentage of the electoral ward with a higher education degree is -26.7 which is very strongly negative compared to the other coefficients. This shows that the model would associate having a higher education degree with being heavily likely to vote remain. This is in accordance with the figure on the guardian website labeled figure 1. In figure 1 from the guardian, as percentage of residents with higher education decreases, the number of constituency to voting to leave increases in a very strong negative correlation manner.

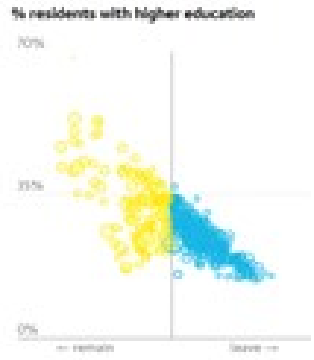


Figure 1: higher education

The coefficient for median annual income is -6.4 which is slightly negative and no where near as negative as the coefficient for higher education. This shows that the model associates having a higher median income with being slightly more likely to vote remain but not at no where near the magnitude it does having higher education qualification. This is in concordance with the guardian website figure labeled figure 2 which shows a weak negative correlation between median annual income of constituency and percentage of residents voting leave.

Figure 2: Median annual income

The coefficient for ABC1 is approximately 17.6 which is strongly positive suggesting that ABC1 has a strong positive correlation with voting leave. This goes against the guardians figure in figure 3 which shows negative correlation between abc1 and boting to leave. This negative correlation can be seen by an increase in vote leave voters as the percentage of residents of ABC1 social grade decreases. The guardian figure data is however very sparse for vote remain and very dense on the vote remain side. It is likely the case that the number of vote leave electoral ward s with around 50% to 60% are very dense, swaying the an increase of abc1 to cause an increase the vote leave campagin vote per electoral ward.

The coefficient for median age of residents is 5.9 which is a small positive value. A small positive value suggests that median age has a slight positive correlation with voting leave where as the median age increases, the likelihood of voting leave also increases. The low positive value of the coefficient suggests that median age does not have an incredibly strong affect on a constituency voting leave, especially compared to abc1. This is in line with the guardian website figure labeled figure 4, which shows median age not having much

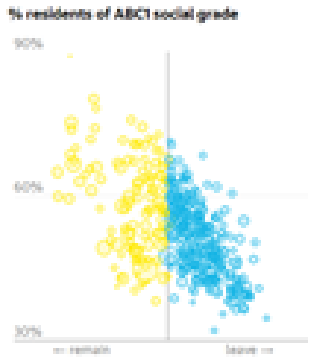


Figure 3: abc1 social group

correlation between median age and voting choice but does show a very slight trend in more electoral ward s with median age over 40 being more inclined to vote leave.

Figure 4: Median age

The coefficient for residents born outside of the uk is 5.7 suggesting that being having a greater percentage of residents born outside the Uk would lead to slightly greater percentage of residents in that electoral ward voting leave. This goes against the figure on the guardian website labeled figure 5 which shows not much difference in voting for electoral ward s with less than 30% residents not born in the uk but electoral ward s with high percentage of residents not born in the being more inclined to vote remain. This again is likely to be due to the density of data points on the vote leave side in the guardian figure bellow being more abundant/densely packed than can be perceived through this figures.

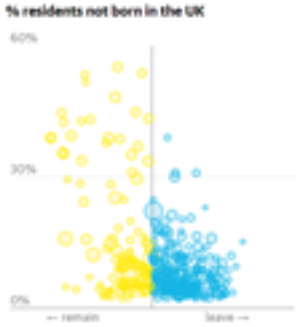


Figure 5: born outside of the uk

Variable strength according to coeffcient

The coefficients with the highest magnitudes in the positive or negative direction will have the strongest effect. The largest magnitude in the logistic regression glm model is the with higher education variable, followed by abc1 and so on, with not being born in the uk having the weakest effect on the way in which an electoral ward would vote.

Coefficient effect from strongest to weakest based solely on coefficient. Estimate withHigherEd -26.7443 abc1 17.5780 medianIncome -6.3857 medianAge 5.9209 notBornUK 5.6861

However, the coefficient come with uncertainty. If uncertainty is high, there is a possibility the true affect a variable and its coefficient would have on the way an electoral ward would vote may differ quite a bit. Furthermore, overlapping confidence intervals would have an affect on ordering the inputs based on strongest affect as it may be unclear where the true value lies.

In order to quantify uncertainty, the 95% confidence interval is taken. Within the calculated confidence interval for each coefficient, the true value of the coefficient is likely to be within that range with a 5% chance of a type 1 error. A type 1 error being when a true null hypothesis is rejected.

