Homework assignment 2

Data Analysis 4: Prediction Analytics with Introduction to Machine Learning 2017/2018 Winter

Peter Paziczki

2018 February 18

1. Predicting firm default

We have been provided the bisnode_all.csv dataset containing more than 150.000 observations, covering financial data from 2011 to 2016, management related information and further information about employment and industrial classification. Our task is to define firm default and build models that can predict it.

1.1 Defining firm default

I am going to consider a firm defaulted if there are no records about its sales results in 2014 and 2015. Having no records means that it is either zero and / or missing. I could see many examples of missing sales records of one year, but two missing records in row is enough for me to say that the company has defaulted.

It also means that I cannot work with companies founded later than 2013. Companies that had zero as record of sales in all years from 2011 to 2013 have also been dropped, those are considered as inactive, irrelevant for our study. I have dropped the year 2016, because when the data was gathered, not all the the companies have submitted their financial records for 2016.

I created a binary flag called `defaulted`, 1 indicates that the company has defaulted.

1.2 Data preparation

I need to find and create relevant predictors to able to build a model with actual predicting power. Let

I am expecting company age to be an relevant predictor, I am assuming that the older a company, the more stable it is and the less likely it is to default.

Sales

Sales plays an important role in our study, so I have computed sales growth for 2011 to 2012 and for 2012 to 2013. The equation was *Current Period Net Sales - Prior Period Net Sales) / Prior Period Net Sales*. Previously I have dropped all the observations that would have NA in sales records in any of the years from 2011 to 2013, but to make sure that that sales growth wouldn't be NA, I have to avoid having zero in the denominator, therefore I have dropped the observations that has zero in sales records for year 2011 and 2012. After calculating sales growth I grouped the sales growth into three categories (by industries) and used that for modelling. It proved to be one of the most performing predictors (see variable importance plot).

Age

I am expecting age to be an important predictor. I have computed it using the year the firm was founded in and substracted it from 2015. In addition to that I have cut the observations into 5 years long age groups, please find a table below showing the number of observations in the different groups.

agegroup	min_age	max_age	n
[20,25)	20	24	2491
[10,15)	10	14	2797
[0,5)	2	4	1348
[15,20)	15	19	3163

agegroup	min_age	max_age	n
[5,10)	5	9	4728
[25,30)	25	29	612

Later there will be more about finding the best functional form of age variable, but in short, a quadratic from described it the best.

Region

Region sounded to be an interesting potential predictor, so I have included that too. Please find a table below showing the regions available and the number of companies in them.

region	N
West	2393
Central	8969
East	3777

It turned out to be one of the least important predcitors. I had assumption that region could have an effect on firm default, but it is far less relevant than I thought it would be.

Income before tax

Similar to sales growth, I also investigated the yearly change in `inc_bef_tax` variable from year 2011 to 2013 and grouped the results into three groups (by industries). It turned out to be an improtant predictor.

Size

I defined a way to group the companies by their size. I chose a year and added the `curr_assets` and `fixed_assets` variables and based on the results I have grouped them into three different size groups (by industries). It also turned out to be one of most important predictors (see variable importance plot).

Firm default

After the data preparation I am checking, how many companies have been flagged as *defaulted*:

0	1
14056	1083

I am still checking the number of defaulted companies but by industry (`ind` variable) and age groups (`agegroup` variable).

```
##    defaulted                    ind    N
## 1:         0 I hotel and restaurant 9658
## 2:         0           C Manuf_auto  365
## 3:         1 I hotel and restaurant  921
## 4:         0      C Manuf_equipment 4033
## 5:         1      C Manuf_equipment  146
## 6:         1           C Manuf_auto   16
```

```
##     defaulted agegroup    N
##  1:         0  [20,25) 2376
##  2:         0  [10,15) 2604
##  3:         0    [0,5) 1228
##  4:         0  [15,20) 2978
##  5:          1   [5,10)  447
##  6:          0   [5,10) 4281
##  7:          0  [25,30)  589
##  8:          1  [10,15)  193
##  9:          1  [15,20)  185
## 10:          1  [20,25)  115
## 11:          1    [0,5)  120
## 12:          1  [25,30)   23
```

Approximately 7% of the firm have been labelled to be defaulted.

