Project #1

Credit card clients default analysis.

Link to the source of dataset:
https://archive.ics.uci.edu/ml/datasets/default+of+credit+card+clients

Variables:

X1 (LIMIT_BAL): Amount of the given credit (NT dollar): it includes both the individual consumer credit and his/her family (supplementary) credit.
X2 (SEX): Gender (1 = male; 2 = female).
X3 (EDUCATION): Education (1 = graduate school; 2 = university; 3 = high school; 4 = others).
X4 (MARRIAGE): Marital status (1 = married; 2 = single; 3 = others).
X5 (AGE): Age (year).
X6 - X11 (PAY_#): History of past payment. We tracked the past monthly payment records (from April to September, 2005) as follows: X6 = the repayment status in September, 2005; X7 = the repayment status in August, 2005; . . .;X11 = the repayment status in April, 2005. The measurement scale for the repayment status is: -1 = pay duly; 1 = payment delay for one month; 2 = payment delay for two months; . . .; 8 = payment delay for eight months; 9 = payment delay for nine months and above.
X12-X17 (BILL_AMT#): Amount of bill statement (NT dollar). X12 = amount of bill statement in September, 2005; X13 = amount of bill statement in August, 2005; . . .; X17 = amount of bill statement in April, 2005.
X18-X23 (PAY_AMT#): Amount of previous payment (NT dollar). X18 = amount paid in September, 2005; X19 = amount paid in August, 2005; . . .;X23 = amount paid in April, 2005.

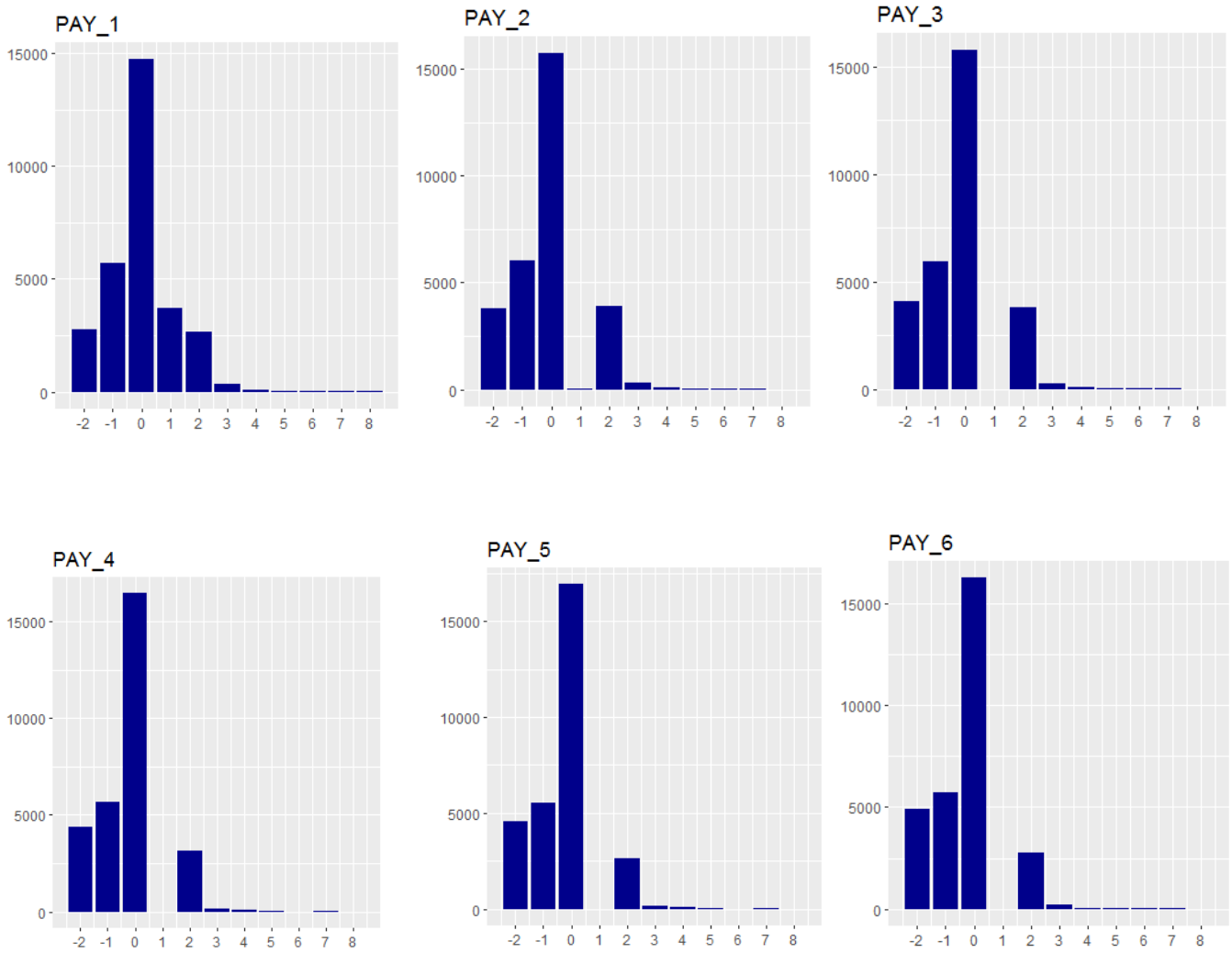
Used methods:

Logistic regression, LDA, Neural Network, Decision Tree

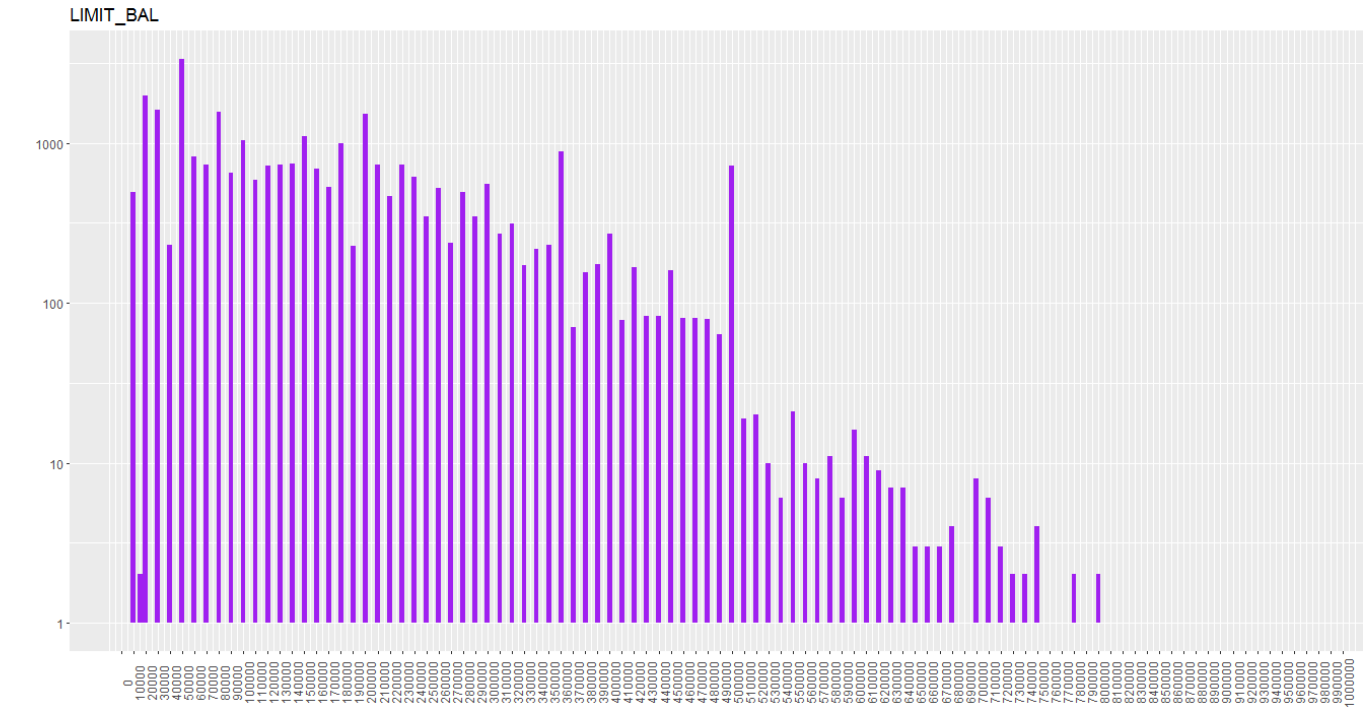
Project author:

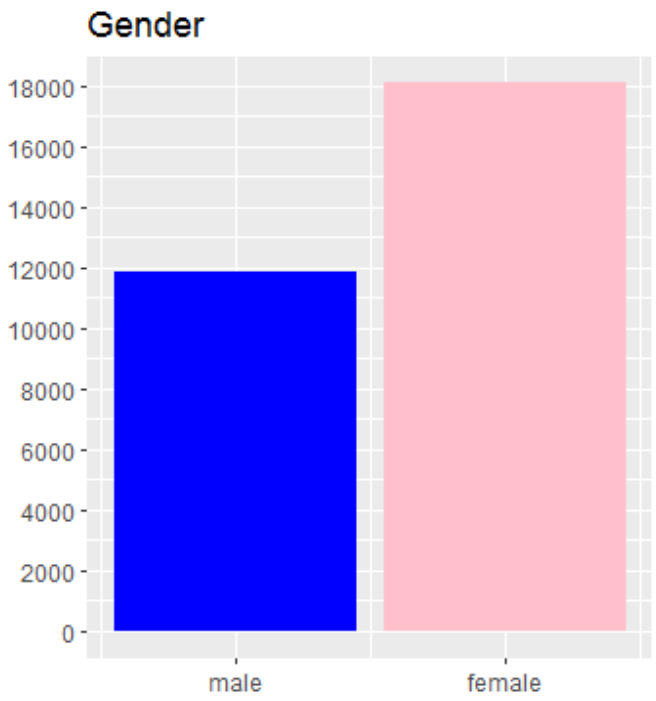
МФН181 Finance, Eremeev Kirill Evgenevich

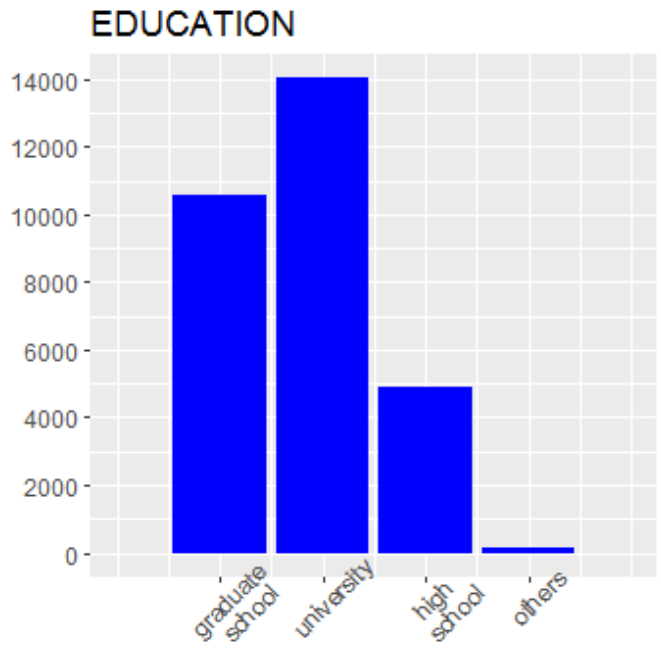
1. Data visualization (data_visualization.R)
   Repayment status