The confidence interval for each variable is shown in the output of the following code.

```
zc = qnorm(0.975)
#Extract estimate and standard error of coefficient from model summary (check above)
estimate = summary(myglm)$coefficients[2,1]
standard_error = summary(myglm)$coefficients[2,2]
#Calculate and print the lower and upper boundaries of the confidence interval
CI_min = estimate - zc*standard_error; CI_max = estimate + zc*standard_error
paste('abc1 Confidence interval:',CI_min,',' ,CI_max)
```

```
## [1] "abc1 Confidence interval: 11.8717241841725 , 23.2842718890574"
```

```
zc = qnorm(0.975)
#Extract estimate and standard error of coefficient from model summary (check above)
estimate = summary(myglm)$coefficients[3,1]
standard_error = summary(myglm)$coefficients[3,2]
#Calculate and print the lower and upper boundaries of the confidence interval
CI_min = estimate - zc*standard_error; CI_max = estimate + zc*standard_error
paste('Not born in the UK Confidence interval:',CI_min,',' ,CI_max)
```

```
## [1] "Not born in the UK Confidence interval: 2.15166883329817 , 9.22060779588884"
```

```
zc = qnorm(0.975)
#Extract estimate and standard error of coefficient from model summary (check above)
estimate = summary(myglm)$coefficients[4,1]
standard_error = summary(myglm)$coefficients[4,2]
#Calculate and print the lower and upper boundaries of the confidence interval
CI_min = estimate - zc*standard_error; CI_max = estimate + zc*standard_error
paste('Median income Confidence interval:',CI_min,',' ,CI_max)
```

```
## [1] "Median income Confidence interval: -10.1522039945514 , -2.61927527070331"
```

```r
zc = qnorm(0.975)
#Extract estimate and standard error of coefficient from model summary (check above)
estimate = summary(myglm)$coefficients[5,1]
standard_error = summary(myglm)$coefficients[5,2]
#Calculate and print the lower and upper boundaries of the confidence interval
CI_min = estimate - zc*standard_error; CI_max = estimate + zc*standard_error
paste('Median age Confidence interval:',CI_min,',' ,CI_max)
```

```
## [1] "Median age Confidence interval: 3.16406662412453 , 8.67768687448136"
```

```r
zc = qnorm(0.975)
#Extract estimate and standard error of coefficient from model summary (check above)
estimate = summary(myglm)$coefficients[6,1]
standard_error = summary(myglm)$coefficients[6,2]
#Calculate and print the lower and upper boundaries of the confidence interval
CI_min = estimate - zc*standard_error; CI_max = estimate + zc*standard_error
paste('With Higher Education qualification Confidence interval:',CI_min,',' ,CI_max)
```

```
## [1] "With Higher Education qualification Confidence interval: -33.7534564486837 , -19.7350619427495"
```

Abc1 has 23 within its confidence interval and the 'with higher education' has -19 within its confidence interval. The magnitude of the values within these confidence intervals overlap meaning there is a possibility that abc1 could have a greater affect on the way an electoral ward would vote and thus the greatest affect on the way an electoral ward would vote. This leads to the feasibility being questioned of with higher education having greater effect on the way electoral ward s would vote.

Median income and abc1 have no overlap in magnitudes within their own confidence intervals so there is at most a 2.5% chance of both values having type 1 errors where median income and abc1 coefficients are outside of the confidence interval giving chance for median income to be greater than the abc1 coefficient. The calculation of the probability of having median income of and abc1 both being type 1 errors can be seen bellow.

$$(type\ 1\ error\ abc1\ 5\%) * (type\ 1\ error\ median\ income\ 5\%) \ = \ 2.5\%$$

Median income, median age and not being born Median income, median age and not being born in the uk have very large overlaps in magnitude of confidence intervals. This suggests that the order of magnitude defined based solely on coefficients out of the logistic regression glm model may be prone to inaccuracy for the order of magnitude of these 3 variables.

Greedy method

Another way to test the strength of the 5 variables would be to build a model gradually, only adding the gives the greatest decrease in AIC the logistic regression glm model. AIC is used as it penalizes too many parameters being used in the model. Too many unnecessary parameters is something to be avoided as too many parameters could lead to overfitting of the regression model. If the model overfits, it will struggle to generalize well when using unseen data to make predictions

The code and output AIC of the logistic regression model with one input is as follows:

```r
# one input
a=c('voteBrexit ~abc1','voteBrexit ~medianIncome','voteBrexit ~withHigherEd','voteBrexit ~medianAge','v
for (i in 1:length(a)){
```

```
  current_model = glm(formula = a[i], family=binomial, data=brexit)
print(a[i])
print(AIC(current_model))}
```

```
## [1] "voteBrexit ~abc1"
## [1] 377.5437
## [1] "voteBrexit ~medianIncome"
## [1] 368.4437
## [1] "voteBrexit ~withHigherEd"
## [1] 313.5604
## [1] "voteBrexit ~medianAge"
## [1] 401.2767
## [1] "voteBrexit ~medianIncome"
## [1] 368.4437
```