ROE

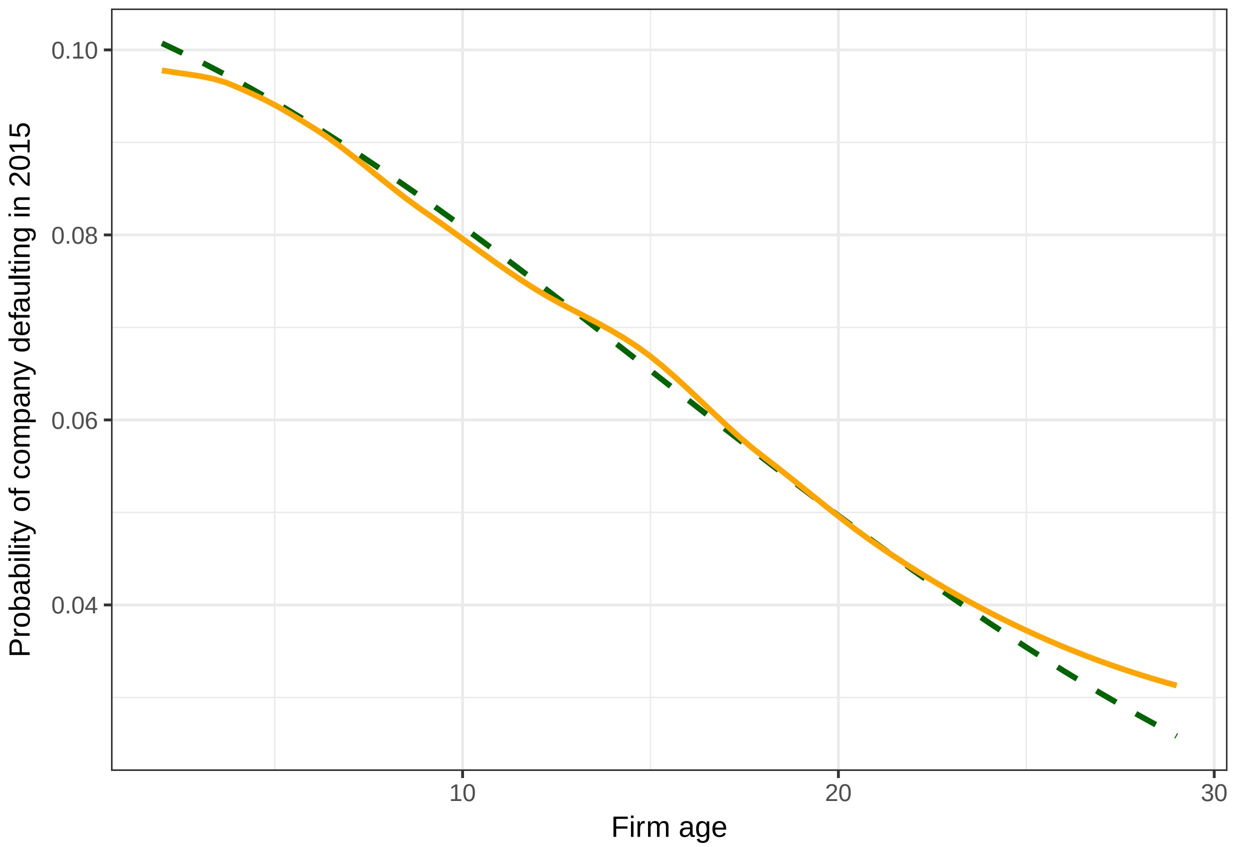
I created one additional measure of company performance,Return on Equity (ROE). ROE is by far the most widespread simple tool to measure the net earnings ratio of a company or an investment. The reason I chose ROE is that it is easy to handle and to calculate, I have all the data needed to calculate it, and it is proven to be a good indicator of companies' profitability. I need just two variables, the Net income (`profit_loss_year`) and Shareholder's Equity (`share_eq` variable) to do the calculation. Calculating ROE for a given year is a "static" value and I calculated it for years 2011 to 2013.

