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, 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.

1.2 Data preparation

I have dropped the year 2016, because when the data was gathered, not all the the companies have submitted their financial records for 2016.

Sales seemed to be an interesting predictor, so I have computed the change in sales from year to year.

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.

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.

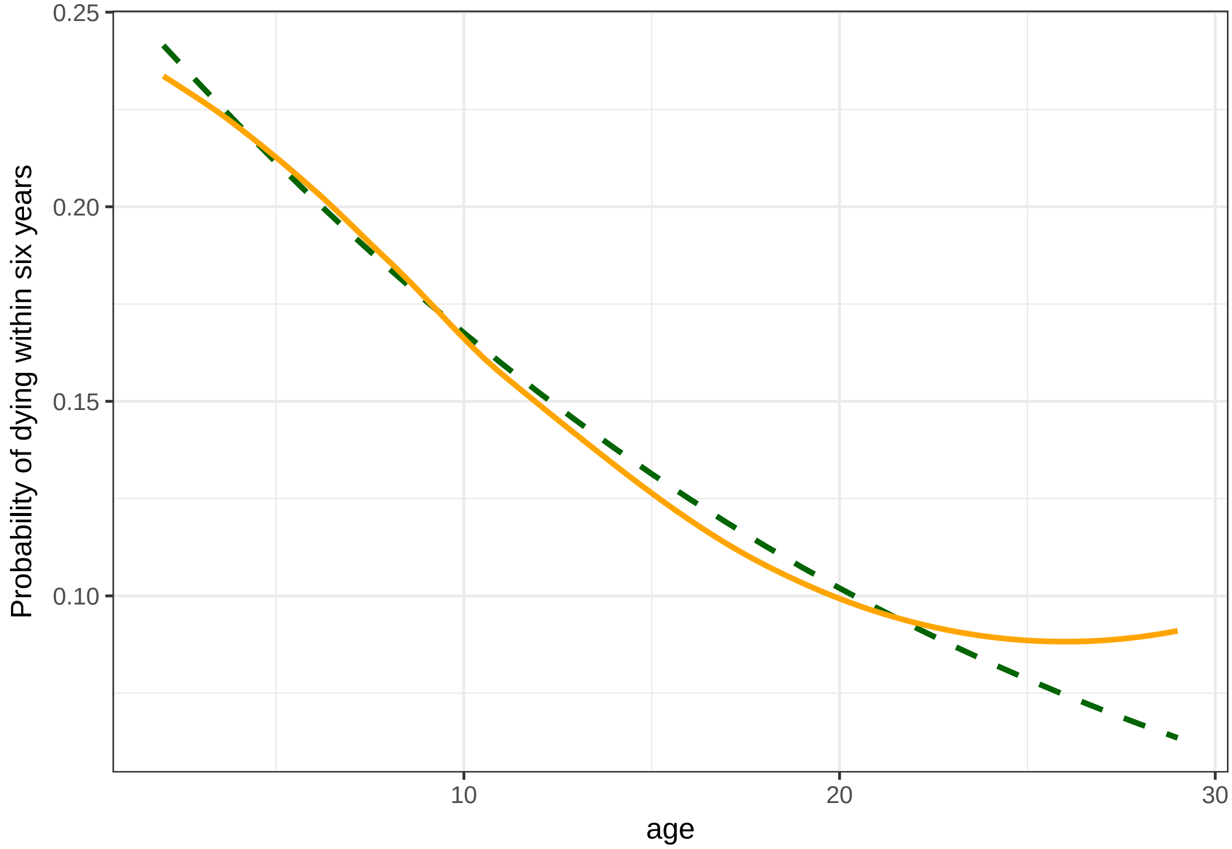
```
##    agegroup min_age max_age    n
## 1:   [5,10)       5       9 6281
## 2:  [20,25)      20      24 3173
## 3:  [10,15)      10      14 3618
## 4:    [0,5)       0       4 2329
## 5:  [15,20)      15      19 4023
## 6:  [25,30)      25      29  776
## 7:       NA     NaN     NaN  232
## 8:  [30,35]      31      35    5
```

```
##  num [1:20437] 7 9 23 14 11 14 4 17 8 5 ...
```