With one input, the logistic regression model built using the glm function shows the model built with just the with higher education variable has the lowest AIC of approximately 313.6

Implementing the greedy method, models with 2 inputs will all contain the wither higher education variable. The code for building a glm logistic regression with 2 variables, one of which being with higher education is as follows:

```
## abc 2 input
a=c('voteBrexit ~abc1+withHigherEd')
for (i in 1:length(a)){

  current_model = glm(formula = a[i], family=binomial, data=brexit)
  print(a[i])
  print(AIC(current_model))}
```

```
## [1] "voteBrexit ~abc1+withHigherEd"
## [1] 286.5454
```

```
# 2inputs median income
a=c('voteBrexit ~medianIncome+withHigherEd')
for (i in 1:length(a)){

  current_model = glm(formula = a[i], family=binomial, data=brexit)
  print(a[i])
  print(AIC(current_model))}
```

```
## [1] "voteBrexit ~medianIncome+withHigherEd"
## [1] 315.5256
```

```
a=c('voteBrexit ~notBornUK+withHigherEd')
for (i in 1:length(a)){

  current_model = glm(formula = a[i], family=binomial, data=brexit)
  print(a[i])
  print(AIC(current_model))}
```

```
## [1] "voteBrexit ~notBornUK+withHigherEd"
## [1] 310.3644
```

```
## 2 inputs  median age
a=c('voteBrexit ~medianAge+withHigherEd')
for (i in 1:length(a)){

  current_model = glm(formula = a[i], family=binomial, data=brexit)
  print(a[i])
  print(AIC(current_model))}
```

```
## [1] "voteBrexit ~medianAge+withHigherEd"
## [1] 303.3091
```

From the 2 input models that all must involve the with higher education variable, the model with abc1 added has the lowest AIC. It is also noted that the model with 2 inputs gives an AIC lower than that used with just one input. Therefore, a model of 3 inputs will be built with abc1 and higher education variables alongside each of the remaining variables and, the AIC will be calculated.

```
# abc1 all combs
a=c('voteBrexit ~abc1+notBornUK+withHigherEd','voteBrexit ~abc1+medianIncome+withHigherEd','voteBrexit
for (i in 1:length(a)){

  current_model = glm(formula = a[i], family=binomial, data=brexit)
  print(a[i])
  print(AIC(current_model))}
```

```
## [1] "voteBrexit ~abc1+notBornUK+withHigherEd"
## [1] 285.2444
## [1] "voteBrexit ~abc1+medianIncome+withHigherEd"
## [1] 275.9339
## [1] "voteBrexit ~abc1+medianAge+withHigherEd"
## [1] 271.9317
```

For the 3 input models, the model with median age as well as abc1 and with higher education variables, has the lowest AIC of approximately 271.9. Therefore, all the models with 4 variables will have abc1, with higher education and median age, and one of the remaining variables which gives the lowest AIC assuming the AIC is lower than that of the 3 variable input model.

```
a=c('voteBrexit ~abc1+notBornUK+medianAge+withHigherEd','voteBrexit ~abc1+medianIncome+medianAge+withHi
for (i in 1:length(a)){
  current_model = glm(formula = a[i], family=binomial, data=brexit)
  print(a[i])
  print(AIC(current_model))}
```

```
## [1] "voteBrexit ~abc1+notBornUK+medianAge+withHigherEd"
## [1] 269.1141
## [1] "voteBrexit ~abc1+medianIncome+medianAge+withHigherEd"
## [1] 266.9488
```

With 4 inputs, the model that adds median income had the lowest AIC of 266.9 which is also lower than the logistic regression AIC that had 3 inputs.

```
print(AIC(glm(formula = voteBrexit ~abc1+notBornUK+medianIncome+medianAge+withHigherEd, family=binomial
```

```
## [1] 259.3851
```

When all 5 inputs are used, the AIC is at its lowest suggesting the best model is that with 5 inputs.

Under the assumption that the variable with lowest AIC when used in the single input model has the strongest effect on the way a constituency would vote and, that the variable added to model to give the lowest AIC for a 2 input model would have the 2nd strongest affect on the way a constituency would vote, the strength of the variables would be as follows:

1 - withHigherEd 2 - abc1 3 - medianAge 4 - Median Income 5 - Not born in the UK

This is mostly inline with the variables ordered by strongest to weakest effect when using the magnitudes calculated based on coefficients when all 5 inputs were fed into the model. The only difference is that median age is stronger than median income when using the 'greedy' method. This is likely to be due to the overlap in confidence intervals calculated for median income and median age in model with all 5 inputs.