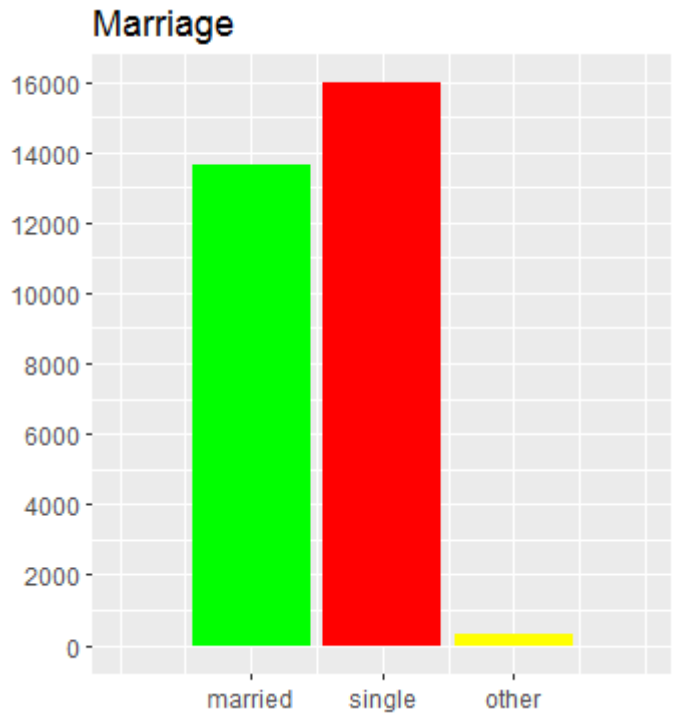


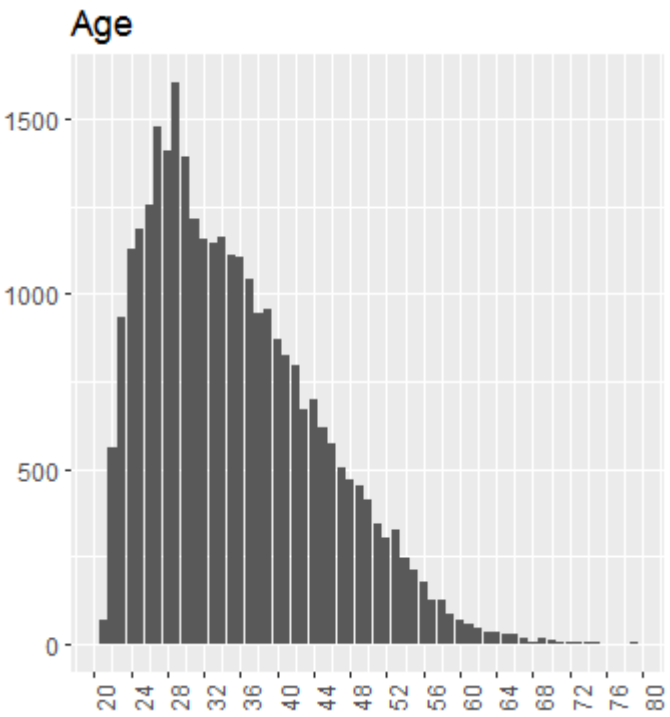
Amount of the given credit (y-axis in log)

