United Kingdom National Census Data

Stat 471 Final Project

By Kit Wiggin

https://github.com/kitwiggin/uk-census-health

1. Executive Summary

This analysis explores data from the 2011 Census of the UK. It considers a set of categorical variables representing an individual's survey responses and fits models that attempt to classify people as healthy or unhealthy, trained on this self-reported metric. The goal is to identify attributes that provide a notable increase in classification accuracy when compared to predictions based primarily on Age, which is intuitively a very significant indicator for this response variable. As well as this, the analysis stipulates that any features that achieve this must not have some direct link to age that might allow them to behave as some kind of proxy.

After some cleaning, the data consists of 457,123 people characterised by 15 socio-demographic, categorical features and a binary response variable representing healthiness. Although there are no null entries, some attributes have an additional level used when some question doesn't apply to someone. For example, a 3 year old would (probably) use this level for a question regarding their current employment.

The analysis fits one model using just the age feature as well as six more using the entire feature set. This list includes a multivariate logistic regression, ridge and lasso-penalised regressions, a standard tree model, random forest, and a tree-based boosting model. Following this, it considers the differences in misclassification error

between the models in order to examine the most significant features of the best performing models and how they relate to age.

Due to the dataset being inherently more liable to bias than variance, the regression penalisation models don't perform nearly as well as the trees, where the boosting method just beats the random forest to claim the top spot. Looking at the most important features in both of these models, four appear to stand out and of these, only one seems to demonstrate a clear risk of proxying the effect of age. Therefore, three features are highlighted as improving prediction accuracy, on top of and independently of age.

2. Introduction

This paper will explore data taken from the National Census in the UK in 2011. This census takes place every 10 years and aims to get a comprehensive set of information on each citizen or resident of the country. Using this information gathered in 2011, I will be considering the effect of 16 socio-demographic features on an individual's health. Specifically, attempting to use these variables to classify them in a binary manner as 'healthy' or not. For additional context, I am from the UK and am therefore included somewhere in this dataset, which is fun.

The goal of this analysis is to try and find out if there exist features, either demographic or social, that can be of significance to a model's predictive capacity despite the presence of the Age factor. That is, I am making the underlying assumption that Age will be a very strong indicator of someone's health and want to find out if there exists information in the remaining features that can improve on the accuracy of a model that relies on Age alone. There are two important issues that will stop me from considering a new feature to be interesting. Firstly, there might exist features that provide some predictive value on their own but their impact is minimised due to the overruling influence that Age has on the model, so the improvement that comes from including them is miniscule. Secondly, many features might be indirectly linked to age and appear valuable in the summary of a model but in actual fact are just riding the age-wave and don't represent any genuinely unique supplemental information. Therefore, the goal of my analysis will be to find the models that

provide the lowest misclassification rates and then examine the features of most significance in these models to work out if they pass the above criteria.

The significance of this analysis is surface level in that increasing the prediction accuracy of machine learning models for something as important as health is a society-wide goal. Specifically, being able to look beyond a feature like age and find much less obvious correlations with poor health can result in more finely tuned characterizations of a population. That is, discovering groups of people prone to poor health that are smaller and more definitively identified can lead to easier and better monitoring which could improve quality of care and even encourage pre-emptive medical action or lifestyle changes. On top of this, using a wide range of unrelated features to try and predict something as crucial as health can lead to the almost accidental discovery of unexpected, unintuitive and potentially frightening relationships. For example, if ethnic group or religion turned out to be a good predictor of health, it would be a strong indication of the existence of institional discrimination and inherent social inequality. This is a benefit from doing this analysis on such a broad and therefore random data set as it minimises the potential for biases that might otherwise be able to weakly explain away such findings.

3. Data

3.1 Data Source

This data was taken from the UK Government's website[1]. However, the data does not contain entries for all 65 million people in the UK. This dataset is published especially for projects like these and contains anonymised data for 1% of all the respondents to the Census. As well as standard information such as data and variable descriptions, since it is itself a subsample of another dataset, there is an interesting page that goes through each category of each feature and highlights any difference in its percentage representation between the teaching data and the whole data.

[1] Microdata Teaching File information and download page
https://www.ons.gov.uk/census/2011census/2011censusdata/censusmicrodata/microdatateachingfile

There are pros and cons about the format of the dataset. On the plus side, it's individualised perfectly and every row represents exactly one unique person from the UK. As well as this, there are no null values since everybody is obligated to fill out the form, even if it means checking an 'Other' or 'Unknown' category. Lastly, since it just represents a questionnaire where people have to choose a category in which to place themselves for every question, the data is inherently clustered into perfect factor variables. However, this self-identified clustering also creates some issues. For one, no question has more than 12 possible options and most non-binary questions have between 5 and 10. Therefore, on a sample size as large as the population of the UK, there are a large number of people who don't fit into any of the options and end up as an entry in the 'Unknown' category of that question. Even worse, many of the questions in the very way that they are designed do not apply to significant portions of the population such as children or short-term renters. On top of that, the questions are also poorly designed in that they are sometimes not independent or even gather any additional information that isn't available from other answers.

3.1.2 Data Download

Frustratingly, it was very difficult to make a script to download the file and eventually I resigned to doing it manually. Here are the instructions on how to do so (for Mac):

1. Click the following link: Download the Teaching File
2. The file "*rft-teaching-file*" should be downloaded and automatically unzipped - open it
3. There should be four files, one of which is called "*2011 Census Microdata Teaching File.csv*", rename this file to "*census_data_raw.csv*".
4. Move this file from the "*rft-teaching-file*" directory in Downloads to the directory storing this project which I have called "*uk-census-health*". Specifically, move it into the otherwise empty raw data directory, *"uk-census-health/data/raw"*.
5. Set "*uk-census-health*" as your working directory and run all scripts normally to replicate the analysis.

3.2 Data Description

The raw data is being described before the cleaning as the process is somewhat nuanced and requires an understanding of the feature set in order to be well understood. Features that undergo significant cleaning will be highlighted with asterisks.

The data contains 569,742 entries representing 1% of the individual responses to the census survey in 2011. There are 18 columns containing 16 categorical variables, one unique person ID for each entry and one categorical response variable, Health.

Here is a list of the features in the dataset along with the number of categories if it's categorical, a brief description of what the feature represents and a link to a table in the appendix that describes the translation between the coded value used in the dataset (a number) and what that code with regard to the feature. As well as this, each feature bullet will state whether or not it has an *"other"* option as well as what it represents.

3.2.1 The "*other*" option

Here's an attempt to clarify what the *"other"* option is and how it's been used in the data. About half the categorical variables have an extra level represented by a "-9". Then, in the table for that feature, there will be a vague explanation about the groups of people that are populating that category. Sometimes it's specific and other times the "-9" entry is used so many times that it's very clearly also being used as a catch-all for a number of other troublesome responses to that question. Lastly, during the data cleaning I converted the "-9" entry to be a "0" in order to more closely line up with the positive integer characterizations of the other variables. Therefore, in the variable bullets this additional option is counted in the levels and is referred to as the 0-level.

3.2.2 Feature Set Description

Variables:

- Person ID
    - Unique Reference ID for each entry

    - *Deleted in cleaning*
- Region
    - Categorical, 10 levels
    - 9 digit codes representing a geographical area in the UK
    - No 0-level
    - Appendix 6.1.1
- Residence Type
    - Categorical, 2 levels
    - Whether or not someone lives in communal housing
    - Appendix 6.1.2
- Family Composition
    - Categorical, 7 levels
    - Characteristics of family situation
    - 0-level: (Resident of a communal establishment, students or schoolchildren living away during term-time, or a short-term resident)
    - Appendix 6.1.3
- Population Base
    - Categorical, 3 levels
    - Residential situation
    - No 0-level
    - Appendix 6.1.4
- Sex
    - Categorical, 1 levels
    - Male or Female
    - No 0-level
    - Appendix 6.1.5
- Age
    - Categorical, 8 levels
    - Age brackets, mostly 10 years wide
    - No 0-level