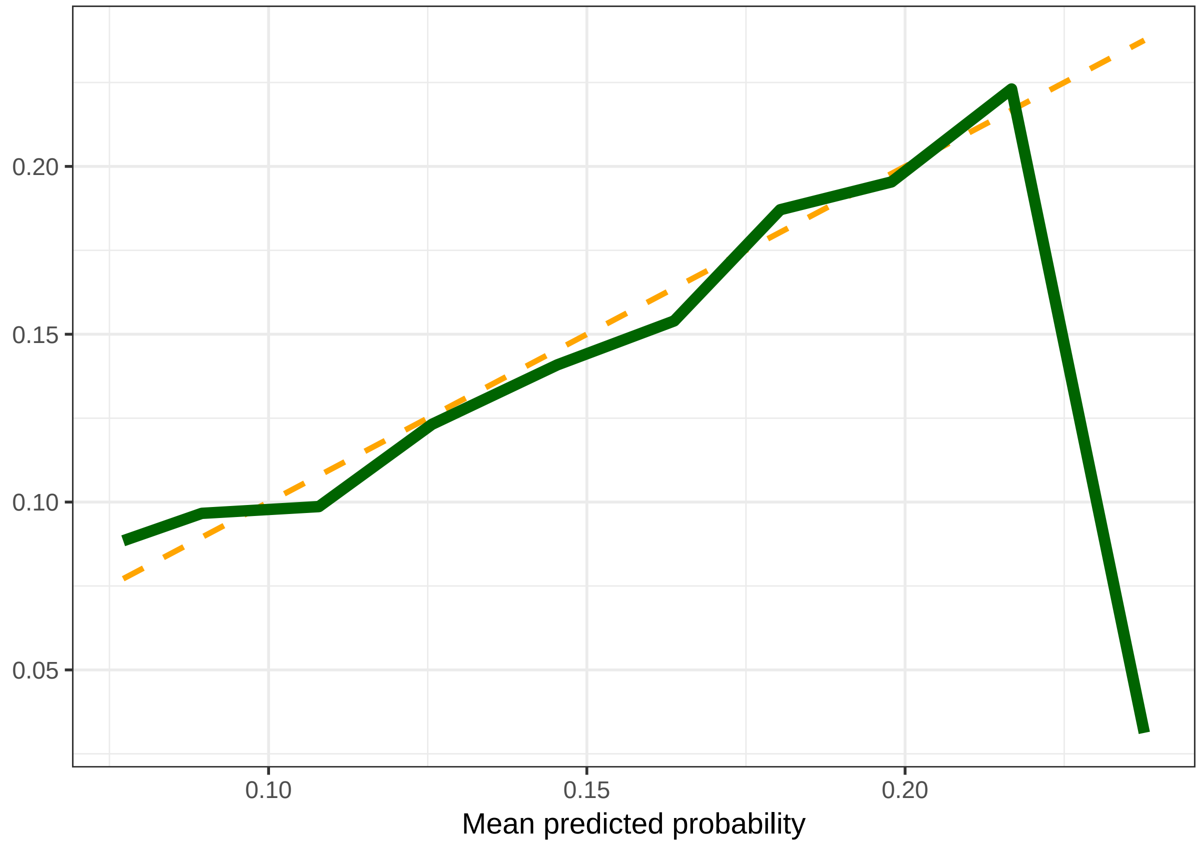
Region would be interesting predictor if it were relevant, I am going to investigate it.

```
##
## Call:  glm(formula = defaulted ~ age, family = "binomial", data = data)
##
## Coefficients:
## (Intercept)          age
```

