RDS Project Draft Report

Tom Yang yy2949, Lyris Wei yw4182

Background

The ADS we propose to analyze in the project is Nicholas's solution for Give Me Some Credit. We found this project in Kaggle competitions. The host has published the training set, test set, sample entry and submission files. Nicholas's solution is the second solution under the Most Comments filter. All his code is on his github and notebook with explicit explanation and running results.

The goal for this ADS is to help borrowers make better financial decisions in loan granting by improving a credit scoring system that makes a guess at the chance of default (delinquency). Participants need to build a model to predict the probability that individuals may have financial problems in the next two years. From Nocholas's github, we could know that he trained a XGBoost model on the dataset, which is able to attain private and public AUC scores of 0.86756 and 0.86104 respectively. One thing we noticed in Nocholas's ADS is that he mentioned younger people are more likely to default. Therefore we want to figure out if his model determines that age is a significant feature in this credit scoring system, and if there are more fairness issues in his model.

Input and Output

a. About the source of the data, it is provided by sponsors who held this Kaggle competition. In the data description, they didn't explicitly mention who collected the data and how, but vaguely said the data are from 250,000 borrowers. It is likely that the data comes from the bank system.

b. Next we look through the features one by one. The table has 11 columns in total, including 10 features and 1 predictor (y variable).

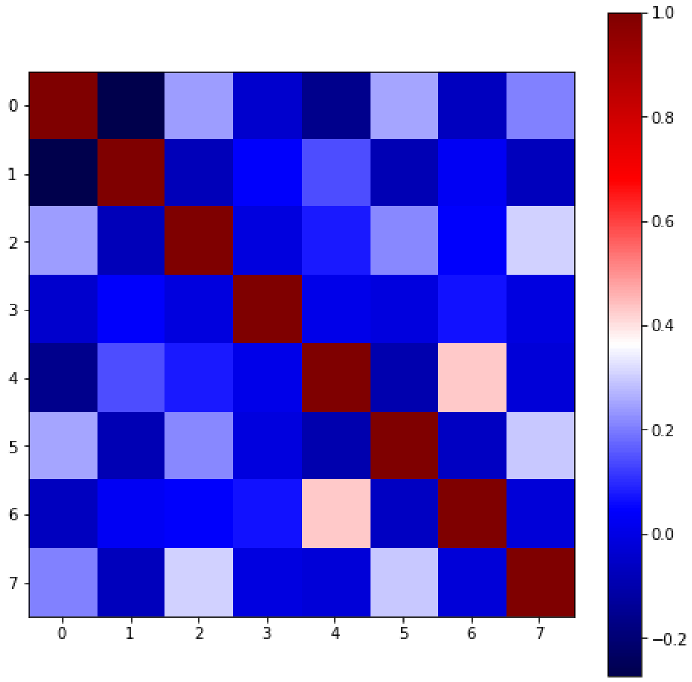
A comparison of all the data is shown in the table below. We draw the distribution of each feature in the notebook. But to make the report look simpler, we collect the data from `df.describe()`, and summarize them in the table below.

Feature Name	**Description**	**Feature Type**	**Distribution**	**Notes**
SeriousDlqin2yrs	Person experienced 90 days past due delinquency or worse	Boolean (1 or 0)	0 (No delinquency):1(experienced delinquency) = 94:6	Predicting Variable
RevolvingUtilizationOfUnsecuredLines	Total balance on credit cards and personal lines of credit except real estate and no installment debt	float	min: 0.0 50%: 0.15 max: 50708 mean: 6.05 std: 249.75	No missing value
Age	Age of borrower in years	integer	An approximately normal distribution for all and for those creditable people; multimodal for those that are not creditable	Can find the graph of each distribution in the code; No missing value
NumberOfTime30-59DaysPastDueNotWorse	Number of times the borrower has been 30-59 days past due but no worse in the last 2 years.	integer	120k are 0 among the 150k lines of data min: 0 max: 98	No missing value
NumberOfTime60-89DaysPastDueNotWorse	Number of times the borrower has been 60-89 days past due but no worse in the last 2 years.	integer	140k are 0 among the 150k lines of data min: 0 max: 98	No missing value
NumberOfTimes90DaysLate	Number of times the borrower has been 90 days or more past due.	integer	140k are 0 among the 150k lines of data min: 0 max: 98	No missing value
DebtRatio	Monthly debt payments, alimony,living	float	mean:353.01 std: 2037.82 min: 0	No missing value

	costs divided by monthly gross income		50%: 0.37 max: 329664	
MonthlyIncome	Monthly Income	float	60% lower than mean. Also very imbalanced.	19.8% percent missing values.
NumberOfOpenCreditLinesAndLoans	Number of Open loans (installment like car loan or mortgage) and Lines of credit (e.g. credit cards)	integer	mean: 8.452760 std: 5.145951 min: 0 median: 8 max: 58	No missing value
NumberRealEstateLoansOrLines	Number of mortgage and real estate loans including home equity lines of credit	integer	mean: 1.01 std: 1.54 min: 0 median: 1 max: 54	No missing value
NumberOfDependents	Number of dependents in a family excluding themselves (spouse, children etc.)	integer	mean: 0.76 std: 1.12 min:0 median: 0 max: 20	2% missing value

Table 1. Data Description

For correlation between each feature, we draw a correlation matrix for every feature that has no missing values. The graph is shown below.



The fourth row, representing `NumberOfOpenCreditLinesAndLoans`, and the sixth row, representing `NumberRealEstateLoansOrLines`, has a rather high correlation (0.43). All other features' correlation are lower than 0.25. So we might need a model more complex than linear regression to make a good prediction.

c. The predictor is the first column (first row of the table above), it represents whether a borrower has enough credits (to be trusted). The model predicts this value as a probability first, and then assigns label 1 if probability is greater or equal to 0.5, and label 0 if smaller than 0.5.

A full code can be found in the two links below: data analysis, and ADS.

Plans for following sections

For part 3 implementation and validation, the original author Nicholas has already presented detailed EDA and preprocessing. In fact, he got private and public AUC scores of 0.86756 and 0.86104, which are relatively high AUC compared to other participants.Tom will try to rerun his code and write the report for this part, and Lyris will be reviewing the works. This part will be finished in the next week.

For part 4 outcome, we are going to analyze the effectiveness of the ADS by comparing its performance across different subpopulations, quantify one fairness measure of this ADS, and finally develop additional methods for analyzing ADS performance. Lyris will mainly work on this part next week, and Tom will review the work.

In the last week, we will finish writing the report and do some proofreading. If time permits, we would like to schedule an appointment with the instructor to discuss our report and see what we can improve. We expect to finish part 3, 4, and 5 before the end of April, and then seek feedback or improvement from Professor Wood and TAs.