    - Appendix 6.1.6
- Marital Status
    - Categorical, 5 levels
    - Marital Situation
    - No O-level
    - Appendix 6.1.7
- Student
    - Categorical, 2 levels
    - Whether or not they are a student
    - No O-level
    - Appendix 6.1.8
- Country of Birth
    - Categorical, 3 levels
    - UK, not UK, other
    - O-level: (Students or schoolchildren living away during term-time)
    - Appendix 6.1.9
- Ethnic Group
    - Categorical, 6 levels
    - Ethnic Identity
    - O-level: (Not resident in England or Wales, students or schoolchildren living away during term-time)
    - Appendix 6.1.10
- Religion
    - Categorical, 10 levels
    - Religious Identity
    - O-level: (Not resident in England or Wales, students or schoolchildren living away during term-time)
    - Appendix 6.1.11
- Economic Activity
    - Categorical, 10 levels

    - Whether or not respondent is economically active and why or why not
    - 0-level: (Aged under 16 or students or schoolchildren living away during term-time)
    - Appendix 6.1.12
- Occupation
    - Categorical, 10 levels
    - Brackets of skill and seniority to categorise employment
    - 0-level: (People aged under 16, people who have never worked and students or schoolchildren living away during term-time)
    - Appendix 6.1.13
    - *Deleted in cleaning*
- Industry
    - Categorical, 13 levels
    - Industry of employment
    - 0-level: (People aged under 16, people who have never worked, and students or schoolchildren living away during term-time)
    - Appendix 6.1.14
- Hours worked per week
    - Categorical, 5 levels
    - Brackets of weekly hours to categorise time spent working
    - 0-level: (People aged under 16, people not working, and students or schoolchildren living away during term-time)
    - Appendix 6.1.15
- Approximated social grade
    - Categorical, 5 levels
    - Official encoding used to identify social grade
    - 0-level: (People aged under 16, people resident in communal establishments, and students or schoolchildren living away during term-time)
    - Appendix 6.1.16
- Health

- The response variable!
- Categorical, 6 levels
- Self-assessed rating of general health level from 1 to 5
- 0-level: (Students or schoolchildren living away during term-time)
- Appendix 6.1.17

3.3 Data Cleaning

This will be a list of any and all alterations made to the raw data set before it was labelled as clean. Each item will contain a brief description as well as an explanation if one is necessary.

1. Manually set first row as column titles and removed each entry's Person ID (569,741 x 17)
2. Set all feature columns to be factors and recoded the factors to be more descriptive when it was simple in order to minimise time spent cross-checking with the variables list. E.g. set Sex to be "Male" or "Female" rather than "1" or "2".
3. Removed a group of 6804 entries with many 0-level entries. Looking at the variable list, it was immediately clear that there were patterns among the groups that were putting down "-9" for different questions. For the sake of the analysis, I wanted to try and remove bad entries that had these uninteresting values across multiple, independent features. That is, try to identify and delete people that fit into groups that contributed no value to the data and only risked weakening potential feature relationships with the response variable. As well as looking at the 0-level descriptions in the variable list, I counted up how many entries there were at this level for each feature (Appendix 6.2). The number 6804 appeared for Health, Country of Birth and Ethnic group. At first thought, I thought that perhaps this was the number of entries that corresponded to "students or schoolchildren living away during term time". Unfortunately, only 6730 entries were labelled as being in this category in Population Base. However, with some experimenting some things became clear. Firstly, the 6804 entries were the same culprits for all three variables.

Secondly, the 6730 live-away students were entirely contained within this 6804. Lastly, the 6804 were largely or entirely contained within every other group of 0-level entries for any other feature. Therefore, the whole group was removed from the dataset since it was clear these entries were missing the majority of the data that might be useful for predictions, given they didn't even have informative health entries.

4. Removed Occupation as a feature. This decision was made after researching what Social Grade was and what the letters meant [2] and it felt as though the two features had significant overlap. Between the two, Occupation felt as though it had slightly more overlap with Industry and Economic Activity primarily just because it had more categories. On top of this, my good friend Occam has taught me that if the simpler option does a similarly good job, pick that!
5. (This is some cleaning that I decided to come back and do retroactively once I started exploring the data) Removed entries for anyone younger than 16. This group of individuals caused me to go back and forth a little bit when deciding whether or not to include them. On one hand, this whole group had 0-level entries for all the economic features and therefore not only provided notably less value to the model than the other age groups but simultaneously undermined those economic variables as they were inevitably overpopulated by 0-level entries (Appendix 6.2). On the other hand, there was nothing actually wrong with these entries. They legitimately had multiple 0-level in the economic fields as these are variables that are completely unrelated to them. As a result, the response variable would be built from fewer independent variables and therefore might provide a uniquely clear insight into the value of some component of the model. Therefore, I decided to keep those entries and try to minimize the dampening effect on the economic variables by splitting their 0-level entries into two, separating the under-16s. However, this perspective changed when I created a histogram of health level faceted by age (see 3.5 - Data Exploration) for two reasons. Firstly, I had not yet considered the fact that

[2] Explanation of UK Social Grade codes:
https://www.ukgeographics.co.uk/blog/social-grade-a-b-c1-c2-d-e

this was the largest age category by a significant margin. 19% of the data set was younger than 16 with the next largest age group representing just 13%. Therefore, it's negative impacts would only be magnified. Secondly and importantly, the histogram made it clear that the number of individuals in this age group who weren't in the healthiest two categories was insignificant. Therefore it became apparent that if I kept this group in the dataset then Age would be the only real determinant of healthiness. Any potential insight into the effect of other factors would be overshadowed by the overwhelming presence of perfectly healthy children randomly distributed across the remaining categories. Therefore, the trajectory of this analysis changed to only include people of working age and above. (See appendix 6.3 for the plot that brought me to this decision - note that it is no longer produced by my code and has been included to explain this retroactive decision.)

6. Last but not least, I converted the response variable from a healthiness index on a five point scale to a binary, 'healthy or unhealthy' indicator. My primary reason for this decision was a concern with the seemingly arbitrary nature of the self-assessed health report. There is no way of ensuring any consistency between different people and can't ever quantify a difference between two points on the scale. Therefore, generalising it as much as possible minimizes these potential negative impacts by lowering the possibility of variation between samples.

Therefore, after the data has been cleaned, there are 457, 123 entries with 16 variables including the response variable. Only four features: Social Grade, Family Comp, Hours/week and Industry still have 0-level entries.

3.4 Data Allocation

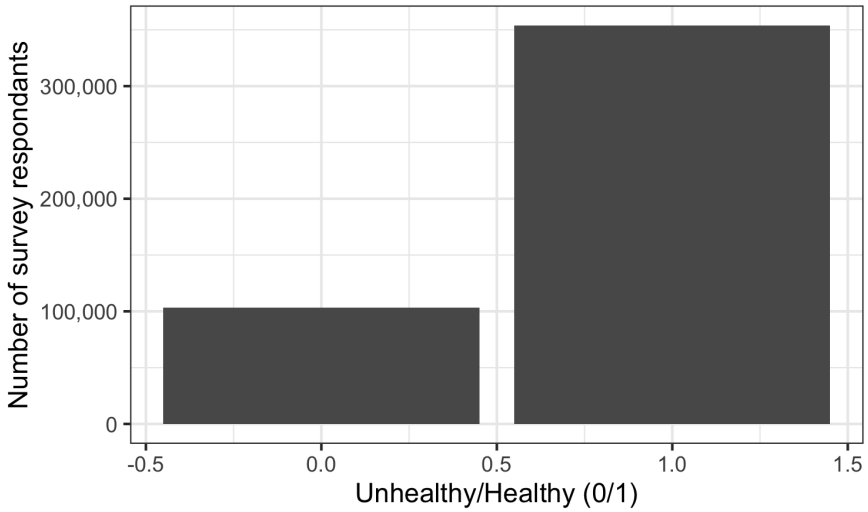
Once the data was fully cleaned and ready to use, the data was split randomly such that 80% of the entries were used in the training set and the remaining 20% were kept aside in the test set. That is, there were 365,698 observations in the training data and 91,425 in the test data. The only additional detail is that for training the random forest and for boosting, the training set was shrunk to 20% and 50% of its original size respectively. This was as a

result of memory constraints on my laptop. The error of these models was still calculated using the same test set.

3.5 Data Exploration

3.5.1 Response Variable

Firstly, let's look at the distribution of the response variable across the entire data set.



There are 353,968 healthy survey participants and 103,155 unhealthy ones. The classes are imbalanced but at a roughly 2:7 ratio addressing potential issues caused by model prioritization of the majority class is not critical to the success of the model.

