Decision making for credit card approval

Elena Shevchenko

Data

The data was taken from kaggle datasets https://www.kaggle.com/dansbecker/aer-credit-card-data. The data has 1319 rows and includes the following variables:

card: Dummy variable, 1 if application for credit card accepted, 0 if not
reports: Number of major derogatory reports
age: Age n years plus twelfths of a year
income: Yearly income (divided by 10,000)
share: Ratio of monthly credit card expenditure to yearly income
expenditure: Average monthly credit card expenditure
owner: 1 if owns their home, 0 if rent
selfempl: 1 if self employed, 0 if not
dependents: 1 + number of dependents
months: Months living at current address
majorcards: Number of major credit cards held
active: Number of active credit accounts

EDA

Summary of the data shows that slightly over 20 % of all applications did not receive an approval. *reports* has more than 3rd of its values as zero.

```
 card        reports             age              income            share             expenditure        owner
 no : 296   Min.   : 0.0000   Min.   : 0.1667   Min.   : 0.210   Min.   :0.0001091   Min.   :   0.000   no :738
 yes:1023   1st Qu.: 0.0000   1st Qu.:25.4167   1st Qu.: 2.244   1st Qu.:0.0023159   1st Qu.:   4.583   yes:581
            Median : 0.0000   Median :31.2500   Median : 2.900   Median :0.0388272   Median : 101.298
            Mean   : 0.4564   Mean   :33.2131   Mean   : 3.365   Mean   :0.0687322   Mean   : 185.057
            3rd Qu.: 0.0000   3rd Qu.:39.4167   3rd Qu.: 4.000   3rd Qu.:0.0936168   3rd Qu.: 249.036
            Max.   :14.0000   Max.   :83.5000   Max.   :13.500   Max.   :0.9063205   Max.   :3099.505
 selfemp      dependents         months         majorcards          active
 no :1228   Min.   :0.0000   Min.   :  0.00   Min.   :0.0000   Min.   : 0.000
 yes:  91   1st Qu.:0.0000   1st Qu.: 12.00   1st Qu.:1.0000   1st Qu.: 2.000
            Median :1.0000   Median : 30.00   Median :1.0000   Median : 6.000
            Mean   :0.9939   Mean   : 55.27   Mean   :0.8173   Mean   : 6.997
            3rd Qu.:2.0000   3rd Qu.: 72.00   3rd Qu.:1.0000   3rd Qu.:11.000
            Max.   :6.0000   Max.   :540.00   Max.   :1.0000   Max.   :46.000
```

Figure 1: Summary

Correlation matrix for numerical features is presented below. Features*expenditure* and *share* have high positive correlation coefficient, meaning that there is strong relationship between these variables.

```
                 reports          age      income       share expenditure  dependents      months   majorcards
reports      1.000000000  0.044088513  0.01102287 -0.15901079 -0.13653760  0.01973090  0.04896762 -0.007303561
age          0.044088513  1.000000000  0.32465320 -0.11569704  0.01494770  0.21214643  0.43642554  0.009776687
income       0.011022871  0.324653199  1.00000000 -0.05442926  0.28110402  0.31760130  0.13034627  0.107137782
share       -0.159010789 -0.115697038 -0.05442926  1.00000000  0.83877932 -0.08261776 -0.05534756  0.051469560
expenditure -0.136537597  0.014947698  0.28110402  0.83877932  1.00000000  0.05266406 -0.02900660  0.077513810
dependents   0.019730896  0.212146432  0.31760130 -0.08261776  0.05266406  1.00000000  0.04651197  0.010284541
months       0.048967618  0.436425540  0.13034627 -0.05534756 -0.02900660  0.04651197  1.00000000 -0.041446883
majorcards  -0.007303561  0.009776687  0.10713778  0.05146956  0.07751381  0.01028454 -0.04144688  1.000000000
active       0.207755016  0.181069715  0.18054026 -0.02347440  0.05472424  0.10713276  0.10002764  0.119602777
                 active
reports      0.20775502
age          0.18106971
income       0.18054026
share       -0.02347440
expenditure  0.05472424
dependents   0.10713276
months       0.10002764
majorcards   0.11960278
active       1.00000000
```

Figure 2: Correlation matrix

According to the scatter plot, there are only few applications that did not receive approval and these are with zero *share*. It can potentially lead to the problem in analysis since denials are determined by zero shares only.