I grouped the values into three groups (by industries), as I did previously.

1.2 Benchmark model

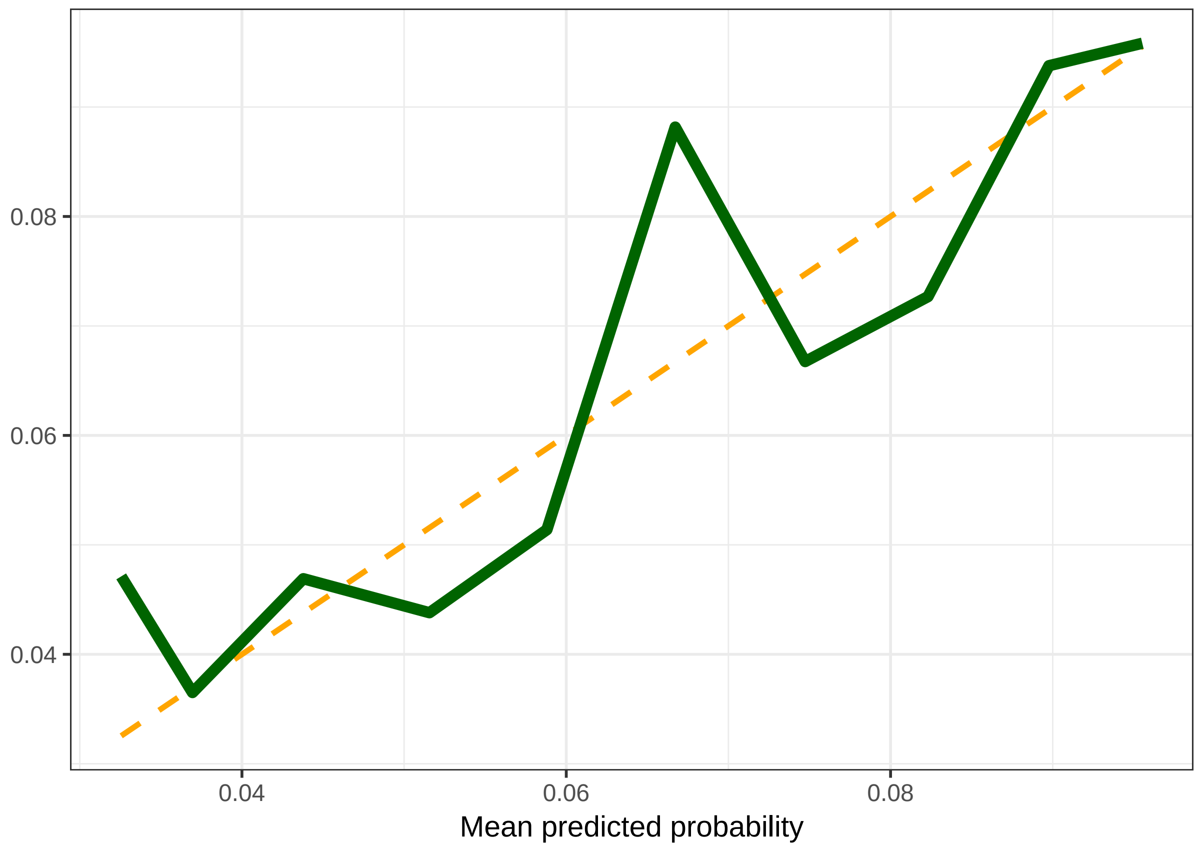
As a benchmark model I am starting with a logit model with one single predictor, square of age.

```
##
## Call:  glm(formula = defaulted ~ poly(age, 2), family = "binomial",
##     data = data)
##
## Coefficients:
##   (Intercept)  poly(age, 2)1  poly(age, 2)2
##        -2.609        -40.833         -5.705
##
## Degrees of Freedom: 15138 Total (i.e. Null);  15136 Residual
## Null Deviance:       7800
## Residual Deviance: 7703  AIC: 7709
```



Please find a calibraton curve below:

```
##        p       mean
## 1: FALSE 0.07153709
```