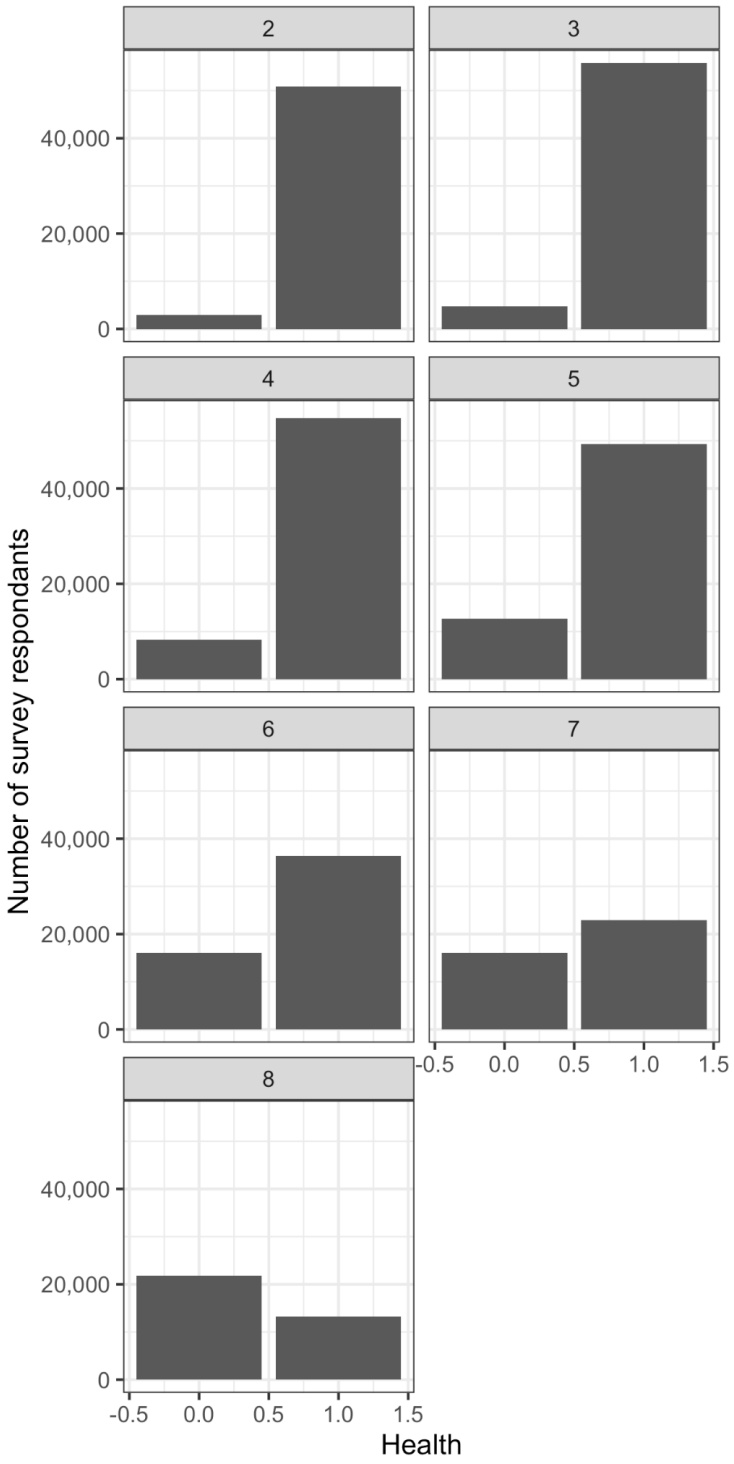
3.5.2 Features

Since all the features are categorical, the extent to which collinearity can be explored is limited. There have been considerations for potential interdependencies among the features such as deleting Occupation due to its definitional overlap with Social Grade. Although some overlap might still exist between certain features and the significance they might hold in a model, I don't believe there is sufficient evidence to make any other significant alterations in this regard. I made histograms for each of the features in order to

examine their categorical distributions and look for any signs of bad data but found nothing of particular interest.

3.5.3 Health Split by Age Category

To round out the data exploration I thought it was important to have a look over how the balance of healthy and unhealthy changes with age. As previously mentioned, I am assuming that there is a strong correlation between increasing age and an increasing proportion of unhealthy subjects. However, I was interested to see what this relationship looked like (see next page). I like this visualization because even though it's not the clearest representation of a changing category balance, we get all the information we need whilst also getting an insight into the change in the total size of the age category. The changing balance aligns with my expectations with the size of the incremental shift between consecutive age groups increasing as the groups age. However, something I found especially interesting to consider was how deaths would influence the perceived proportion of healthy people in an age group. That is, since the deceased were not asked to submit Census information, they are absent from the data, despite being definitively 'unhealthy'. The best demonstration of this effect is in category 7. Although it appears that the majority is still healthy, upon closer inspection one can see that the actual number of healthy people is just 20,000 or so. This is notably less than half of 60,000 which is roughly the size of the 5 younger age groups. This 'death effect', for want of a better term, is something that I will not actively investigate but that might be interesting for future analysis in this area.


2
3
4
5
6
7
8
40,000
20,000
0
Number of survey respondants
-0.5
0.0
0.5
1.0
1.5
Health

4. Modelling

4.1 Regression Techniques

4.1.1 Univariate Logistic Regression on Age

This model was run just to confirm and quantify the assumption that has been repeatedly made up to this point that age is going to be a strong predictor of healthiness. The model indeed confirmed that age was a highly significant predictor.

We run a logistic regression since we are trying to predict a categorical outcome.

