Group Work Handout - Credit Scoring with R

SBD2 - Data-driven Visualization and Decision Making

October 2, 2023

The main objective of the group work assignment is to produce a R markdown file on credit scoring using a data set uploaded on Moodle.

Note: Each group works with a different version of the dataset. Namely, each group should work with the dataset that corresponds to their group number in Moodle. For example: group 1 should work with the dataset "loan_sample_1".

Throughout the script, you should use comments to include descriptions/interpretation of results as well as answers to specific questions. The target variable of the analysis is the "Status" feature included in the dataset which contains the values 0 if the loan contract did not default and 1 if the loan contract defaulted. The description of the other variables included in the data set is provided in Table 1.

Exercise 1

Using the data set "loan_data.csv", please go through the following tasks:

- Describe the data. Specifically:
  - Check and report the structure of the data set.
  - How many numeric and how many categorical variables are included in the data? What categorical variable has the most levels in it?
  - Summarize the variables. Discuss the summary statistics obtained.
  - Check the levels of the target variable by choosing the appropriate visualization. Is the target variable balanced?
  - Check the distribution of the numeric variables in the data set (include different visual representations).
- Investigate whether certain variables contain outliers (hint: what does a box plot show?). Elaborate your view on how to proceed in dealing with the outliers and – if necessary – take appropriate action.
- Choose the appropriate visualization to investigate the distribution of the numeric features per the two levels of our target feature (i.e. default vs non-default). Discuss the visualizations. Which variables seem to be relevant in predicting the target feature?

Variable	Description
annual_inc	The self-reported annual income provided by the borrower during registration.
application_type	Indicates whether the loan is an individual or joint application
dti	Borrower's total monthly debt payments divided by monthly income.
grade	Assigned loan grade by the financial service provider
home_ownership	The home ownership status
int_rate	Interest Rate on the loan
loan_amnt	The listed amount of the loan applied for by the borrower
open_acc	Number of open trades in last 6 months
purpose	A category provided by the borrower for the loan request.
revol_bal	Total credit revolving balance
revol_util	Revolving line utilization rate.
tot_cur_bal	Total current balance of all accounts
total_acc	The total number of credit lines currently in the borrower's credit file
total_rec_int	Interest received to date
total_rev_hi_lim	Total revolving high credit/credit limit.
verification_status	Indicates if the co-borrowers' joint income was verified

Table 1: Variables' description

- Use a bar plot visualization to investigate the associations between the categorical variables and the target feature.
- Visualize the correlations that emerge between the numerical features. Discuss the results. Which variables are highly correlated? Decide whether you keep all variables.
- Plot an interactive scatter plot of the association between the loan amount requested and the annual income of the borrower. Discuss the plot. What can you tell about the association?
- Create a new balanced data set where the two levels of the target variable will be equally represented; Create a bar plot of the newly created target variable. Why is this step necessary?

Exercise 2

Using the new balanced data set:

- Train and test a logistic classifier. Specifically:
  - Divide the sample into training and testing set using 70% for training the algorithm.
  - Train the classifier and report the coefficients obtained and interpret the results.
  - Plot the ROC and the Precision/Recall Curve and interpret the results.
  - Produce the confusion matrix and interpret the results.
  - Report the AUC values and the overall accuracy and interpret the results.

Exercise 3

Thinking about the pre-processing steps that you carried out before training the logistic classifier:

- Can you think of a way to improve the predictive performance of your data?
- What can you do differently? (hint: Feel free to be creative and discuss any additional step in data collection and/or data pre-processing that you might try so to improve the results)

Exercise 4

Finally, thinking about putting your model into action and basing credit decisions on the prediction that it generates:

- What kind of challenges may a company face if it would use your model in their daily business, in particular in regard to ethical challenges and moral obligations companies have? Please refer to the „common ethical issues in the context the creation of value from data" (see slides week 11) in your answer.
- Can you think of a way how companies can overcome or at least mitigate the issues that you described above?