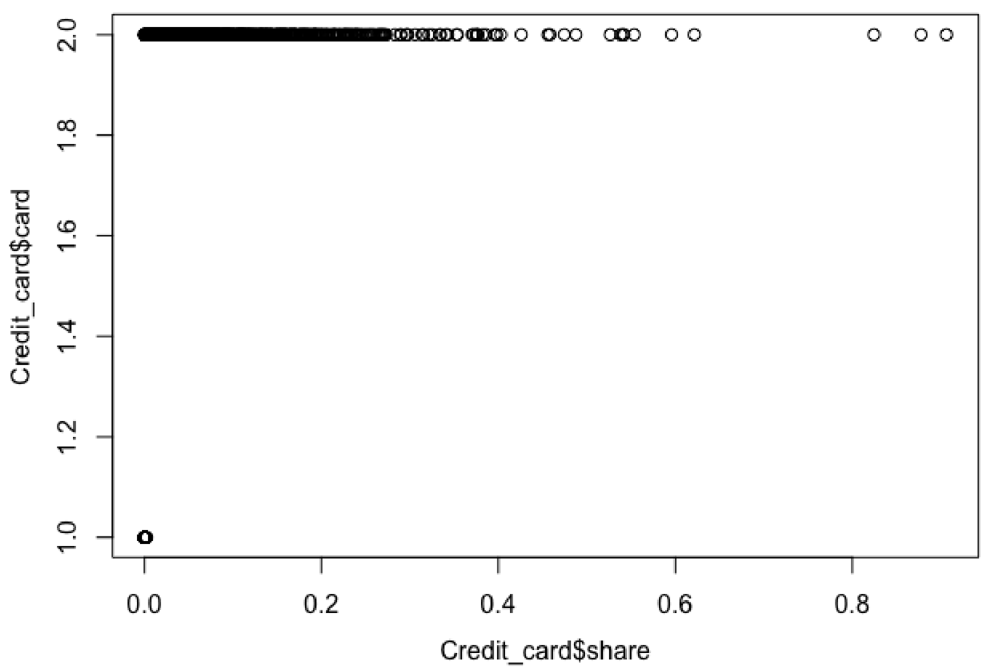

Figure 3: Scatter plot between *share* and *card*

Similar scatter plot is shown below for *expenditure*.

