Project Proposal

Loan approval prediction for the U.S. Small Business Administration: Performance analysis of different machine learning algorithms

The U.S. Small Business Administration (SBA) was founded in 1953 on the principle of promoting and assisting small enterprises in the U.S. credit market. Fostering small business formation and growth has had social benefits in the United States by creating job opportunities and reducing unemployment. Therefore, one way SBA assists these small business enterprises is through a loan guarantee program which is designed to encourage banks to grant loans to small businesses. SBA acts much like an insurance provider to reduce the risk for a bank by taking on some of the risk through guaranteeing a portion of the loan. In the case that a loan goes into default, SBA then covers the amount they guaranteed (Li et al., 2018).

Although there have been many success stories of start-ups receiving SBA loan guarantees, there have also been some start-ups that have defaulted on their loans. 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. Because of the high risk of default, banks are still faced with a difficult choice as to whether such a loan should be granted. One way to inform their decision making is through analyzing relevant historical data, such as the datasets provided in the article, and use data science and machine learning techniques to build efficient and robust models that can predict whether a small business will default on its loan.

This data science project, therefore, addresses the following questions:

- How can machine learning help in making decisions about whether a loan should be approved or denied by the U.S. SBA?
- How do different machine learning algorithms compare in terms of predictive ability, model complexity and computational intensity?
- What are the important data features that affect a model's performance?
- If necessary, what other data should be collected to improve a model's performance and help in accurate decision making?

Criteria for an optimal machine learning model: a simple model that makes accurate predictions of whether a small business shall default its loan or not; is fast to train and make predictions; and is robust to new data. The dataset used in the project is a real dataset from the U.S. Small Business Administration (SBA). The dataset is available here: https://doi.org/10.1080/10691898.2018.1434342.

The project is organized as follows:

1. Data wrangling (`StepOne_DataWrangling.ipynb`)

2. Exploratory data analysis and preprocessing (`StepTwo_EDA_Preprocessing.ipynb`)
3. Training and modeling (`StepThree_Training_Modeling.ipynb`)
4. Project report
5. Project presentation

References

Min Li, Amy Mickel & Stanley Taylor (2018) *"Should This Loan be Approved or Denied?": A Large Dataset with Class Assignment Guidelines*, Journal of Statistics Education, 26:1, 55-66, DOI: 10.1080/10691898.2018.1434342.