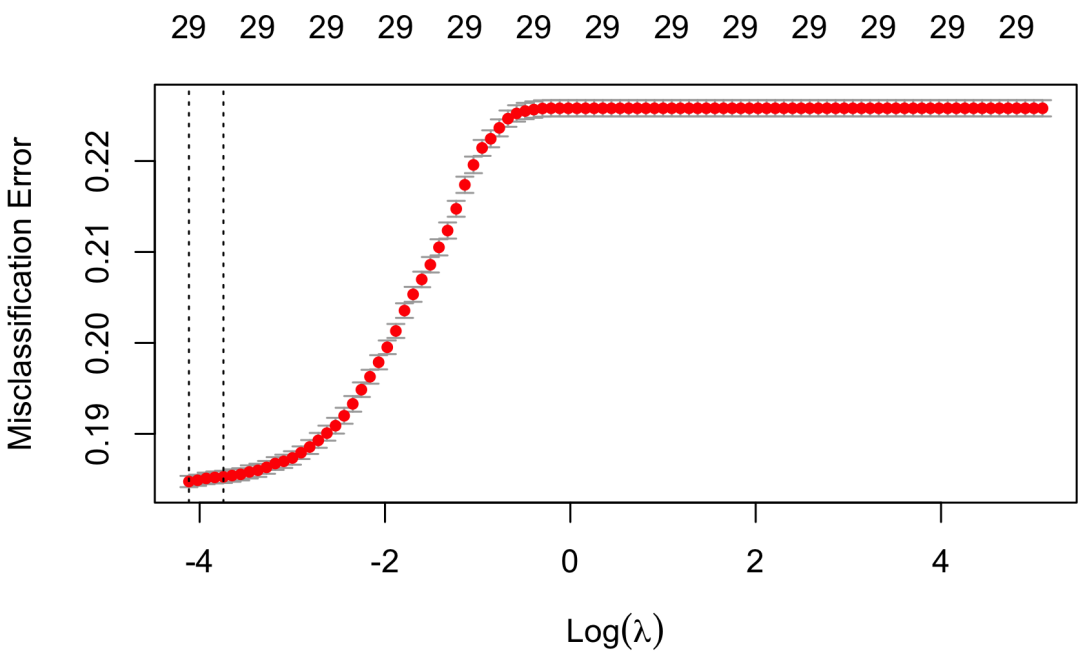
4.1.2 Multivariate Logistic Regression

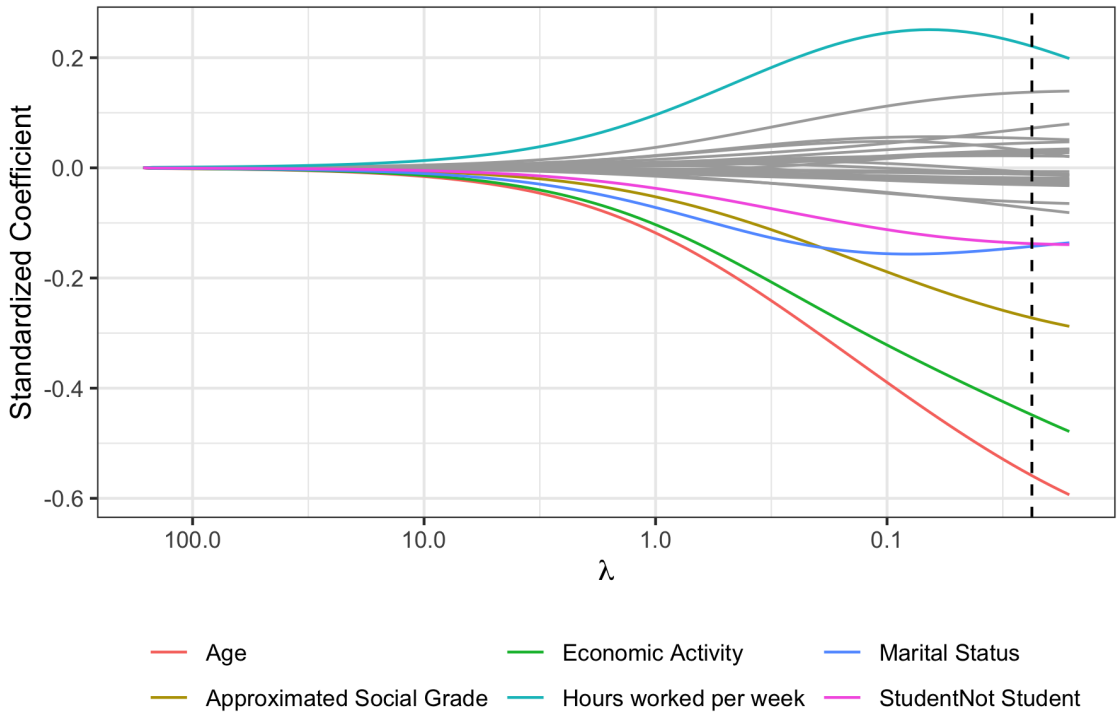
This is the first model using all 16 features to find a best fit for Age. Nothing seemed to explode which is nice.

Multiple logistic regression is to categorical response variables what Ordinary Least Squares regression is for continuous ones. It's a basic regression with no penalisation that attempts to find optimum intercept and feature coefficient values according to a standard, logarithmic formula.

4.1.3 Ridge Penalised Logistic Regression

This is the point at which things start to get a little bit more interesting! The ridge penalisation process was run with 10 folds for cross validation purposes. Consider the resulting CV plot and 6-feature trace plot.

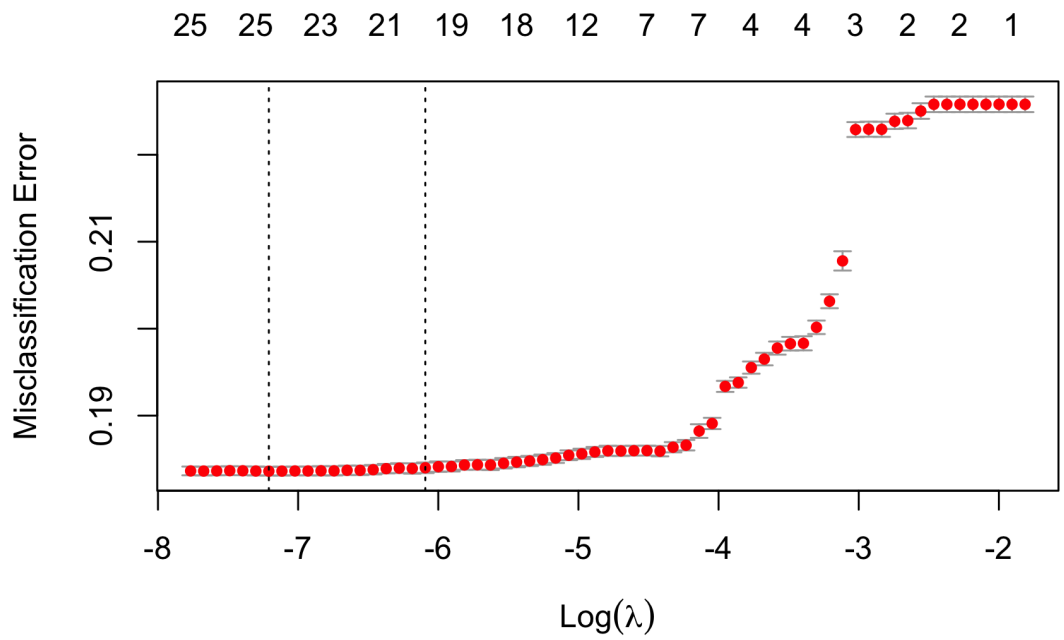

29 29 29 29 29 29 29 29 29 29 29 29 29
Misclassification Error
0.19 0.20 0.21 0.22
-4 -2 0 2 4
Log(λ)

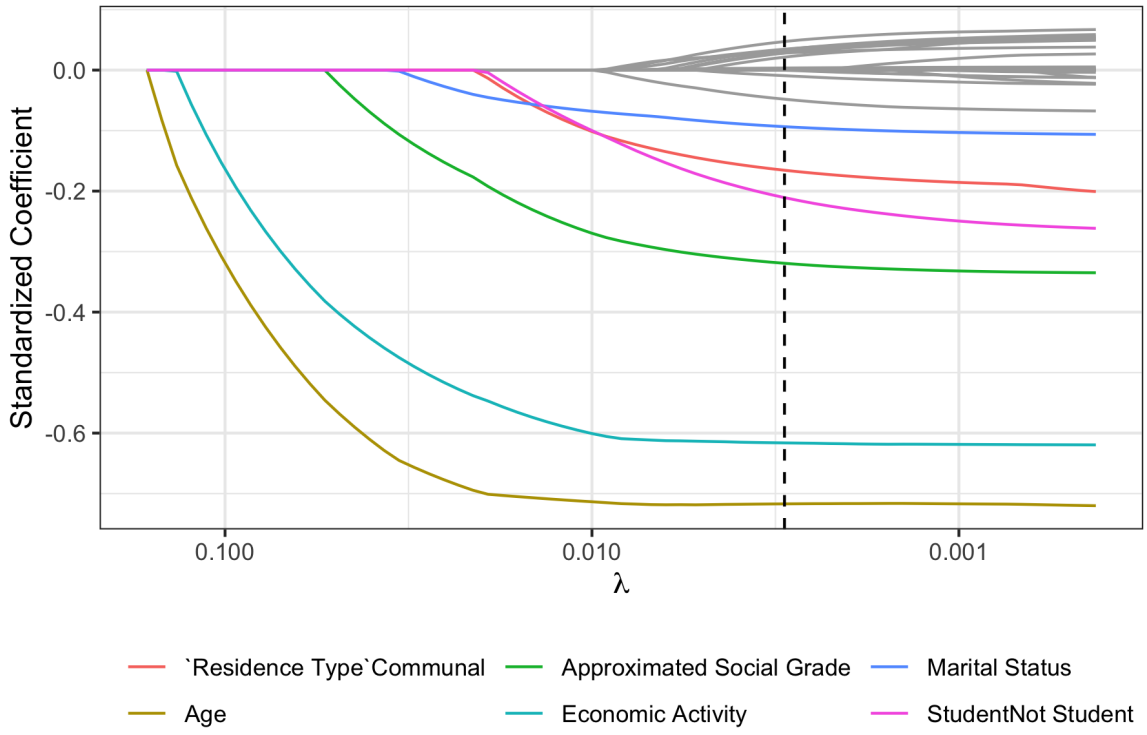

Standardized Coefficient
0.2 0.0 -0.2 -0.4 -0.6
100.0 10.0 1.0 0.1
λ
Age
Economic Activity
Marital Status
Approximated Social Grade
Hours worked per week
StudentNot Student

The trajectory of the CV plot makes it quite clear that this model favours values of lambda closer to zero, with a 1se value of lambda equal to 0.0237. This indicates that it is unlikely penalisation is going to drastically improve the quality of this model, although it hasn't been completely rejected. This actually makes good sense since penalisation exists to reduce the negative effects of high variation in a dataset. My intuition would be that this data is much more likely to suffer from implicit bias than from variation due to the lack of opportunities for massively erroneous entries. That is, it's a survey populated by multiple choice questions. People have no incentive to completely randomise their answers and might even face some repercussions if they did. As well as this, the respondents should more or less know all the answers and not guess. This eliminates the prospect of data governed by probabilities that could lend itself to higher variance. Instead, issues are more likely to occur from more consistent and small mistakes such as a poorly written question causing people to be more likely to put themselves in the wrong category. People could even display their own implicit biases for example over reporting the number of hours they work in a week if they are between two categories. Therefore, given that I would think this data is more prone to bias than variance issues, my guess would be that the misclassification error of the penalisation methods does not demonstrate a significant improvement from the ordinary logistic regression.