Decison tree

Another classification model that can be used to assess which variables had the strongest effect on the brexit vote would be a decision tree.

The code and decison tree outline can be seen bellow.

```
#Training a decision tree with formula Y ~ X using mydata
library(rpart)
```

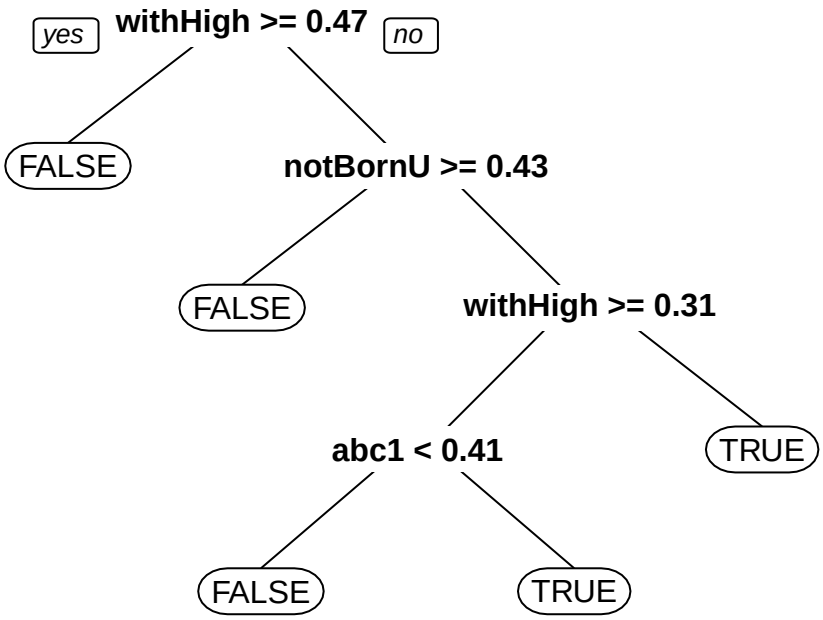
```
## Warning: package 'rpart' was built under R version 4.0.4
```

```
library(rpart.plot)
```

```
## Warning: package 'rpart.plot' was built under R version 4.0.4
```

```
mytree = rpart(voteBrexit ~ ., data=brexit, method='class')

#Visualising a decision tree
# library(rpart.plot)
prp(mytree)
```



The tree splits higher education first with any electoral ward with equal to or greater than 0.47 with higher education variable, being assumed to be remain voters. The tree chooses to split with higher education at 0.47 as this must give the lowest gini impurity value. The decision tree then splits the tree with electoral wards with values less than 0.47 having higher education qualifications with not being born in the UK. If the not born in the UK value greater than or equal to 0.43, this would be classified as remain and vice versa. The gini impurity will be zero at the final split (abc<0.41)

Model selection

It can be assumed that splitting at higher education first and more than once suggests that with higher education has the strongest effect on the way an electoral ward would vote. This is in line with the coefficient magnitudes calculated when all inputs were put into the logistic regression glm and with the order of magnitudes determined by the greedy method.

It may be inferred that the classification tree assumes not being born in the uk and the abc1 variables have the 2nd and 3rd strongest effect on the way an electoral ward would vote but this may be miss leading as not all the variables are in the tree and with higher education appears more than once.

Explanation of results

If explaining the results in a way that non technical as well as technical people could understand such as a news paper article, a classification tree would be the least likely to be used as it is the least clear about how each variable effected the way an electoral ward would vote.

The greedy method takes into account how much each variable effects the way an electoral ward would vote. The greedy method builds a model not containing all the variables available. This suggests that the greedy

method is better at identifying how variables effect the way an electoral ward would vote if the model was independent from the other variables.

However, the coefficients of the logistic regression models with all inputs used quantify the direction and magnitude the way an electoral ward would vote. This allows more inferences to be made about the correlation between that variable and the areas voting.

Conclusion

This report assessed how coefficients when using all 5 inputs, the greedy method and classification trees could be used to assess the strength of the 5 variables affected the brexit vote. From the 3 evaluation methods used, a clear conclusion that having a higher education degree not only had the greatest effect on the way an electoral ward would vote, it also was very strongly correlated with voting to remain. This makes a lot of sense seeing as how much higher education involves international collaboration. People immersed to higher education not only have the chance to meet europeon people that come to study at british universities but also the chance to see the great positive impacts their europeoen counterparts provide for Britain. Positive impacts such as the erasmus scheme which sees british students go off to study in europeon countries with funding from the europeon union. Positives such as top academics from euopeoen countries coming to teach them at their higher education institute in britain. positives such as the funding the EU provides for the countless research projects at british universities.