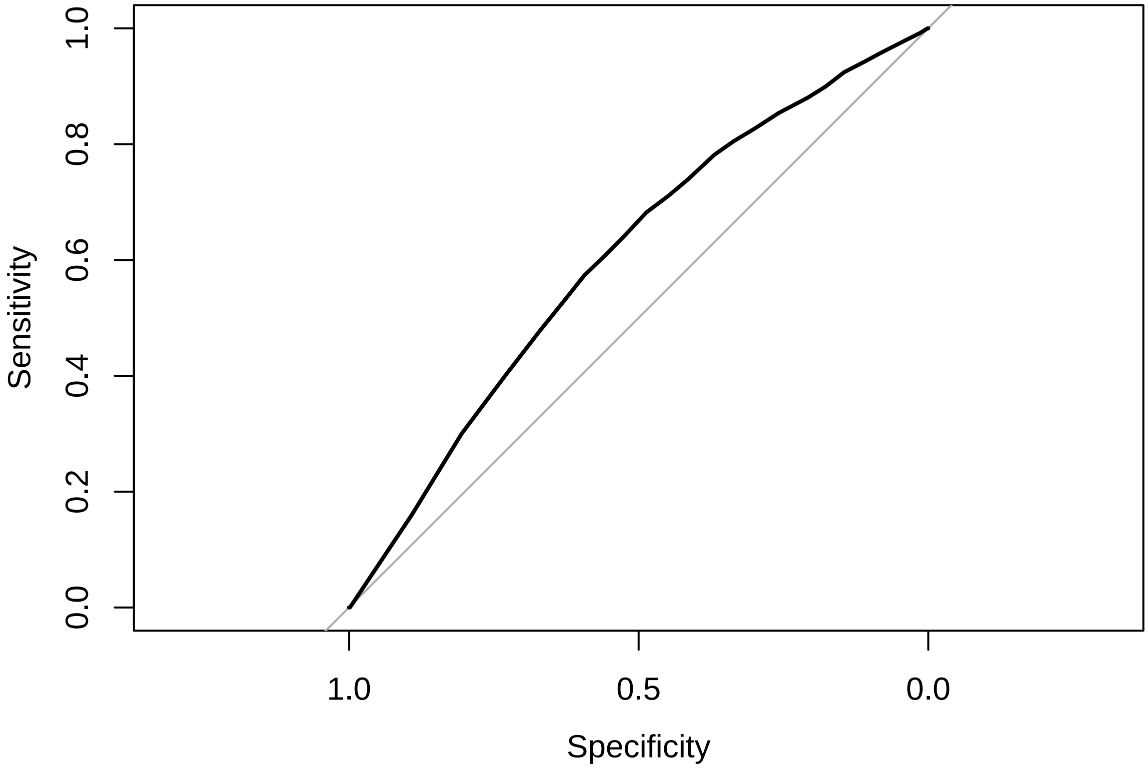
```
##    -1.03016     -0.05726
##
## Degrees of Freedom: 19485 Total (i.e. Null);  19484 Residual
## Null Deviance:       16860
## Residual Deviance: 16510     AIC: 16520
```




```
##        p      mean
## 1: FALSE 0.1558042
```




```
## [1] 0.1292153
## [1] 0.1315292
```