4.1.4 Lasso Penalised Logistic Regression

Lastly, the lasso penalisation was also run with 10 folds of cross validation.


25 25 23 21 19 18 12 7 7 4 4 3 2 2 1
Misclassification Error
0.21
0.19
-8 -7 -6 -5 -4 -3 -2
Log(λ)

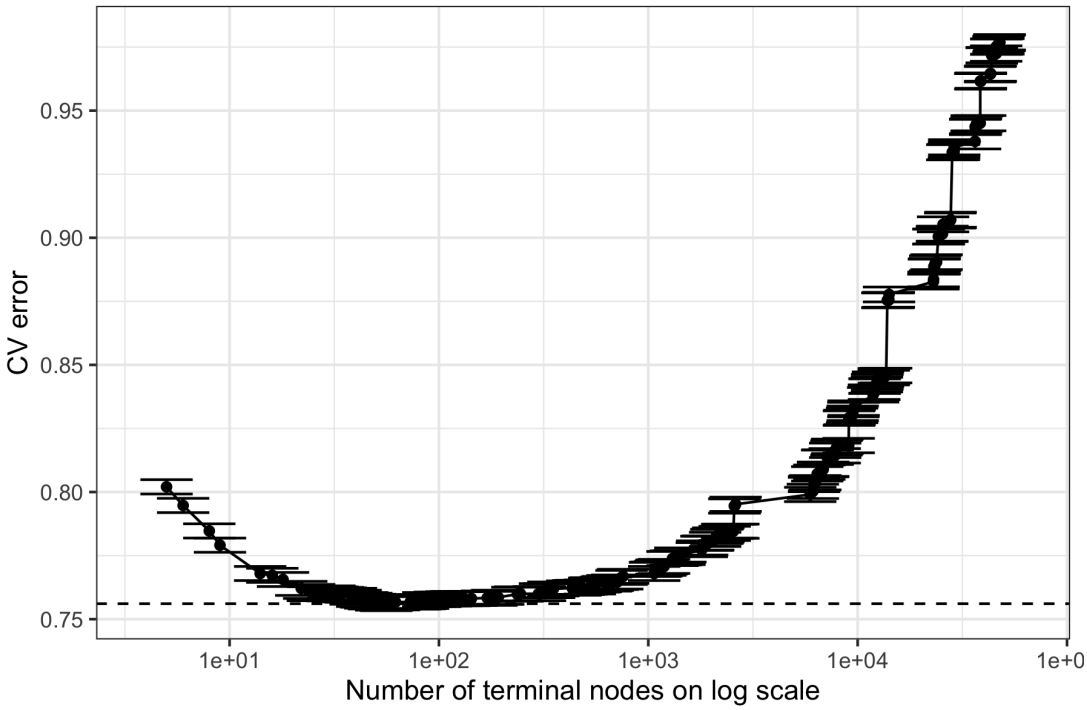

Standardized Coefficient
0.0
-0.2
-0.4
-0.6
0.100
0.010
0.001
λ
`Residence Type`Communal
Approximated Social Grade
Marital Status
Age
Economic Activity
StudentNot Student

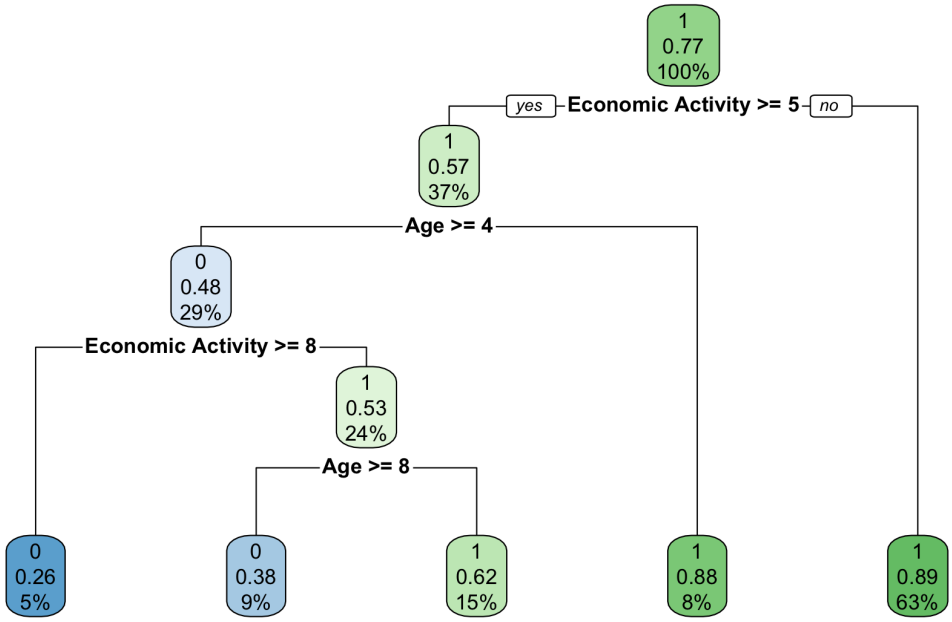
The lasso fit was even closer to rejecting the penalisation with a tiny lambda 1se equal to 0.0023. My thoughts on these penalisation attempts mentioned above remain just as relevant. What both the ridge and lasso models show is that although there are many features having a significant effect, Age is still the most significant predictor of health. Economic Activity and its significance looks like an exciting prospect for exploration and I look forward to trying to find models that might be able to better incorporate it.

4.2 Tree-based Techniques

4.2.1 Ordinary Tree

To fit the optimum tree I started by finding the deepest possible tree and then, using the 1-standard error rule, found the simplest tree such that the error was within a standard deviation of the minimum. I could then prune a default tree using these optimal control parameters to get the optimal tree. Here is the CV plot of the deepest tree with a line demonstrating the optimal parameters it provided followed by a graphic of the final tree itself.



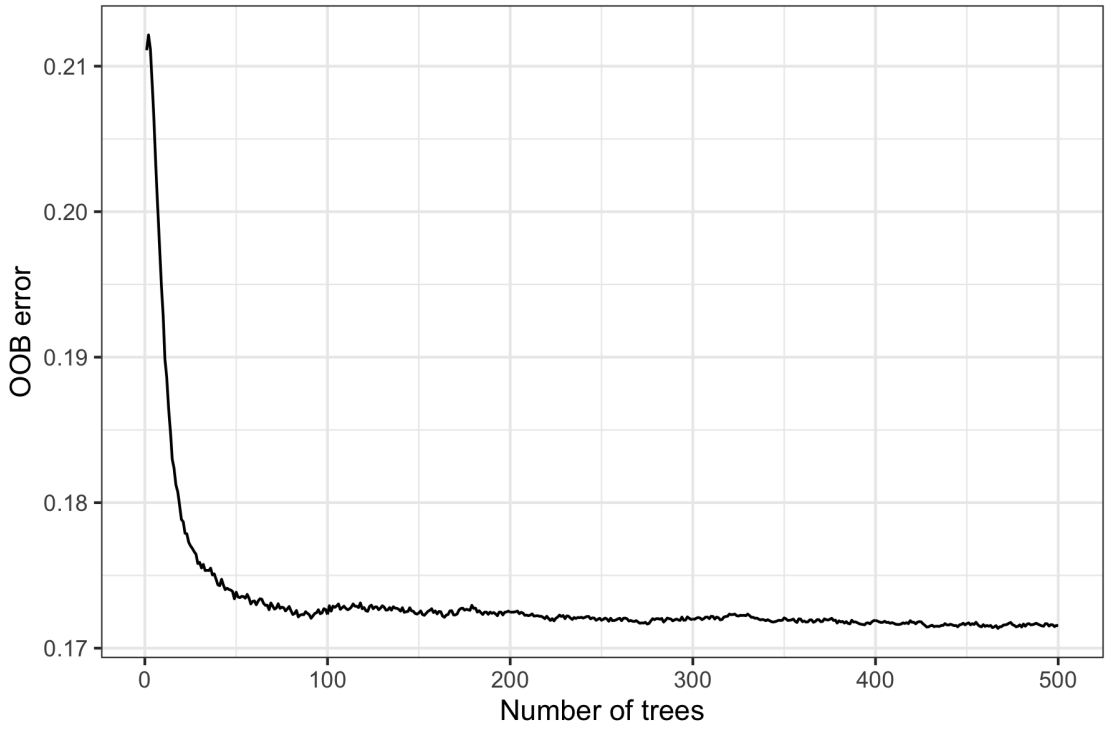


The simplicity of the output tree is satisfyingly readable. The first split after the route node is very encouraging. It immediately puts 73% of the data into a group in which 89% of people are healthy. It does this based on their lack of participation in 3 categories of economic activity. These categories are “Economically Inactive: Retired”, “Economically Inactive: Long-term sick or disabled” and “Economically Inactive: Other”. Clearly, the first two categories are expected to contain very high proportions of unhealthy people! The exciting aspect of this tree is its ability to expand on each feature and consider the individual levels of the categorical variable to identify the most generally insightful survey responses.