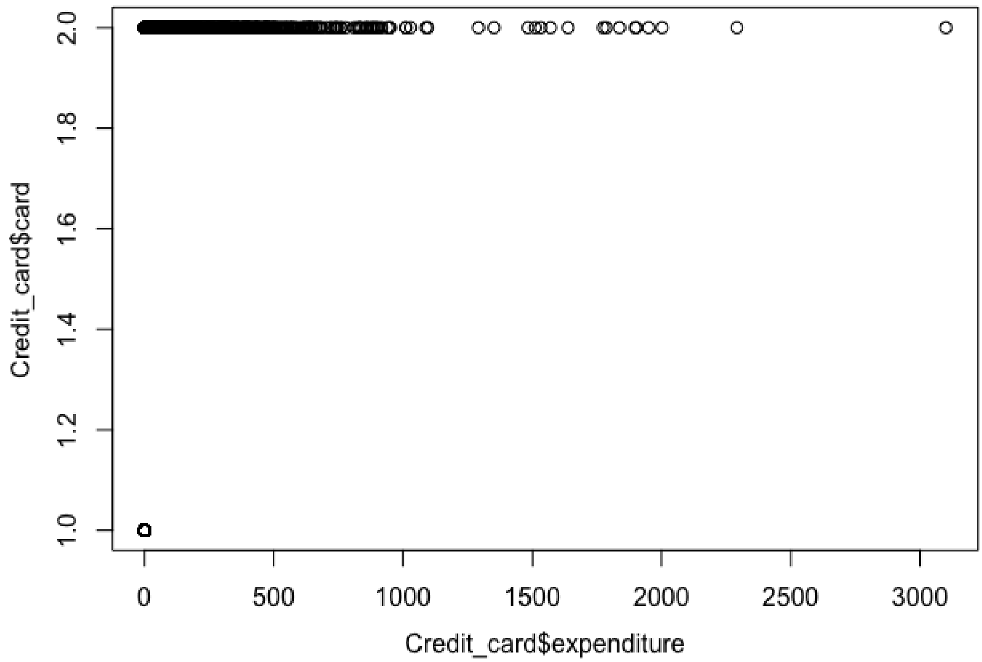

Figure 4: Scatter plot between *expenditure* and *card*

To investigate this situation further it is necessary to create new dummy variable, taking 0 if expenditure is 0 and 1 otherwise. Confusion matrix below shows that all applications with positive expenditure received an approval.

	0	1
no	296	0
yes	21	1002

Figure 5: Confusion matrix for dummy *expenditure* and *card*

Logistic regression

EDA shows that everyone who has positive expenditure receives an approval. Thus, approval can be determined by only *expenditure*. Hence, it is only sufficient to make analysis for those who do not have expenditure as these applications have both approvals and denials. Then there are 2 steps in analysis using logistic regression:

1. Logistic regression for all observations using all features except for *expenditure* and *share*.

2. Logistic regression using only observations whose applications were denied.

1. Logistic regression for all observations

Features *share* and *expenditure* are excluded from logistic model as for these variables denials are determined by zero values only. Almost all coefficients are significant.

```
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.5460   0.1644   0.4045   0.6148   2.8284

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.6272828  0.3286879   1.908 0.056334 .
reports     -1.7516736  0.1410727 -12.417  < 2e-16 ***
age         -0.0125143  0.0095948  -1.304 0.192140
income       0.2262948  0.0642351   3.523 0.000427 ***
owneryes     0.4782723  0.2001855   2.389 0.016888 *
selfempyes  -0.7573433  0.2890634  -2.620 0.008793 **
dependents  -0.2423072  0.0691878  -3.502 0.000461 ***
months       0.0005106  0.0013941   0.366 0.714180
majorcards   0.5053449  0.1907830   2.649 0.008078 **
active       0.1322955  0.0188275   7.027 2.11e-12 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1404.57  on 1318  degrees of freedom
Residual deviance:  980.33  on 1309  degrees of freedom
AIC: 1000.3
```

Figure 6: Logistic regression summary

Results of step-wise selection show that the final model does not include *age* and *months*, which is a consistent result according to the significance of features in the original model.

```
Step:  AIC=998.02
card ~ reports + income + owner + selfemp + dependents + majorcards +
    active

             Df Deviance     AIC
<none>            982.02  998.02
- owner       1   986.91 1000.91
- majorcards  1   988.76 1002.76
- selfemp     1   988.91 1002.91
- income      1   994.73 1008.73
- dependents  1   995.05 1009.05
- active      1  1041.49 1055.49
- reports     1  1340.36 1354.36

Call:  glm(formula = card ~ reports + income + owner + selfemp + dependents +
    majorcards + active, family = binomial, data = Credit_card)

Coefficients:
(Intercept)      reports       income     owneryes    selfempyes    dependents    majorcards       active
     0.3307      -1.7574       0.2124       0.4149       -0.7792       -0.2508        0.5017       0.1317

Degrees of Freedom: 1318 Total (i.e. Null);  1311 Residual
Null Deviance:      1405
Residual Deviance: 982  AIC: 998
```