Abc1 may also be concluded to be the 2nd strongest variable in affecting the way an electoral ward would vote. This is as both the coefficient method and greedy method found that to be the case. This also makes sense as social class, like most countries, would divide opinions strongly. People from more middle class backgrounds would be assumed to want britain to not be overly influenced by other culture and 'keep its sovereignty'. This idea is reinforced by the coefficient of the model built using all inputs being heavily positive showing middle class/upper class dense electoral wards were more likely to vote leave.

The model that orders the strength of each variable with regards to the way an electoral ward would vote has been assumed to be the one built using greedy method.

Appendix

knitr::opts_chunk$set(echo = TRUE, fig.align="center")

knitr::include_graphics() ## Introduction

In 2016, the United Kingdom voted to leave the EU in one the of the greatest upsets in 21st century politics. With vote remain being the heavy favorite, many people have questioned the reasons for the great upset. This report will look at data obtained from the guardian website with the 5 following variables across all the electoral wards within the Uk:

abc1 - The ratio of individuals who were upper to middle class in that electoral

medianIncome - the median income of all the residents within that electoral ward

medianAge - the median of age of residents in that electoral ward

WithHigherEd - The proportion of residents with university level education in that electoral ward

notBornUk - The proportion of residents not born in the Uk in that electoral ward

All the variables have been normalized so that they can be assessed against each other fairly without scaling bias in any model to be built.

The effect each of these variables had on the way an electoral ward would vote will be analyzed using logistic regression and assessed against the results found on the Guardian page.

The variables will also be assessed against each other with the aim of understanding which variables had the strongest effect and which had the least. Logistic regression coefficients, 'greedy' logistic regression building and decision tree models will all be built and compared to understand the effect of each variable.

#importing the gurdain data brexit=read.csv2('C:/Users/Ahmed/Desktop/Statistical Learning/Practical assessment 2/brexit.csv', header = TRUE,sep = ",", quote = "",dec = ".")

Logitic regression - all 5 inputs

The first model built will be the logistic regression with all 5 variables as inputs.

The logistic regression takes normal linear regression equation as shown bellow where there are 5 variables labeled x1 to x5 and beta 1 to beta 5 representing the variables in this data set (abc1, median income etc)

$$Z = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 \quad (eq\ 1)$$

This linear regression input labeled z is fed into a sigmoid function shown in equation 2 bellow

$$p = \frac{1}{1 + \exp(-z)} \quad (eq\ 2)$$

$$p = \frac{1}{1 + \exp(-(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5))} \quad (eq\ 3)$$

In order to build the logistic regression, a glm function with a family input equal to binomial will be used. The coefficients will be solved numerically by the glm function used bellow

myglm = glm(voteBrexit ~ ., family=binomial, data=brexit)

summary(myglm)

From the logistic regression model built by the glm, the following coefficient (beta) is found.

	Estimate
(Intercept)	-0.1386
abc1	17.5780
notBornUK	5.6861
medianIncome	-6.3857
medianAge	5.9209
withHigherEd	-26.7443

The coefficients show how strongly a variable effects an electoral ward to vote remain or leave in the brexit vote in 2016. The more negative the coefficient, the more this variable pushes to suggest this electoral ward would vote remain and, the more positive a coefficient, the more this variable pushes to suggest this electoral ward is in favor of leave. The closer a coefficient is to 0, the weaker the variables impact on the way a constituency would vote.

The coefficient for percentage of the electoral ward with a higher education degree is -26.7 which is very strongly negative compared to the other coefficients. This shows that the model would associate having a higher education degree with being heavily likely to vote remain. This is in accordance with the figure on the guardian website labeled figure 1. In figure 1 from the guardian, as percentage of residents with higher education decreases, the number of constituency to voting to leave increases in a very strong negative correlation manner.

The coefficient for median annual income is -6.4 which is slightly negative and no where near as negative as the coefficient for higher education. This shows that the model associates having a higher median income

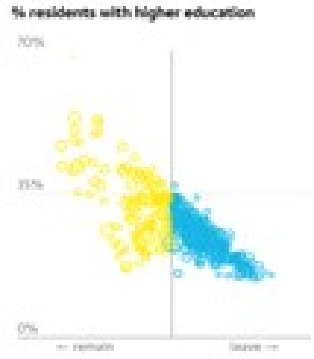


Figure 6: higher education

with being slightly more likely to vote remain but not at no where near the magnitude it does having higher education qualification. This is in concordance with the guardian website figure labeled figure 2 which shows a weak negative correlation between median annual income of constituency and percentage of residents voting leave.