4.2.2 Random Forests

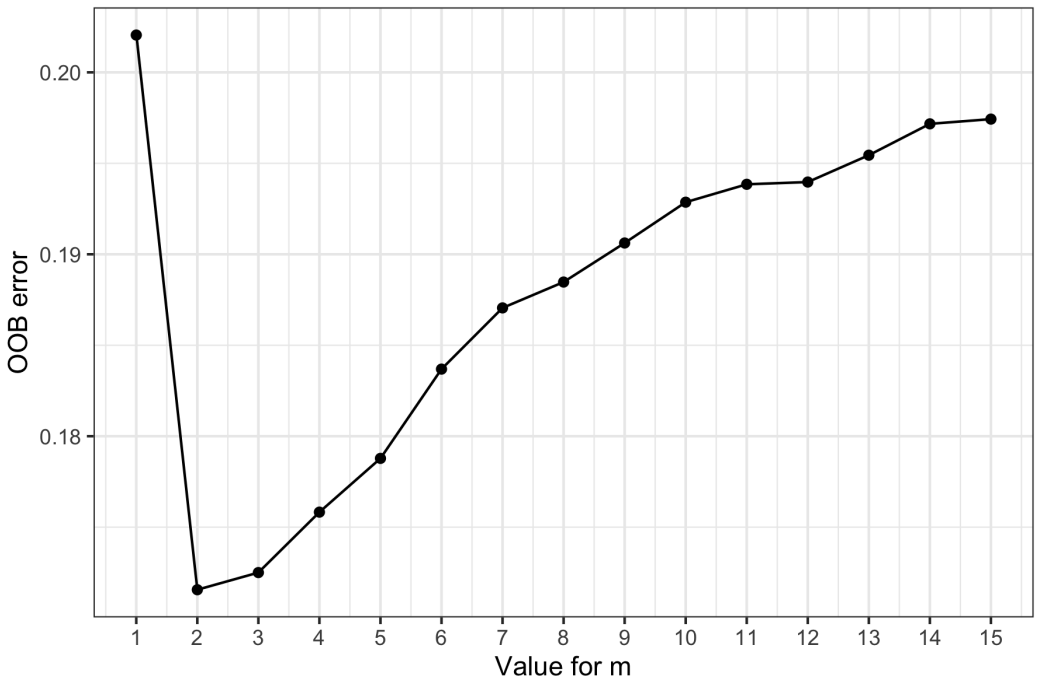
The process of fitting the Random Forest model included fitting the default to a subset of the data (20% of train for memory constraint reasons), iterating over potential ntree and m values and then using these optimal parameters to refit a better, tuned forest. Firstly, ntrees was tuned by calculating the oob error generated by every ntrees value between 1

and 500. This is just to find the value at which the error inevitably plateaus while also keeping the value as low as possible for simplicity's sake.



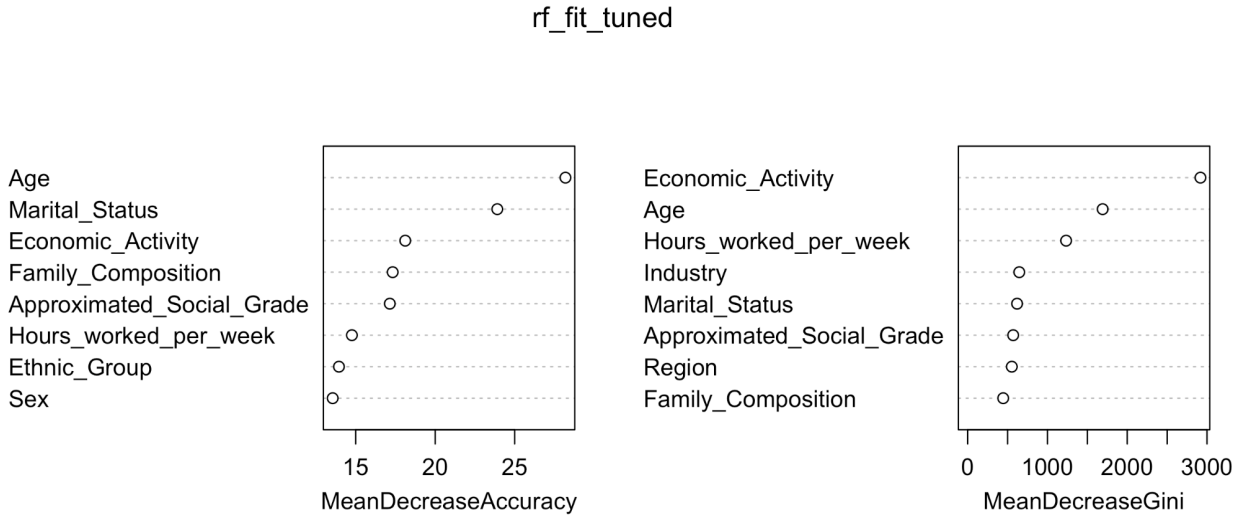
My impression was that the plateau clearly begins before 100 and therefore this was a nice, round, simplistic parameter to choose for the optimal pruning.

Next we had to tune the number of features used for each bootstrap sample. Since there are only 15 features in the data, excluding the response, my little laptop was just about able to make it through calculating the out of back error for a model tuned using each possible value of m.



This graph is interesting because it shows that the RF really benefited from using bootstrap samples with multiple features, but the error then increases consistently as the value of m increases above 2. It also shows that the time put into tuning the RF was well worth it! The optimal value for m shown by this graph, 2, is different from the bootstrap sample size the default would have used: num_features/3 or 15/3 = 5.

Therefore, using 100 trees and a bootstrap sample size of 2, the tuned random forest returned the following breakdown of features and their variable importance by purity based importance and OOB error importance.

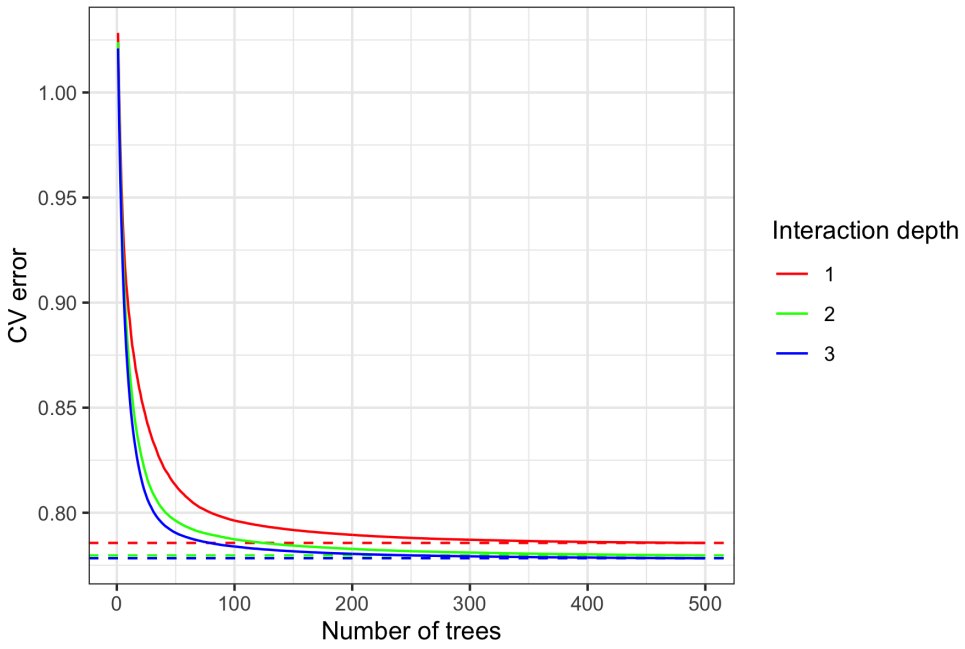


There are several exciting insights to be gained here. Firstly, we see Age and Economic Activity vying for top spot. Knowing the significance of age before this analysis even began, the fact that there are any situations at all where Economic Activity could be more important to healthiness is a bit of a surprise! Secondly, the significance of Marital Status is highlighted by its placements of 2nd and 5th. This is a feature that would have been much lower down on my list of potentially significant variables beforehand. However, after further consideration it makes perfect sense that people who are widowed are much more likely to be unhealthy than those in any of the other categories of that feature.

