CHE1147H – Data Mining in Engineering

Project Progress Report

Gabriel Wong

ID: 1006593018

Project Description

LendingClub loan data set and its' feature descriptions are provided by Wendy Kan from Kaggle, which is an online data science community platform. The data set contains the payment information of clients, ranging from 2017 to 2018. There are 74 features and 887,379 samples, with 17,998,490 missing data points. 49 features (66%) are float64 types, 23 features (31%) are object types, and 2 features (3%) are int64 types. With regards to the missing data, removing features with 80% or more data missing would reduce 90% of total missing data without a significant loss in data quality. The remainder may be addressed by filling them with either empty strings, maximum or minimum feature values depending on their context while erring on the conservative side. For example, missing values in feature *emp_length*, which describes the employment length in years, should be replaced with minimum values, assuming that the applicants have never worked. The prediction target will be the loan status of applicants, i.e., good loan (fully paid) versus bad loan (default, charged off, and late), which is a binary classification problem. For this problem set, it would be appropriate to use and compare multiple supervised machine learning algorithms in the classification sub-domain such as logistic regression, decision trees, ensemble learning or random forest. Model evaluation, feature importance, and fine-tuning can be performed with the help of metrics such as ROC/AUC curves, precision, recall, and confusion matrix.

Executive Summary

LendingClub is a San Francisco-based company currently operating one of the largest lending marketplace platforms in the United States, connecting borrowers and investors. Borrowers access installment loans such as personal, business, patient, and auto refinancing, through an online and mobile interface. Investors provide capital to fund loans in exchange for the opportunity to earn attractive returns. The company generates most of its revenue from transaction fees received from providing loan approval service to customers on behalf of the bank partners to enable loan originations.

LendingClub employs proprietary algorithms to assess the risk profile of borrowers by leveraging data such as behavioral, transactional, bank, and employment history. This reduces the borrower's credit cost and shortens the loan approval process; thereby improving customer experience. The algorithms, therefore, have to be accurate and reliable, as undetected errors may cause loan mispricing, erroneous loan approvals or denials, which may harm the company's reputation and its trust with customers.

The objective is to develop a supervised machine learning algorithm to predict whether a borrower is qualified to receive a loan or not, with high accuracy and robustness, to minimize forecasted losses. The data set selected for modeling comprises of 74 features involving client information and payment histories, along with approximately 900,000 observations from 2017 to 2018. The ratio of numerical to categorical variables is approximately 2:1. The target output will be applicant loan status, which is either a good loan or a bad loan; hence, a binary classification problem.