Figure 7: Step-wise selection

Cross-validation is applied to Lasso, where 2 coefficients are set to zero. The results are the same as for step-wise selection.

```
                      1
(Intercept)  0.67828530
reports     -1.36290282
age          .
income       0.11892330
owneryes     0.29688016
selfempyes  -0.43132249
dependents  -0.12926670
months       .
majorcards   0.35801165
active       0.09229623
```

Figure 8: Lasso coefficients

Finally, predictions are made using the final model. The accuracy of prediction is 86 %. According to ROC curve, optimal threshold is 0.6.

```
Confusion Matrix and Statistics

          Reference
Prediction  no yes
       no  150  44
       yes 146 979

               Accuracy : 0.856
                 95% CI : (0.8358, 0.8745)
    No Information Rate : 0.7756
    P-Value [Acc > NIR] : 1.284e-13
```

Figure 9: Confusion matrix

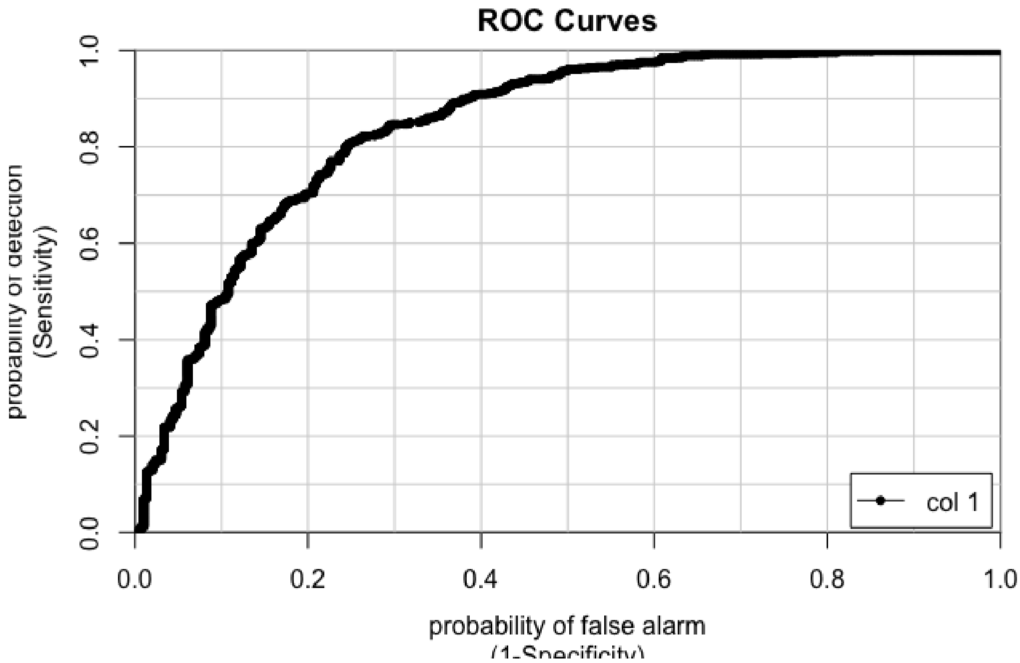

Figure 10: ROC curve

2. Logistic regression for 317 observations

First make a scatter plot for *reports* and *card*. There is similar problem with *reports* as with *expenditure* and *share*. Zero reports almost determine chance of approval for applications without expenditures. Therefore, it is excluded from further analysis.