4.2.3 Boosting

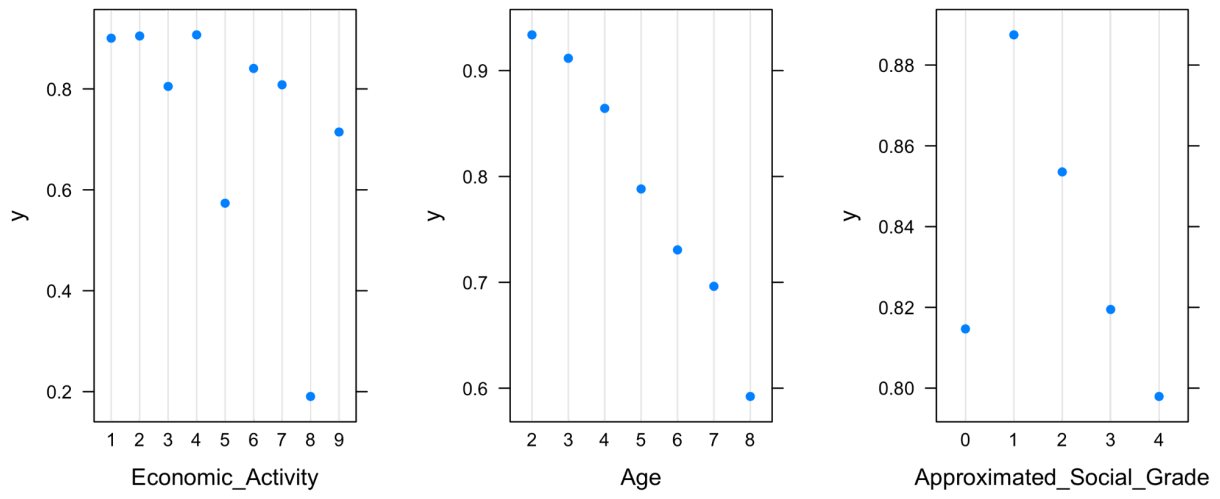
Last, but very much not least, we will consider a boosted tree model. Once again, I had to use a subsample of the training data although this time it was 50% rather than the measly 20% used for training the RF.

The process of fitting the boosted model was simple in that interaction depth was tuned by simply training three trees with depths 1, 2 and 3 and using the best model. Unfortunately, this process took so long that my laptop and I had to negotiate a peace treaty in which I was forced to concede the opportunity to also tune shrinkage and so chose 0.1 as my best guess.



This graph of showing the CV error vs ntrees for each of the three fits demonstrates that using an interaction depth of 3 was optimal. With this optimised model, the next step was to examine the relative influence of the variables, considering the most significant in order to extract and examine their partial dependence plots.

Economic_Activity	75.54909707249790
Age	12.88427040878960
Approximated_Social_Grade	5.467786742476810
Marital_Status	1.7345399886855100
Family_Composition	1.2096380573971300
Industry	0.9639854793489970
Region	0.7133905578721120
Religion	0.6703280075832520
Residence_Type	0.23967869372233800
Hours_worked_per_week	0.19645795177054000
Ethnic_Group	0.15148772211100600
Country_of_Birth	0.09961396993730270
Sex	0.07002052511026130
Population_Base	0.032278573172857900
Student	0.017426249524405400



There are a handful of very interesting observations to be made from these plots. The first is the clear drop in the expected healthiness of category 3 when compared to the first four

categories. The first four categories are 'Economically Active' and third is for people who are economically active but unemployed which I would assume implies they are between jobs. Therefore, I am surprised that there is enough evidence in the data to support this lowered probability of healthiness for this group. The second thing of note is simply the nice and consistent age relationship that still serves well as the benchmark for feature importance on health. Lastly, my favourite observation is the incredibly consistent downwards trend among the social grades as they transition from high to low skill work. Although, the magnitude of the differences are somewhat small, if they were any larger it would bring some serious concern about the state of the country's access to healthcare and maybe even insufficient health and safety laws to protect those in low skilled work!

5. Conclusion

5.1 Model Comparisons

The simplest way to compare each of the 7 models is to compare their misclassification errors and determine which ones made the fewest mistakes when predicting the health of the test set.

Model	Misclass_Err	FPR	FNR
Uni-Log(Age)	0.202	0.736	0.047
Multi-Log	0.184	0.551	0.077
Ridge	0.186	0.596	0.067
Lasso	0.185	0.563	0.075
Ordinary Tree	0.180	0.599	0.058
Random Forest	0.172	0.555	0.061
Boosting	0.170	0.535	0.064

Overall I'm incredibly pleased with these results. It's very exciting that this completely unprepared dataset is able to be grappled with in such a way that these models can consistently have such a high level of accuracy. I'm also relieved that the univariate logistic

regression using just age has a noticeably worse error than all of the models that use the full feature set as this entire analysis would have otherwise been pointless!

The best performing model was the tree-based boosting method, only rivalled by the Random Forest. This makes sense given that these are traditionally the most accurate predictive models as well as the most computationally intensive (of those we used). Another reason this outcome is intuitive comes from hypotheses I made earlier about this being a bias-heavy dataset and therefore deriving little value from penalised regression. As a result, not only does the ordinary tree beat both ridge and lasso, but so does the ordinary multivariate logistic regression! Technically, I don't think this should be since the lasso and ridge should break down into the multi-log by setting lambda equal to 0 when it's being tuned. However, I think the issue lies with the glm libraries that don't always allow this to occur exactly, instead setting a small but non-zero lambda which then decreases the model accuracy.

As well as this, in order to be sure that the class imbalance wasn't constraining the model accuracy, I lazily re-evaluated the predictions of the tree-based boosting model using slightly different threshold values, 0.4 and 0.6. Both saw notable increases in misclassification error, confirming my earlier thought that the imbalance is small enough that there is no significant need for threshold tuning.

5.2 Takeaways

This analysis set out to find features that could be used to improve health level predictions even in the presence of age, a very strong and intuitive indicator. Therefore, I will now analyse features that fit my standards of significance and see if they pass the remaining criteria. I will consider features from the two strongest tree models and disregard the regression penalisations due to their higher misclassification errors. Aside from age, the most significant variables in the random forest were Economic Activity, Marital Status and Hours worked per week. On top of that, the tree-based boosting model derived a lot of value from Social Grade.

In my introduction I mentioned two tests. The first was that the features must provide value above and beyond what age can do alone. The fact that both these models have a significantly lower misclassification error than the Age-only model while lending the highest priority to some or all of these features is enough for them to all pass this test. However, the second test stipulates that their value must not be directly derived from age. That is, at least one level of that feature implies a solid indication of the age of that person. For example, one of the features, marital status, has a category for 'Widowed'. Intuitively, I am assuming that it is this category that is providing the majority of the insight garnered from this feature since being widowed makes it incredibly likely that that person is in old age. That is, since couples are about the same age and the vast majority of deaths happen in old age, a widow is very likely to also be old. Therefore, this feature is insufficiently independent from age. The Economic Activity feature is interesting because at first glance it's category for Retired should be an immediate failure of the age-independence rule. However, although I'm sure this category is contributing to the significance of this feature, its partial dependence plot for the tree-based boosting model saves it. In fact, it's the very existence of the 'Retired' classification that provides additional value to its remaining categories that have been disassociated from age due that distinction. That is, people who ticked 'Inactive:Long term injury or health issue' had to choose to click that over the retired option. Therefore, it's much less likely these people are also retired and therefore associated with age. As such, the value of this age-independent category, amongst others in the feature, means that Economic Activity passes the tests! The remaining two, Hours per week and Social Grade also pass as there is absolutely no direct link to age. If anything, working more hours or in a less skilled industry probably makes an individual less likely to be old. My only concern is the potential for some overlap between the two features themselves as the level of skill required in an industry is likely to have some effect on the standard number of working hours or vice versa. However, this issue is not in the requirements for acceptance and therefore three features, Economic Activity, Hours Worked per Week and Social Grade have stood out as age-independent attributes that can be used to more accurately and deterministically characterise people as healthy or not than age alone.

