UniMi DSE Credimi Challenge Lab - Shihab Hamati (Matricola: 985941)

submitted 16-May-2022 by email and through OpenDataPlayground, score 302000

Exploring the Training Data

The data suffered from missing values that I categorized into 2 sections. More than a third of the companies were missing financial (numeric) features. For those, I chose not to drop the rows since they are a significant chunk of the data and might contain useful information through the other columns. As such, the missing values were imputed by the median (I also tried mean and settled on the median due to a slightly better model performance).

The 2nd section were only 8 rows that were missing industry and the scores. I chose to drop these rows in the model training since all my attempted imputations worsened the model performance (which is possibly due to the high correlation and importance of the score features so filling inaccurate estimates affect the model strongly).

On separate and messier notebooks, I started exploring both the numeric and categorical data in relation to the target variable. For example, the target variable was not roughly distributed on opposite sides for any of the numeric variables.

The variables did not produce easily recognizable patterns to my eyes (hence the strong reliance on machine learning approaches moving forward).

Feature Engineering

I tried to engineer new features based on some online reading into the financial field. The 2 engineered features do not necessarily correspond strictly to established financial ratios. I tried over a dozen combinations and mathematical operations, and finally selected for the 2 used in the final code based on a balance between the highest correlation with the target variable and least correlation with existing features (to avoid introducing or worsening collinearity).

I was curious to explore if combinations of categorical features would present correlation with the target feature (and some slightly did). However, using a tree-based classifier later would discover such insights so I did not create new features as new columns.

I also scaled the numeric features using a PowerTransformer rather than the standard MinMaxScaler to better handle the outliers (this improved the outcome slightly). I avoided using the StandardScaler since the features were not close to a normal distribution.

Model

I arrived at the model's parameters through trial and error while keeping an eye at both the confusion matrix and the total profit function. The reason I favored using an LGBM classifier is because of its renowned power, its ability to handle categorical data effectively, and its reliance on ensemble techniques.

Insights

None of the classical financial features ranked high among the important variables for my model (or in the correlation during the exploratory phase). This would seem to indicate that the developed 'scoring' metrics (along with non-financial features like industry, dossier type and application source) are much more effective at predicting loan repayment of Credimi's lenders than the classical accounting ratios.

Test Predictions

The same preprocessing was applied to the test data. There was a leakage from the training data into the testing data in the area of imputation of missing values (i.e., filling with the training data median). This is a tolerable data leakage as in real life Credimi already has knowledge about its historical data. A practical implication of this would be the need to update the medians of the concerned columns regularly (which is not difficult given the relatively manageable scale of the data).

About the situation of missing industry and scoring data for a few rows, they were given directly a target prediction of 1 since the model could not have enough data on these very important features. This is the majority outcome of the target variable for the 8 similar rows in the training set (5/8 defaulted).