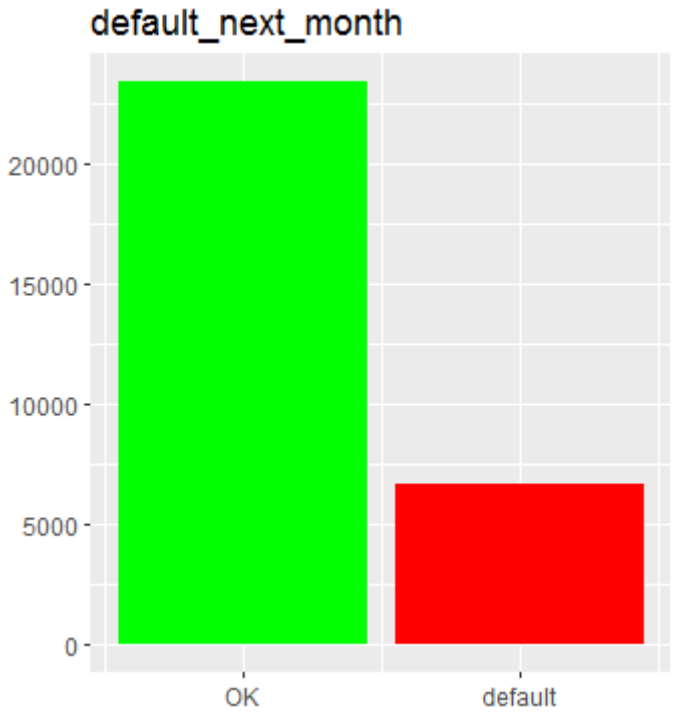



Gender
18000
16000
14000
12000
10000
8000
6000
4000
2000
0
male
female


EDUCATION
14000
12000
10000
8000
6000
4000
2000
0
graduate school
university
high school
others


Marriage
16000
14000
12000
10000
8000
6000
4000
2000
0
married
single
other


Age
1500
1000
500
0
20
24
28
32
36
40
44
48
52
56
60
64
68
72
76
80



Visualization summary:

From PAY_1-PAY_6 graphs we can see that the most of people pay duly and even pay in advance up to 2 months.

From LIMIT_BAL graph we can observe that lower limit balance was offered to more people.

According to gender graph the data contain more females – 60,37%. Also the sample consist mostly of people with higher education ~82%.

Approximately 45,5% of people are married, 53,2% are single. Average age is 35,49.

77,88% of people didn't default.

2. Comparison of different models

   2.1. Logistic regression (classification_methods_logistic_regression.R)

   Firstly I created model (logisticReg.model1) with all dependent variables (23 variables) and checked their significance to result.

   Second model (lr.model) was simplified (10 variables) according to significance of response variables.

   Third model (lr.model2) was created based on hypothesis that we don't need data about previous payments and only can use data like Balance limit, gender, age, education and marriage status.

   Test set accuracy of 3 models provided below:

logisticReg.model1	lr.model	lr.model2
0,808	0,8055	0,779

lr.model2 has sufficient accuracy, but we can't use it due to model doesn't predict default situations.

```
lr.model2 Contingency table (test set):
         Predicted
Observed    0
       0 6804
       1 2051
```

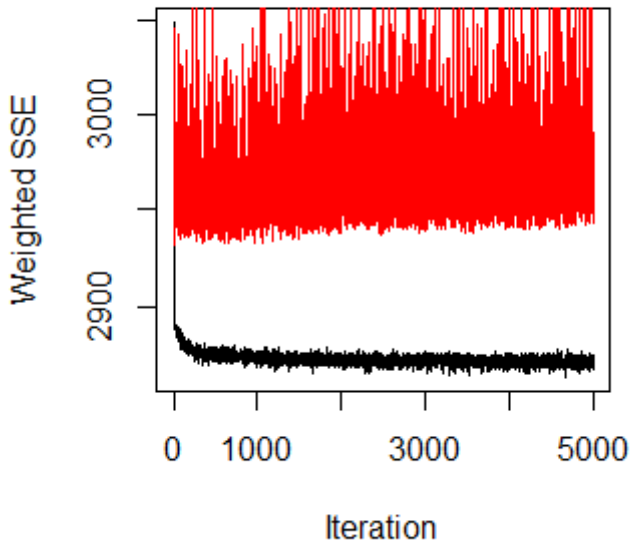
Further I decided to compare models with set of variables like in lr.model (LIMIT_BAL + SEX + EDUCATION + MARRIAGE + AGE + PAY_1 + PAY_2 + BILL_AMT1 + PAY_AMT1 + PAY_AMT2)

2.2. LDA (classification_methods_LDA.R)

Accuracy (test set): 0.8058

2.3. NN (classification_methods_NN.R)

The best result I could achieve with 5 units in the hidden layer was Accuracy (test set): 0.8174.