5.3 Limitations

One limitation of this analysis is the relationship between the class imbalance and categorical data. Either one is not a considerable issue by itself. However, the potential for a few highly predictive but not exhaustive feature-levels can lead to model-breaking exaggerations of the imbalance. For the sake of an explanatory example, let's say that the "Economic Activity: Inactive - Injury" feature-level is exclusively unhealthy people and contains 7% of the population. As well as this, the two extremes of the age categories each contain 10% of the population and are also exclusively populated by healthy and unhealthy people respectively. Lastly, let's assume that the starting class imbalance is 20% unhealthy, 80% healthy. Now consider a diagram representing a tree-based model where the model, since these categories are definitive, immediately makes splits on them to reduce the remaining dataset by getting them into leaf nodes quickly. However, the class imbalance in the remaining data to be classified by the model has changed. The original 80%/20% healthy/unhealthy split is now 70%/3%. In just a few splits, the model is now trying to fit a very imbalanced dataset and therefore has a large incentive to favour the majority classification or even just lazily classify the entire remaining as healthy and only suffer a small increase in misclassification error. This concept of a lazy model can be demonstrated somewhat by the optimum tree diagram or the fact that the RF model had consistently increasing error if it used a bootstrap sample size greater than two. The combination of an abundance of categorical variables containing at least some levels with very good predictions for the minority class, and an imbalance that allows for a minority class in the first place means that this analysis is more exposed to lazy models.

5.4 Follow-ups

There are two primary extensions or new strategies that I would employ to improve or progress this analysis. Firstly, one could split the categorical variables into sets of binary variables. That is, instead of having 9 categories for the 'Religion' variable, have 9 variables with binary categories that represent each religion. As well as allowing for entries to exist in more than one category for some feature, such as race, it would allow the models to have

greater variability in the extent to which different levels of a feature are significant. For example, perhaps one category of a 12-level feature is a very strong indicator of Health but the remaining 11 have almost no effect. Currently, the models are less able to differentiate within a variable's levels and therefore might provide a weaker fit.

Secondly, in order to really examine the effect of the chosen variables compared to that of age as well as the potential effects of some of the additional features that overlap with age, I would like to play around with removing age from the model creation process. It would be insightful to see the strength of our predictive variables without age to possibly dilute their predictive power. As well as this, one could examine the nature of the interdependence between age and these variables by excluding or adapting specific troublesome categories such as 'widowed' in Marital Status.

Finally, I think a logical follow-up would be to run this entire analysis again on a new dataset and attempt to verify the results. Luckily, there's no shortage of data considering that what we used is just 1% of the 2011 Census. Not only could we repeat this on other samples from the same year but another Census has taken place this year and there's no reason that I can think of that would lead to a significant change in which features are good predictors of health. Therefore, assuming the data is in a similar format, repeating this analysis on data from different decades could be a good way of tuning the techniques that were employed as well as an opportunity to possibly discover interesting changes that occur between decades.

6. Appendix

6.1 Categorical feature level translations

6.1.2 Region

Code	Meaning
E12000001	North East
E12000002	North West

E12000003	Yorkshire and the Humber
E12000004	East Midlands
E12000005	West Midlands
E12000006	East of England
E12000007	London
E12000008	South East
E12000009	South West
W92000004	Wales

6.1.2 Residence Type

Code	Meaning
C	Resident in a communal establishment
H	Not resident in a communal establishment

6.1.3 Family Composition

Code	Meaning
1	Not in a family
2	Married/same-sex civil partnership couple family
3	Cohabiting couple family
4	Lone parent family (male head)
5	Lone parent family (female head)
6	Other related family
0	No code required

6.1.4 Population Base

Code	Meaning
1.	Usual resident
2.	Student living away from home during term-time

0.	Short-term resident

6.1.5 Sex

Code	Meaning
1	Male
2	Female

6.1.6 Age

Code	Meaning
1	0 to 15
2	16 to 24
3	25 to 34
4	35 to 44
5	45 to 54
6	55 to 64
7	65 to 74
8	75 and over

6.1.7 Marital Status

Code	Meaning
1	Single (never married or never registered a same-sex civil partnership)
2	Married or in a registered same-sex civil partnership
3	Separated but still legally married or separated but still legally in a same-sex civil partnership
4	Divorced or formerly in a same-sex civil partnership which is now legally dissolved
5	Widowed or surviving partner from a same-sex

	civil partnership

6.1.8 Student

Code	Meaning
1.	Yes
2.	No

6.1.9 Country of Birth

Code	Meaning
1	UK
2	Non UK
0	No Code required

6.1.10 Ethnic Group

Code	Meaning
1.	White
2.	Mixed
3.	Asian and Asian British
4.	Black or Black British
5.	Chinese or Other ethnic group
0.	No code required

6.1.11 Religion

Code	Meaning
1.	No religion
2.	Christian
3.	Buddhist
4.	Hindu

5.	Jewish
6.	Muslim
7.	Sikh
8.	Other religion
9.	Not stated
0.	No code required

6.1.12 Economic Activity

Code	Meaning
1.	Economically active: Employee
2.	Economically active: Self-employed
3.	Economically active: Unemployed
4.	Economically active: Full-time student
5.	Economically inactive: Retired
6.	Economically inactive: Student
7.	Economically inactive: Looking after home or family
8.	Economically inactive: Long-term sick or disabled
9.	Economically inactive: Other
0.	No code required

6.1.13 Occupation

Code	Meaning
1.	Managers, Directors and Senior Officials
2.	Professional Occupations
3.	Associate Professional and Technical Occupations

4.	Administrative and Secretarial Occupations
5.	Skilled Trades Occupations
6.	Caring, Leisure and Other Service Occupations
7.	Sales and Customer Service Occupations
8.	Process, Plant and Machine Operatives
9.	Elementary Occupations
0.	No code required

6.1.14 Industry

Code	Meaning
1.	Agriculture, forestry and fishing
2.	Mining and quarrying; Manufacturing; Electricity, gas, steam and air conditioning system; Water
3.	Construction
4.	Wholesale and retail trade; Repair of motor vehicles and motorcycles
5.	Accommodation and food service activities
6.	Transport and storage; Information and communication
7.	Financial and insurance activities; Intermediation
8.	Real estate activities; Professional, scientific and technical activities; Administrative and s
9.	Public administration and defence; compulsory social security
10	Education
11	Human health and social work activities
12	Other community, social and personal service activities; Private households employing

	domestic
0	No code required

6.1.15 Hours Worked per Week

Code	Meaning
1.	Part-time: 15 or less hours worked
2.	Part-time: 16 to 30 hours worked
3.	Full-time: 31 to 48 hours worked
4.	Full-time: 49 or more hours worked
0.	No code required

6.1.16 Approximated Social Grade

Code	Meaning
1.	AB
2.	C1
3.	C2
4.	DE
0	No code required

6.1.17 Health

Code	Meaning
1.	Very good health
2.	Good health
3.	Fair health
4.	Bad health
5.	Very bad health
0.	No code required

6.2 Counts by feature of 0-level entries in raw data

name	Informative	-9
Age	569741	NA
Approximated Social Grade	445638	124103
Country of Birth	562937	6804
Economic Activity	457123	112618
Ethnic Group	562937	6804
Family Composition	550890	18851
Health	562937	6804
Hours worked per week	267420	302321
Industry	419757	149984
Marital Status	569741	NA
Occupation	419757	149984
Population Base	569741	NA
Region	569741	NA
Religion	562937	6804
Residence Type	569741	NA
Sex	569741	NA
Student	569741	NA

6.3 Retroactive age-health plot