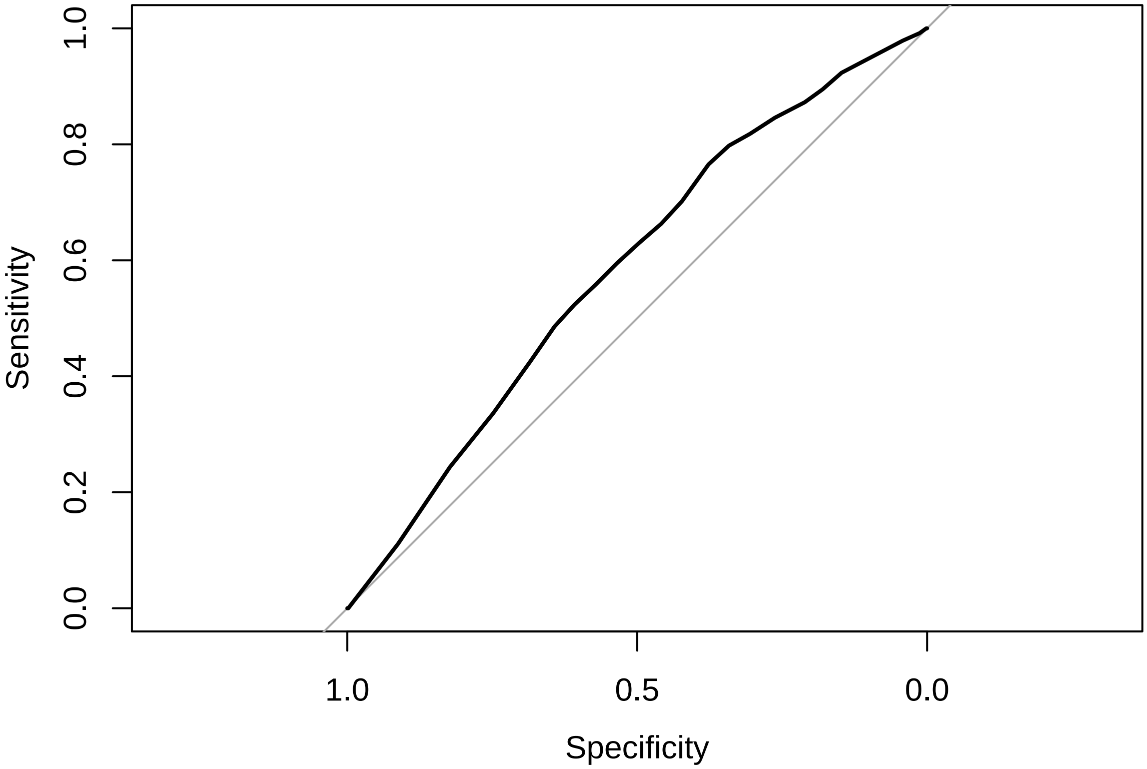
Brier score

Let's see the Brier score of the benchmark model. The Brier score is a way to measure the accuracy of the prediction. It is the mean squared difference of predicted probability and the actual outcome.

```
## [1] 0.06601458
```

ROC curve

The ROC curve shows the true positive rate for a given false positive rate. There is also a line on the plot with a slope of 1, it shows the case when the are under the ROC curve is 0.5, that is the case when the prediction is not better than a random guess. The point is to cover the highest possible AUC (area under the ROC curve).



```
##
## Call:
## roc.default(response = data$defaulted, predictor = data$p, plot = T)
##
## Data: data$p in 14056 controls (data$defaulted 0) < 1083 cases (data$defaulted 1).
## Area under the curve: 0.5857
```

In case of the benchmark model the ROC curve is not far from the 45 degree line, the prediction has to be improved.

1.3 Creating training and test set

I am creating training and test data sets to train and validate the models. The training set 80% of the whole data, randomly chosen. The other 20% goes to the test set.