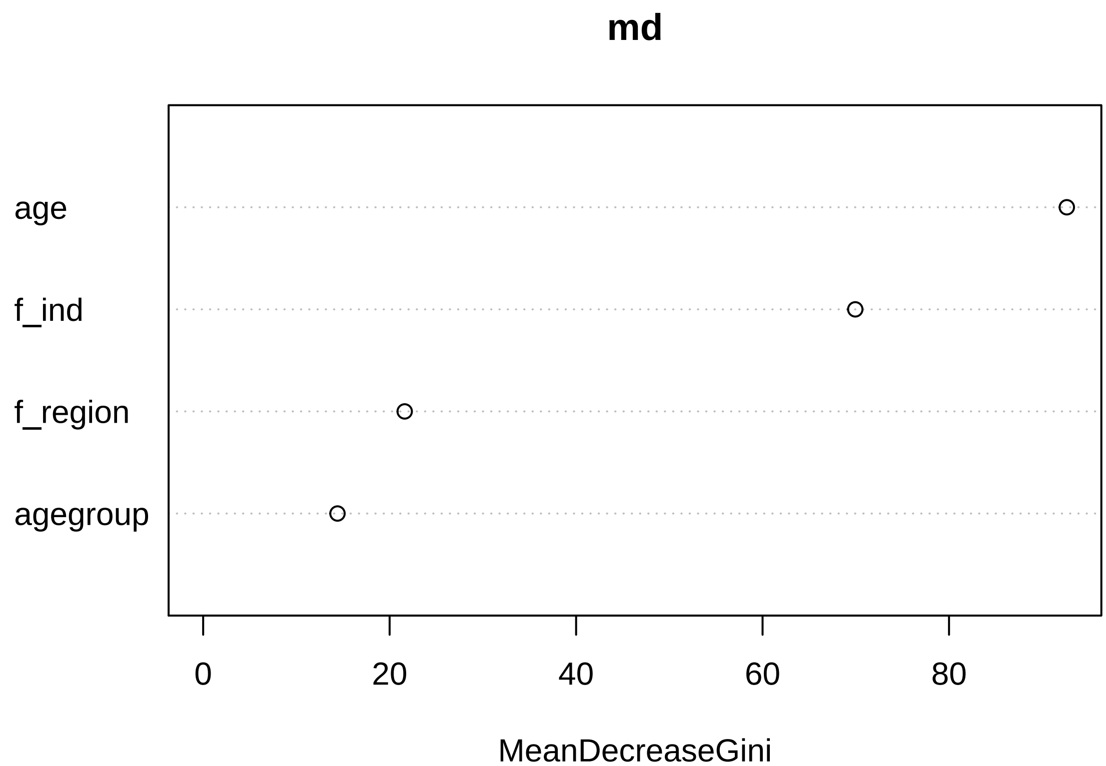



```
##
## Call:
## roc.default(response = data$defaulted, predictor = data$p, plot = T)
##
## Data: data$p in 16450 controls (data$defaulted 0) < 3036 cases (data$defaulted 1).
## Area under the curve: 0.6046
```

```
## $`model 1`
## BRIER_SCORE    ROC_AREA
##   0.1296012   0.6065831
##
## $`model 2`
## BRIER_SCORE    ROC_AREA
##   0.1295537   0.6062904
##
## $`model 3`
## BRIER_SCORE    ROC_AREA
##   0.1278739   0.6402647
##
## $`model 4`
## BRIER_SCORE    ROC_AREA
##   0.1297521   0.6065831
```

```
##
## Call:
##  randomForest(formula = mrf, data = d_train, ntree = 200, mtry = 3)
##                Type of random forest: classification
```

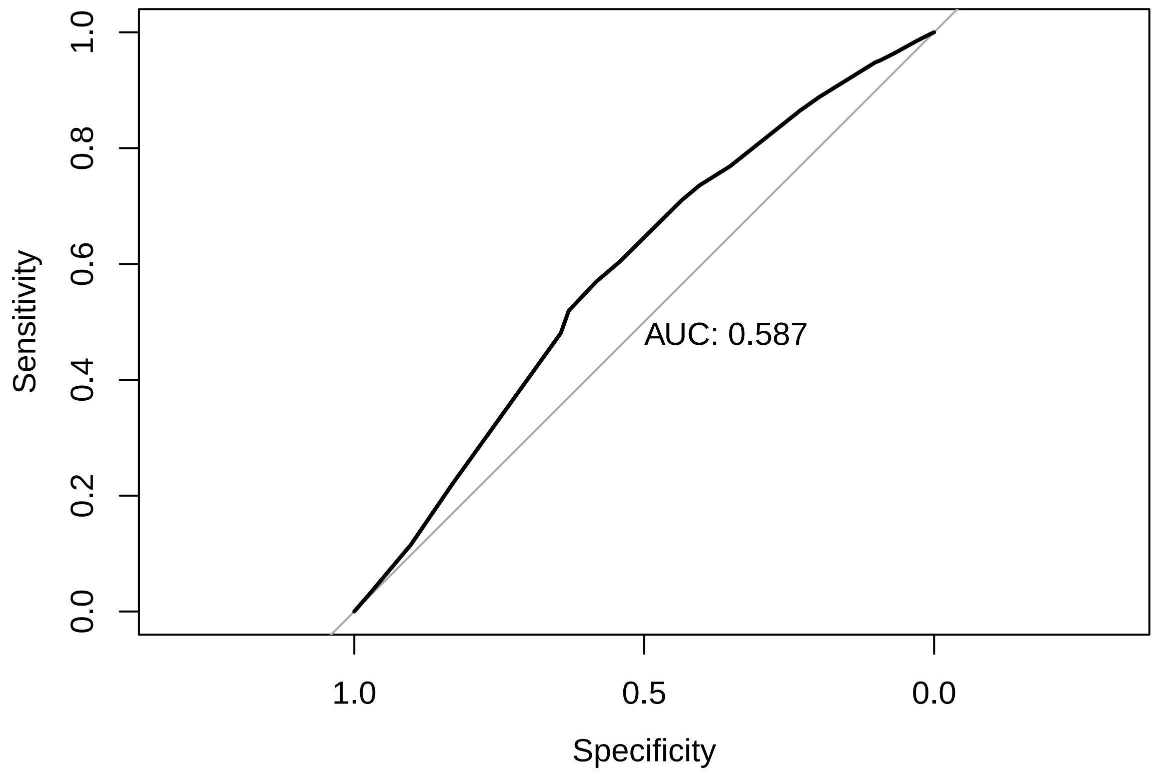
```
##                      Number of trees: 200
## No. of variables tried at each split: 3
##
##         OOB estimate of  error rate: 15.65%
## Confusion matrix:
##       0 1   class.error
## 0 13148 1 0.00007605141
## 1  2439 0 1.00000000000
```