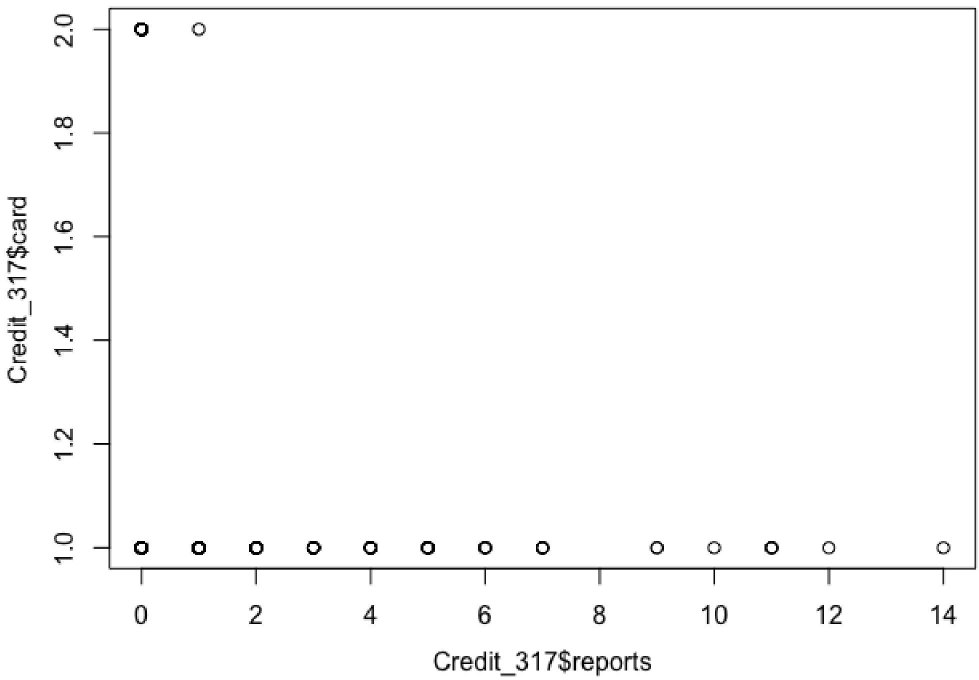

Figure 11: Scatter plot for *reports* and *card*

Only one coefficient for *dependents* is significant. Step-wise model selection leaves only *dependents* variable. However, Lasso method with cross-validation leaves additional features as can be seen below. Hence, the next step is to use decision tree and random forests.

```
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.9181  -0.4282  -0.3054  -0.2054   2.8619

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -3.388424   1.014797  -3.339 0.000841 ***
age          0.028298   0.022556   1.255 0.209640
income      -0.156262   0.187572  -0.833 0.404802
owneryes     0.686951   0.553556   1.241 0.214613
selfempyes   0.457269   0.688749   0.664 0.506747
dependents  -0.722086   0.298267  -2.421 0.015480 *
months      -0.004180   0.004365  -0.958 0.338189
majorcards   0.808769   0.658071   1.229 0.219072
active       0.003020   0.031874   0.095 0.924513
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 154.58  on 316  degrees of freedom
Residual deviance: 141.09  on 308  degrees of freedom
AIC: 159.09
```

Figure 12: Logistic regression summary

```
                      1
(Intercept) -2.65083090
reports     -0.61014546
age          0.01040599
income       .
owneryes     .
selfempyes   .
dependents  -0.37512827
months       .
majorcards   0.30876052
active       0.03576109
```

Figure 13: Lasso results for reduced model

Tree for the whole data set

First, the tree is constructed for all variables including *expenditure* and *share*. The tree is represented below. There are only 3 terminal nodes. According to the tree, card approval is fully determined by 2 variables *expenditure* and *reports*. The structure of the tree support results from previous analysis. Nevertheless, the main variable that determines approval is *expenditure*.