2. Building models

2.1 Building logit models

I am building four different logit models, first is the most simple one and the last is the most complex one. I am expecting the last to perform the best, we will see that a bit later. Logit is a regression model that has the property of yielding results between zero and one, which is its advantage over linear regressions. These are the models:

m1 ~ age

m2 ~ poly(age, 2) + region + f_ind

m3 ~ poly(age, 2) + region + f_ind + size_cat + ROE_2011_cat + ROE_2012_cat + ROE_2013_cat

m4 ~ age + poly(age, 2) + poly(age, 3) + region + f_ind + size_cat + sales_cat_2012 + sales_cat_2013 + income_cat_2012 + income_cat_2013 + ROE_2011_cat + ROE_2012_cat + ROE_2013_cat

2.2 Running logit models in-sample

First I am running the models without cross-validation on the training set. Please find the results (Brier score and AUC) of the four models in the following table:

```
##    in_sample_brier in_sample_auc
## m1 0.06559726      0.5910671
## m2 0.06504761      0.6365239
## m3 0.0600577       0.7917305
## m4 0.04929394      0.90141
```

The most complex model yielded the highest AUC and the best Brier score.

2.3 Running logit models with 5-fold cross-validation

Now let's run the same four models on the training set but now with cross-validation.

```
## $`model 1`
## BRIER_SCORE    ROC_AREA
##  0.06563421  0.59080511
##
## $`model 2`
## BRIER_SCORE    ROC_AREA
##   0.0651199   0.6329050
##
## $`model 3`
## BRIER_SCORE    ROC_AREA
##   0.0602839   0.7890776
##
## $`model 4`
## BRIER_SCORE    ROC_AREA
##  0.04981937  0.89835251
```

The results are similar in a sense that more complex models provided better results.

3 Using machine learning - Random Forest

Random forest is a machine learning method, growing many regression trees, 200 in our case. It uses the training sample to create many similar but slightly different samples with the use of bootstrapping. After creating 200 trees it averages out the results and yield one final tree. This procedure turned out to reduce overfitting successfully. In addition to that I am using another setting, three is set as the number of variables tried at each split.

I used the following variables for the prediction:

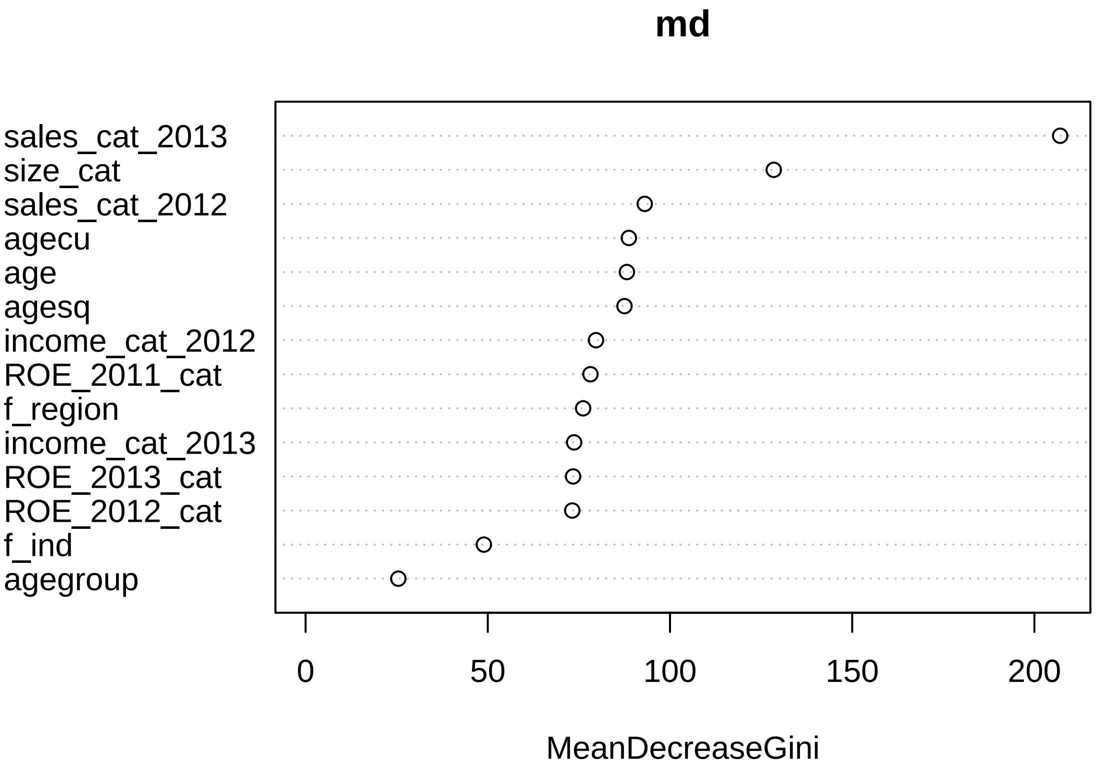
age + agesq + agecu + f_region + agegroup + f_ind + size_cat + sales_cat_2012 + sales_cat_2013 + income_cat_2012 + income_cat_2013 + ROE_2011_cat + ROE_2012_cat + ROE_2013_cat