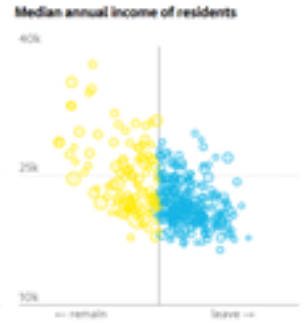


Figure 7: Median annual income

The coefficient for ABC1 is approximately 17.6 which is strongly positive suggesting that ABC1 has a strong positive correlation with voting leave. This goes against the guardians figure in figure 3 which shows negative correlation between abc1 and boting to leave. This negative correlation can be seen by an increase in vote leave voters as the percentage of residents of ABC1 social grade decreases. The guardian figure data is however very sparse for vote remain and very dense on the vote remain side. It is likely the case that the number of vote leave electoral ward s with around 50% to 60% are very dense, swaying the an increase of abc1 to cause an increase the vote leave campagin vote per electoral ward.

The coefficient for median age of residents is 5.9 which is a small positive value. A small positive value suggests that median age has a slight positive correlation with voting leave where as the median age increases, the likelihood of voting leave also increases. The low positive value of the coefficient suggests that median age does not have an incredibly strong affect on a constituency voting leave, especially compared to abc1. This is in line with the guardian website figure labeled figure 4, which shows median age not having much correlation between median age and voting choice but does show a very slight trend in more electoral ward s with median age over 40 being more inclined to vote leave.

The coefficient for residents born outside of the uk is 5.7 suggesting that being having a greater percentage of residents born outside the Uk would lead to slightly greater percentage of residents in that electoral ward voting leave. This goes against the figure on the guardian website labeled figure 5 which shows not much difference in voting for electoral ward s with less than 30% residents not born in the uk but electoral ward s with high percentage of residents not born in the being more inclined to vote remain. This again is likely to be due to the density of data points on the vote leave side in the guardian figure bellow being more

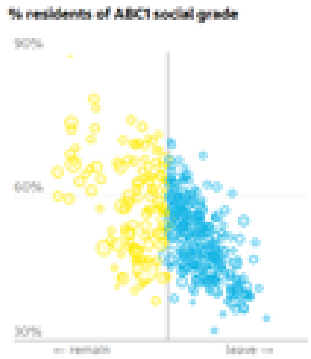


Figure 8: abc1 social group

Figure 9: Median age

abundant/densely packed than can be perceived through this figures.

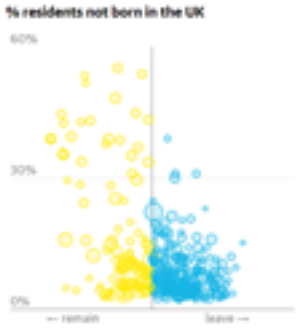


Figure 10: born outside of the uk

Variable strength according to coeffcient

The coefficients with the highest magnitudes in the positive or negative direction will have the strongest effect. The largest magnitude in the logistic regression glm model is the with higher education variable, followed by abc1 and so on, with not being born in the uk having the weakest effect on the way in which an electoral ward would vote.

Coefficient effect from strongest to weakest based solely on coefficient. Estimate withHigherEd -26.7443 abc1 17.5780 medianIncome -6.3857 medianAge 5.9209 notBornUK 5.6861

However, the coefficient come with uncertainty. If uncertainty is high, there is a possibility the true affect a variable and its coefficient would have on the way an electoral ward would vote may differ quite a bit. Furthermore, overlapping confidence intervals would have an affect on ordering the inputs based on strongest affect as it may be unclear where the true value lies.

In order to quantify uncertainty, the 95% confidence interval is taken. Within the calculated confidence interval for each coefficient, the true value of the coefficient is likely to be within that range with a 5% chance of a type 1 error. A type 1 error being when a true null hypothesis is rejected.

The confidence interval for each variable is shown in the output of the following code.

zc = qnorm(0.975) #Extract estimate and standard error of coefficient from model summary (check above) estimate = summary(myglm)$coefficients[2,1]$standard_error = summary(myglm)coefficients[2,2] #Calculate and print the lower and upper boundaries of the confidence interval CI_min = estimate - zc*standard_error; CI_max = estimate + zc*standard_error paste('abc1 Confidence interval:',CI_min,',' ,CI_max)

zc = qnorm(0.975) #Extract estimate and standard error of coefficient from model summary (check above) estimate = summary(myglm)$coefficients[3,1]$standard_error = summary(myglm)coefficients[3,2] #Calculate and print the lower and upper boundaries of the confidence interval CI_min = estimate - zc*standard_error; CI_max = estimate + zc*standard_error paste('Not born in the UK Confidence interval:',CI_min,',' ,CI_max)

zc = qnorm(0.975) #Extract estimate and standard error of coefficient from model summary (check above) estimate = summary(myglm)$coefficients[4,1]$standard_error = summary(myglm)coefficients[4,2] #Calculate and print the lower and upper boundaries of the confidence interval CI_min = estimate - zc*standard_error; CI_max = estimate + zc*standard_error paste('Median income Confidence interval:',CI_min,',' ,CI_max)