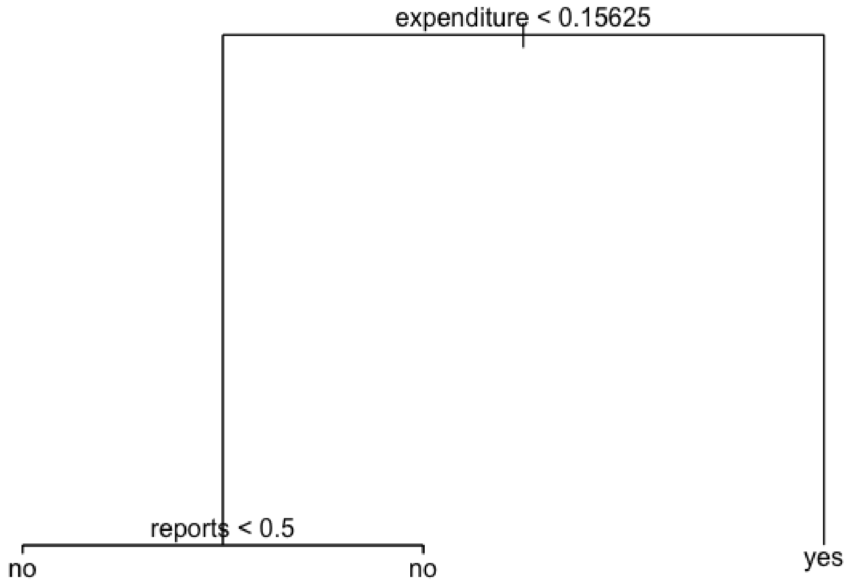

Figure 14: Tree for the full data set

Then the data was divided into training and test sets with 660 cases in training set. Tree was constructed using training set and then predictions were made using test set. The results of prediction are show below, where prediction is accurate in 98%. Pruning does not lead to an improvement even when cross-validation is used.

```
          card_test
tree_pred  no yes
      no  142  12
      yes   1 504
```

Figure 15: Confusion matrix using training and test sets

Tree for the reduced data set

The tree for 317 observations is shown below and is consistent with the results of lasso. The accuracy of the prediction is 93%. Splitting reduced data set into training and set sets does not lead to improvement. Pruning the tree also does not lead to an improvement.

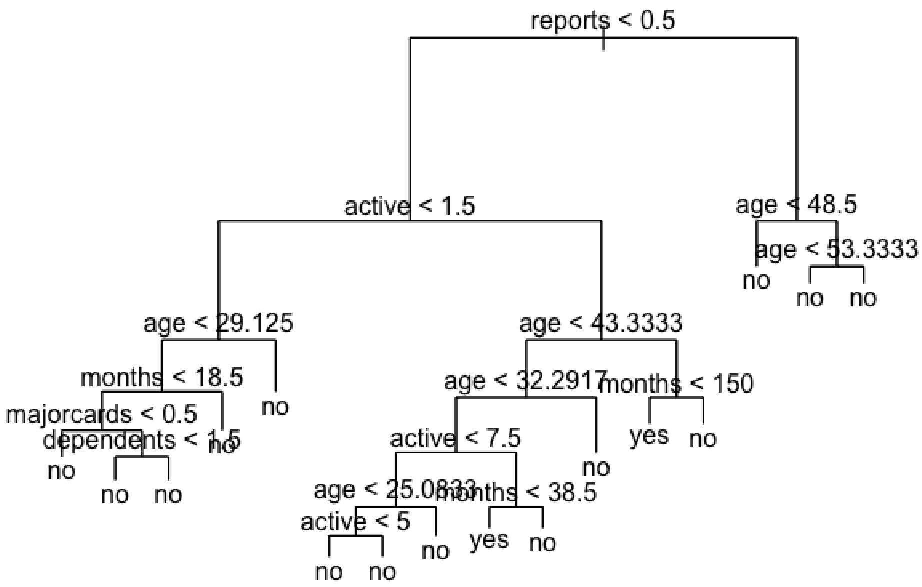

Figure 16: Tree for the reducedl data set

Bagging and random forests

For the original data set bagging with constructing 500 trees using training and test sets leads to 98% accuracy for predictions. Random forests has similar results and does not lead to an improvement.

Similar results hold for reduced data set using 317 observations. Prediction is accurate in 92%.

Conclusion

Credit card approval is one of the most important decision-making tasks for banks. According to analysis, the main features that determine approval are *expenditure* and *share*. Applications that show positive expenditures are certainly to be approved. Among those who do not have expenditures *reports* determine the results of decision. Extra variables, such as *active*, *age*, *months*, *dependents* and *majorcards*, are considered in decision-making based on information of the reports.