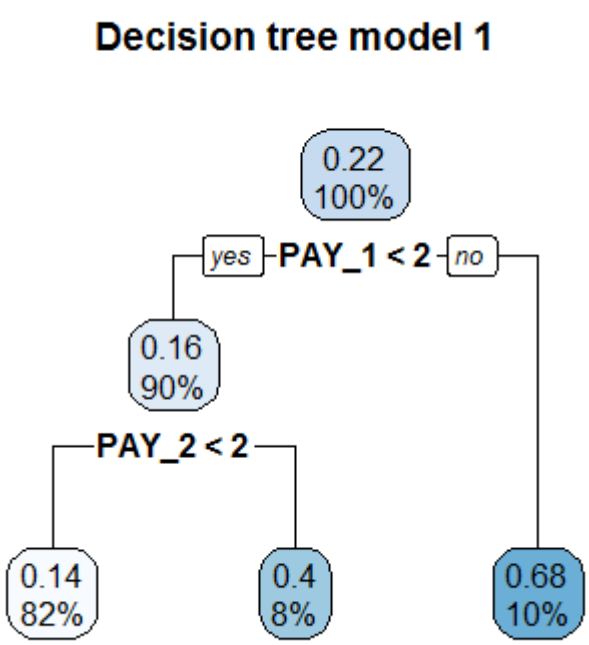


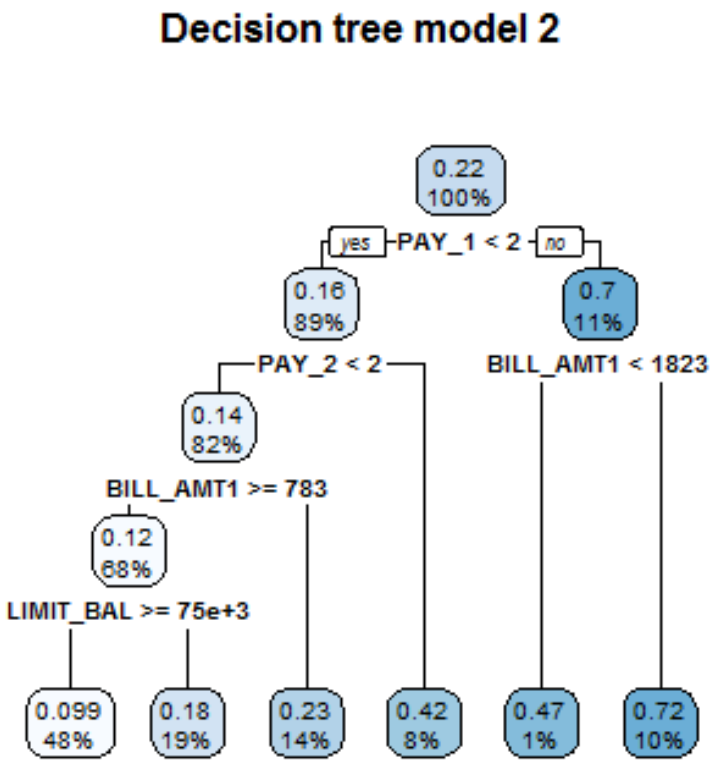
Despite Weighted SSE is relatively high, I can't conclude that neural network structure is too simple because while decreasing or increasing number of units in hidden layer accuracy is decreasing.

2.4. DT (classification_methods_DT.R)

Two models were created, the difference was in minsplit and complexity parameters. Nevertheless the best result was achieved with default model (minsplit and cp).

The structure of first model is very simple.





But Accuracy of first test model: 0.8161 is higher than Accuracy of the second: 0.8142. Also according to contingency table the amount of false negative results is higher in model 2. Which mean that model doesn't predict default, but it happened. In this case it is crucial to reduce false negative results, otherwise bank will lose money.

3. Summary

Method / Parameter	LR	LDA	NN	DT
Accuracy	0,8055	0,8058	0,8174	0,8161
False negative results	1523	1485	1317	1351

Among all used methods NN was the most accurate and had the minimal false positive results, therefore it could be used to predict default with such data provided.

Speaking from the perspective of bank and risk management we can improve our model using probability of default rather than discrete results (1 – default/0 – not default). Because we can't maximize profit with false positive result (model predict default, but it isn't true), for instance NN has 310 false predicted defaults.