zc = qnorm(0.975) #Extract estimate and standard error of coefficient from model summary (check above) estimate = summary(myglm)$coefficients[5,1]$standard_error = summary(myglm)coefficients[5,2] #Calculate and print the lower and upper boundaries of the confidence interval CI_min = estimate - zc*standard_error; CI_max = estimate + zc*standard_error paste('Median age Confidence interval:',CI_min,',' ,CI_max)

zc = qnorm(0.975) #Extract estimate and standard error of coefficient from model summary (check above) estimate = summary(myglm)$coefficients[6,1]$standard_error = summary(myglm)coefficients[6,2] #Calculate and print the lower and upper boundaries of the confidence interval CI_min = estimate - zc*standard_error; CI_max = estimate + zc*standard_error paste('With Higher Education qualification Confidence interval:',CI_min,',' ,CI_max)

Abc1 has 23 within its confidence interval and the 'with higher education' has -19 within its confidence interval. The magnitude of the values within these confidence intervals overlap meaning there is a possibility that abc1 could have a greater affect on the way an electoral ward would vote and thus the greatest affect on the way an electoral ward would vote. This leads to the feasibility being questioned of with higher education having greater effect on the way electoral ward s would vote.

Median income and abc1 have no overlap in magnitudes within their own confidence intervals so there is at most a 2.5% chance of both values having type 1 errors where median income and abc1 coefficients are outside of the confidence interval giving chance for median income to be greater than the abc1 coefficient. The calculation of the probability of having median income of and abc1 both being type 1 errors can be seen bellow.

$$(type\ 1\ error\ abc1\ 5\%) * (type\ 1\ error\ median\ income\ 5\%) \ = \ 2.5\%$$

Median income, median age and not being born Median income, median age and not being born in the uk have very large overlaps in magnitude of confidence intervals. This suggests that the order of magnitude defined based solely on coefficients out of the logistic regression glm model may be prone to inaccuracy for the order of magnitude of these 3 variables.

Greedy method

Another way to test the strength of the 5 variables would be to build a model gradually, only adding the gives the greatest decrease in AIC the logistic regression glm model. AIC is used as it penalizes too many parameters being used in the model. Too many unnecessary parameters is something to be avoided as too many parameters could lead to overfitting of the regression model. If the model overfits, it will struggle to generalize well when using unseen data to make predictions

The code and output AIC of the logistic regression model with one input is as follows:

one input

a=c('voteBrexit ~abc1','voteBrexit ~medianIncome','voteBrexit ~withHigherEd','voteBrexit ~median-Age','voteBrexit ~medianIncome') for (i in 1:length(a)){

current_model = glm(formula = a[i], family=binomial, data=brexit) print(a[i]) print(AIC(current_model))}

With one input, the logistic regression model built using the glm function shows the model built with just the with higher education variable has the lowest AIC of approximately 313.6

Implementing the greedy method, models with 2 inputs will all contain the wither higher education variable. The code for building a glm logistic regression with 2 variables, one of which being with higher education is as follows:

abc 2 input

a=c('voteBrexit ~abc1+withHigherEd') for (i in 1:length(a)){

current_model = glm(formula = a[i], family=binomial, data=brexit) print(a[i]) print(AIC(current_model))}

2inputs median income

a=c('voteBrexit ~medianIncome+withHigherEd') for (i in 1:length(a)){

current_model = glm(formula = a[i], family=binomial, data=brexit) print(a[i]) print(AIC(current_model))}

a=c('voteBrexit ~notBornUK+withHigherEd') for (i in 1:length(a)){

current_model = glm(formula = a[i], family=binomial, data=brexit) print(a[i]) print(AIC(current_model))}

2 inputs median age

a=c('voteBrexit ~medianAge+withHigherEd') for (i in 1:length(a)){

current_model = glm(formula = a[i], family=binomial, data=brexit) print(a[i]) print(AIC(current_model))}

From the 2 input models that all must involve the with higher education variable, the model with abc1 added has the lowest AIC. It is also noted that the model with 2 inputs gives an AIC lower than that used with just one input. Therefore, a model of 3 inputs will be built with abc1 and higher education variables alongside each of the remaining variables and, the AIC will be calculated.

abc1 all combs

a=c('voteBrexit ~abc1+notBornUK+withHigherEd','voteBrexit ~abc1+medianIncome+withHigherEd','voteBrexit ~abc1+medianAge+withHigherEd') for (i in 1:length(a)){

current_model = glm(formula = a[i], family=binomial, data=brexit) print(a[i]) print(AIC(current_model))}

For the 3 input models, the model with median age as well as abc1 and with higher education variables, has the lowest AIC of approximately 271.9. Therefore, all the models with 4 variables will have abc1, with higher education and median age, and one of the remaining variables which gives the lowest AIC assuming the AIC is lower than that of the 3 variable input model.

