Project: Loan Prediction

Ref. <https://datahack.analyticsvidhya.com/contest/practice-problem-loan-prediction-iii/#DiscussTab>

Introduction

This project is based on the competition Analytics Vidhya [1] had published about Loans from the company “Dream Housing Finance”, which deals with all kind of home loans spread across urban a rural area. The data that had been collected has some features related to demographic information about clients and some of their financial data. The dataset is a small one, around 600 rows, and also the number of features is small, just around 12 features without the dataset id and the target.

The purpose of the competition is to identify loan eligibility based on the mentioned data. The result of the project was uploaded in order to score the predictions, which was 0.7986, which is a good result.

Dataset description

The data for the training has 614 rows and 12 columns without the target. The description of each column is found in [1]. The name of the target is “Loan\_Status” that has two possibilities “Yes” and “No”.

Exploratory Data Analysis

The exploratory data analysis was performed by analyzing the context of the problem and the characteristics of the variables as observed in the project directory, in this case the file “Exploration.ipynb” In the directory named “notebooks”.

The target variable is *Loan\_Status* which has just to outcomes “Yes” an “No”, and is not balanced, that is why it had been used a process specific to balance the data. Also in some cases the data has *NaN* values for which, depending on the case, they had been imputed.

The strategy followed was mainly to transform the categorical variables with zero a one only if there were only two possible values. In other cases, the strategy was to apply one hot encoding to process them. These features were processed easily with very simple transformation. The categorical features processed in the mentioned way were:

*Gender*

*Married*

*Dependents*

*Education*

*Self\_Employed*

In the case of numeric features, the story was different because most of them were skewed and in some cases with lots of *NaN* values. In this case, there were no clue in the way to process them, that is why the strategy in this case was a little more complicated because it has been used various ways to try to get the best possible results.

*ApplicantIncome*

*CoapplicantIncome*

*LoanAmount*

*Loan\_Amount\_Term*

*Credit\_History*

One of the most important problems found in the dataset was that the target feature was unbalanced, that is why a balanced technique had been used, which helped a lot in the accuracy calculation.

Evaluation metrics

The competition metric for evaluation was “accuracy”. That is why the dataset had been balanced to get good results with this metric.

Imbalance data

One of the most important problems found in this dataset was the imbalance data fount in the target feature *Loan\_Status* specifically in the case of the performance metric that was “Accuracy”, in this particular case, the metric needs to have a balanced target. That is why it had been used the function SMOTE which balances the target values in order to have the same quantities in every target feature value. In this case it had been used oversampling for the processing.

Bibliography

[1] Dream Housing Finance & Analytica Vidhya: Loan Prediction, <https://datahack.analyticsvidhya.com/contest/practice-problem-loan-prediction-iii/#DiscussTab>