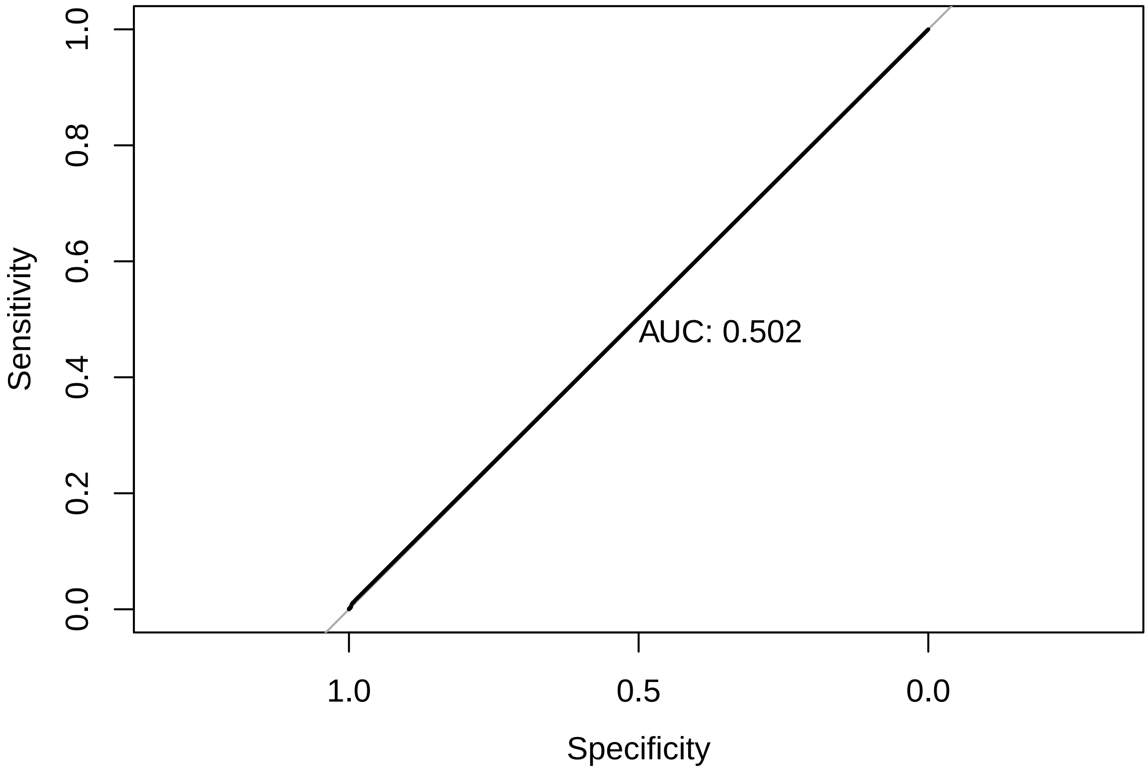

```
##
##         0    1
##   0 2652  446
##   1   10    9
```

```
##
##         0    1
##   0 2658  449
##   1    4    6
```

```
## [1] 0.1281235
```



Sensitivity
1.0
0.8
0.6
0.4
0.2
0.0
AUC: 0.587
1.0
0.5
0.0
Specificity


1.0
0.8
0.6
0.4
0.2
0.0
Sensitivity
AUC: 0.502
1.0
0.5
0.0
Specificity