Title

Should This Loan be Approved or Denied

Introduction

The dataset is initially from the U.S. Small Business Administration (SBA) while I have acquired it from https://www.kaggle.com/mirbektoktogaraev/should-this-loan-be-approved-or-denied. Starting sample has 27 columns and 899164 rows. The software used was Python and all needed libraries are mentioned in the *requirements.txt* file. Previous, and all yet-to-be-mentioned files are all located in *https://github.com/Vanja2610/should-this-loan-be-approved-or-denied* repository.

SBA is a United States government agency that acts much like an insurance provider to reduce the risk for a bank by taking on some of the risks by guaranteeing a portion of the loan. Since SBA loans only guarantee a portion of the entire loan balance, banks will incur some losses if a small business defaults on its SBA-guaranteed loan. The aim of this project is to predict the probability for the company to default on a loan. The project is structured in three parts: Exploratory data analysis, model building, and model deployment; in the reading below, more detail will be presented.

Exploratory Data Analysis (EDA)

In my humble opinion, EDA is the heart of the project. The main goal in this part was to understand the data, clean it and decide on the predictors that should be used in the upcoming modeling. For a comprehensive description of the variables please check the *Should This Loan be Approved or Denied.pdf* document. A keep-in-mind fact is that the project was approached from a loan officer perspective in deciding whether to grant a loan to a business; this played a significant role in the variable selection. In other words, predictors that would not be available or would not contribute in any way to the officer's decision were discarded. On the other hand, variables that seemed logical to include and would probably be asked by the same officer were prioritized.

Before filtering the variables, it was necessary to deal with the missing or bad data, do feature engineering, and do some basic plotting and calculations for better interpretation. Processed data has 897167 rows, with six predictors and one dependent variable.

Code along with the step-by-step description of the procedure and comments can be found in **EDA.ipynb** file.

Model Building

The idea here was to decide between two classification algorithms: Random Forest (RF) and Extreme Gradient Boost (XGB) classifiers. Since our dependent variable is skewed, instead of

AUC-ROC curve, f1 score was used for the valuation. The first step for both of the algorithms was to find a range of optimal parameter values and then use a grid-search for finding the exact hyperparameters that should be used. In the case of XGB, the default parameters already gave more than satisfying train and test scores, hence, my goal was to try improving both of them by few percentages. On the other hand, RF has shown slight overfitting as well as a lack of speed. In order to deal with the issues, I have tried improving the test score with the price of reducing train one and simplifying the RF algorithm.

After hyperparameter tuning, both models provide us with almost the same test score, XGB: 96.9 %, RF: 96.7 %. Out of the two, I have gone with the XGB and the argument for that is that the XGB algorithm is still a bit faster which might prove to be useful for even larger datasets. The process is documented in **classification.py** file.

Model Deployment

Finishing touch was deploying the model. For the cloud platform I have decide on the Heroku and as for the needed API, Flask has served the purpose. Please find all necessary files in above mentioned repository.

Acknowledgments

I would like to thank Mirbek Toktogaraev (https://www.kaggle.com/mirbektoktogaraev) for uploading data along with the Should This Loan be Approved or Denied.pdf document on the Kaggle platform.

Furthermore, I would like to give credit to my friend Goran Vezmar () who is responsible for css code needed for the project as well as providing me help with the html code.

Application Url Address

https://sba-loan-approval.herokuapp.com/

Contact

For any questions or discussions, please feel free to contact me on:

✉ vanjaandrejev@yahoo.com

in