The model yielded the following confusion martix:

```
##
## Call:
```

```
##  randomForest(formula = mrf, data = d_train, ntree = 200, mtry = 3)
##                Type of random forest: classification
##                      Number of trees: 200
## No. of variables tried at each split: 3
##
##         OOB estimate of  error rate: 7.03%
## Confusion matrix:
##       0   1 class.error
## 0 11095 155  0.01377778
## 1   697 164  0.80952381
```

Please find a variable importance plot below. It captures the contribution of different variables to predicting power, the variable importance is plotted in a descending order. `agegroup` seems to contribute the less.



After training the random forest, let's try it on test set with different threshold, please find two confusion matrices below. I was trying out different thresholds, 0.05 and 0.10 respectively. The second is better in a sense that the number of false negative are much less.

```
##
##         0    1
##    0 1912    1
##    1  333  176
```

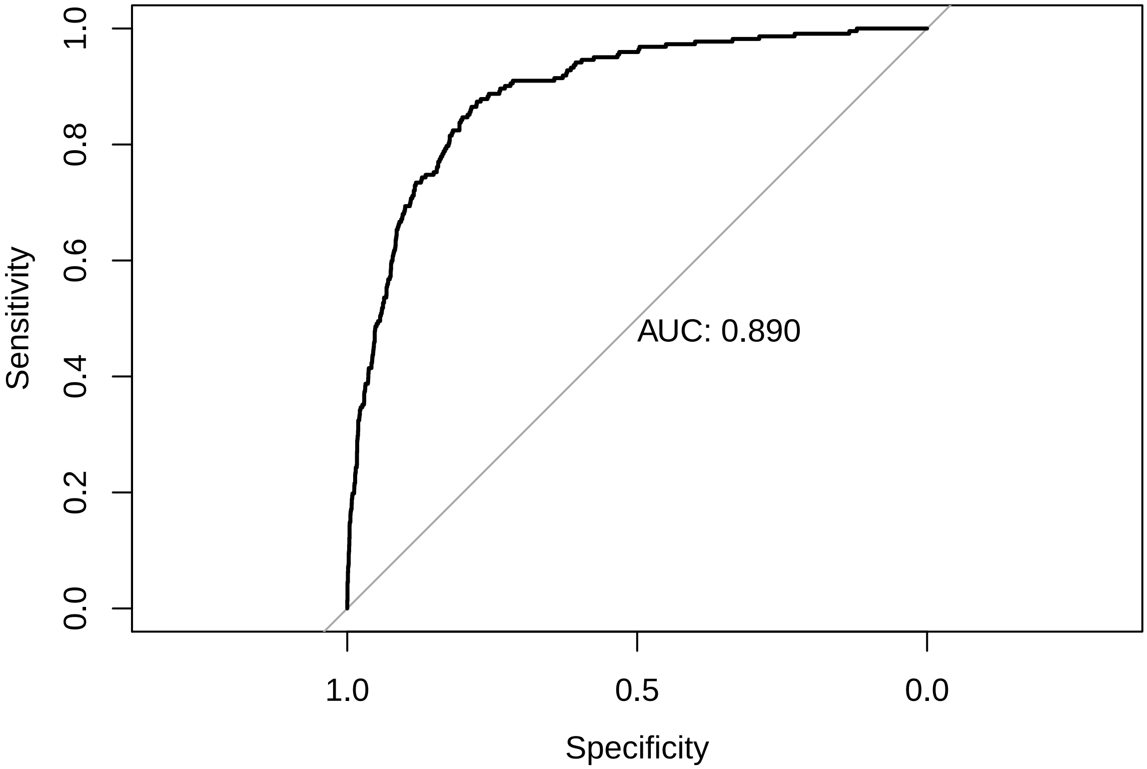
```
##
##         0    1
##    0 2052    2
##    1  193  175
```