a=c('voteBrexit ~abc1+notBornUK+medianAge+withHigherEd','voteBrexit ~abc1+medianIncome+medianAge+withHighe for (i in 1:length(a)){ current_model = glm(formula = a[i], family=binomial, data=brexit) print(a[i]) print(AIC(current_model))}

With 4 inputs, the model that adds median income had the lowest AIC of 266.9 which is also lower than the logistic regression AIC that had 3 inputs.

print(AIC(glm(formula = voteBrexit ~abc1+notBornUK+medianIncome+medianAge+withHigherEd, family=binomial, data=brexit)))

When all 5 inputs are used, the AIC is at its lowest suggesting the best model is that with 5 inputs.

Under the assumption that the variable with lowest AIC when used in the single input model has the strongest effect on the way a constituency would vote and, that the variable added to model to give the lowest AIC for a 2 input model would have the 2nd strongest affect on the way a constituency would vote, the strength of the variables would be as follows:

1 - withHigherEd 2 - abc1 3 - medianAge 4 - Median Income 5 - Not born in the UK

This is mostly inline with the variables ordered by strongest to weakest effect when using the magnitudes calculated based on coefficients when all 5 inputs were fed into the model. The only difference is that median age is stronger than median income when using the 'greedy' method. This is likely to be due to the overlap in confidence intervals calculated for median income and median age in model with all 5 inputs.

Decison tree

Another classification model that can be used to assess which variables had the strongest effect on the brexit vote would be a decision tree.

The code and decison tree outline can be seen bellow.

#Training a decision tree with formula Y ~ X using mydata library(rpart) library(rpart.plot) mytree = rpart(voteBrexit ~ ., data=brexit, method='class')

#Visualising a decision tree # library(rpart.plot) prp(mytree)

The tree splits higher education first with any electoral ward with equal to or greater than 0.47 with higher education variable, being assumed to be remain voters. The tree chooses to split with higher education at 0.47 as this must give the lowest gini impurity value. The decision tree then splits the tree with electoral wards with values less than 0.47 having higher education qualifications with not being born in the UK. If the not born in the UK value greater than or equal to 0.43, this would be classified as remain and vice versa. The gini impurity will be zero at the final split (abc<0.41)

Model selection

It can be assumed that splitting at higher education first and more than once suggests that with higher education has the strongest effect on the way an electoral ward would vote. This is in line with the coefficient

magnitudes calculated when all inputs were put into the logistic regression glm and with the order of magnitudes determined by the greedy method.

It may be inferred that the classification tree assumes not being born in the uk and the abc1 variables have the 2nd and 3rd strongest effect on the way an electoral ward would vote but this may be miss leading as not all the variables are in the tree and with higher education appears more than once.

Explanation of results

If explaining the results in a way that non technical as well as technical people could understand such as a news paper article, a classification tree would be the least likely to be used as it is the least clear about how each variable effected the way an electoral ward would vote.

The greedy method takes into account how much each variable effects the way an electoral ward would vote. The greedy method builds a model not containing all the variables available. This suggests that the greedy method is better at identifying how variables effect the way an electoral ward would vote if the model was independent from the other variables.

However, the coefficients of the logistic regression models with all inputs used quantify the direction and magnitude the way an electoral ward would vote. This allows more inferences to be made about the correlation between that variable and the areas voting.

Conclusion

This report assessed how coefficients when using all 5 inputs, the greedy method and classification trees could be used to assess the strength of the 5 variables affected the brexit vote. From the 3 evaluation methods used, a clear conclusion that having a higher education degree not only had the greatest effect on the way an electoral ward would vote, it also was very strongly correlated with voting to remain. This makes a lot of sense seeing as how much higher education involves international collaboration. People immersed to higher education not only have the chance to meet europeon people that come to study at british universities but also the chance to see the great positive impacts their europeoen counterparts provide for Britain. Positive impacts such as the erasmus scheme which sees british students go off to study in europeon countries with funding from the europeon union. Positives such as top academics from euopeoen countries coming to teach them at their higher education institute in britain. positives such as the funding the EU provides for the countless research projects at british universities.

Abc1 may also be concluded to be the 2nd strongest variable in affecting the way an electoral ward would vote. This is as both the coefficient method and greedy method found that to be the case. This also makes sense as social class, like most countries, would divide opinions strongly. People from more middle class backgrounds would be assumed to want britain to not be overly influenced by other culture and 'keep its sovereignty'. This idea is reinforced by the coefficient of the model built using all inputs being heavily positive showing middle class/upper class dense electoral wards were more likely to vote leave.

The model that orders the strength of each variable with regards to the way an electoral ward would vote has been assumed to be the one built using greedy method.