Loan Data Analysis

EE 660 Course Project

Project Type (1) Design a system based on real-world data

Number of student authors: 1

Weiding Huang

[weidingh@usc.edu](mailto:weidingh@usc.edu)

03/12/2020

1. Abstract
2. Introduction
   1. Problem Type, Statement and Goals

Whether to issue a loan to a person requires a lot of investigation and time, and it will cause losses to the loaner if a lender is behind the due and the debt must be charged off. Typically, banks spend lots of resources and time to review loan applications. In order to help banks to expedite the process, this project is aimed to build a model and classify whether a loan is going to be paid off or charged off. The data set used in the project contains 18 feature spaces and 10000 data points. Among these 18 features, 8 features are categorical data and 6 features contain missing data points. Due to the large data quantity, mixing of categorical and numeric data and many missing data points, a comprehensive preprocessing step needs to be conducted in order to get the best result.

* 1. Literature Review

The project was previously done by Victor Hugo Pereira on Kaggle (<https://www.kaggle.com/panamby/bank-loan-status-dataset>). He filled in missing data points using the mean values, and then removed collinear features. For classification models, he used logistic regression, k-nearest neighbors, support vector machine, naïve Bayes, gradient boosting, and random forest. In this project, the best model was gradient boosting, resulting in an 81.7% accuracy on test set.

* 1. Our Prior and Related Work

None

* 1. Overview of Our Approach

In the project,

1. Implementation
2. Final Results and Interpretation
3. Contributions of each team member
4. Summary and conclusion
5. References