4. Comparing results

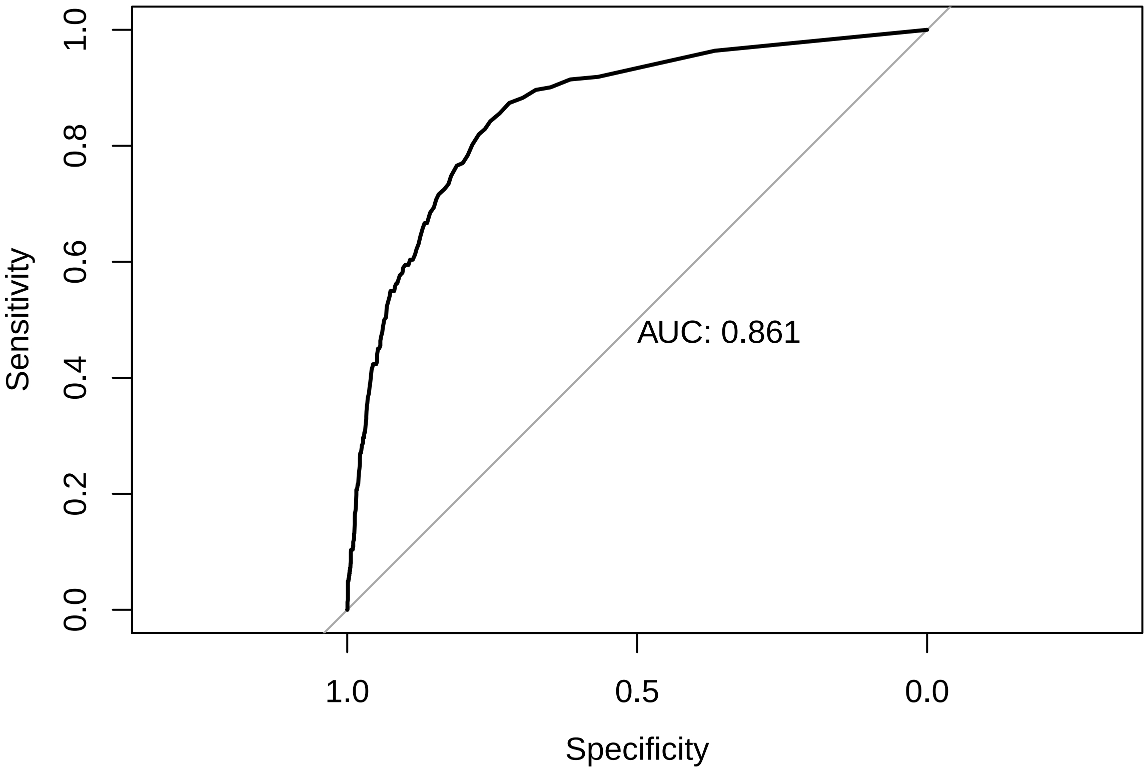
Now I am going to compare the performance of the best logit model and the random forest forest method on the test set. The Brier score of the logit model is the following. It is lower

```
## [1] 0.05136391
```

The logit model has the provided the following ROC curve. It seems to provide a high AUC value.



The random forest gave the following ROC curve. Compared to logit model it provided a lower AUC value.


Sensitivity
0.0
0.2
0.4
0.6
0.8
1.0
Specificity
1.0
0.5
0.0
AUC: 0.861