Credit Risk Analysis in Social Lending Ecosystem

Aveek Choudhury - Harshita Ved - Sagar Singh - Sarang Pande

A key issue of loan origination in the financial services industry is balancing credit risk while increasing credit accessibility. Credit risk is the risk of default as a result of borrowers failing to make required payments, leading to loss of principal and interest amounts. All lenders, especially lending investors in a peer to peer setting with a high level of information asymmetry, must carefully consider credit risk when making informed investment decisions. In this project, we aim to analyze historical loan data of LendingClub, the largest P2P lending company in the world with over $11 billion originated loans, to understand key nuances of the loans and derive insights to support decision making.

The dataset comprises roughly 2.25M loans backed by LendingClub between 2007 & 2018. The variables in the dataset present an overall view of borrower characteristics and loan characteristics at the time of application and loan issuance. The raw data consists of 145 variables, of which many features have more than 50% of the data missing. These along with variables like ID, URL, etc. might not be intuitively useful for further analysis and can be dropped if they don't appear significant while exploration. The borrower characteristics are presented using variables like address, state, annual income, fico scores and debt-to-income ratio (dti) which can be leveraged to create a rich customer profile for our analysis. On the other hand, variables reflecting loan characteristics represent the term of the loan, amount, interest rates and the current status of the loan amongst other features. The current status of the loan indicates whether a loan is currently ongoing or complete and in the case of the latter, was it successfully repaid or defaulted. In terms of successfully repaid and defaulted loans, the data is significantly imbalanced and needs to be treated appropriately.

Using features related to loan issuance, income and years of experience we intend to draw conclusions based on region-wise loan targets answering to questions like why do we have a higher level of operating activity in particular regions such as Southwest and west regions? Leveraging borrowers' characteristics we intend to perform interesting hypothesis tests (using p-value, F-statistics, and R2) to answer questions like whether young people tend to take loans with higher instalments? Preliminary exploration of attributes related to borrower and loan characteristics along with loan grade through visual charts like box plots and histograms combined with correlation statistics can further help in answering questions which can help LendingClub identify specific markets for expansion and also design schemes with flexible interest rates and longer terms to attract young borrowers. Our exploration and analysis of the dataset will also include finding if the borrower attributes can be used to satisfactorily estimate the interest rate at which a loan might be offered for a particular borrower. Furthermore, we can use Principal Component Analysis (PCA) to identify features explaining maximum variability and use regression methods coupled with approaches like bootstrap, validation set, and cross-validation to report realistic evaluation metrics.

LendingClub assigns a grade (A-D) and sub-grade (1-5) to each borrower, which reflects their assessment of the credit risk of the corresponding loans. Once the grades and sub-grades are assigned, these are available as notes for the investors to analyze. Even when this information is available, lenders may not know how to extract useful knowledge from the data, and manually assessing a borrower's credit risk is rarely a practical alternative given the high level of expertise it requires. This motivates us to build machine-learned classification models that can evaluate the credit risk with knowledge from the historical loan data to help investors make an informed decision regarding the investments. For this, we could leverage the completed loan information from the dataset and identify trends separating good or successfully paid loans from the bad or charged-off loans.

Given an imbalanced dataset, we would identify a suitable resampling technique to aid in our modeling for credit risk. Different classification techniques like logistic regression, decision trees, random forest and support vector machines would be evaluated in conjunction with the resampling techniques to conclude on the best approach to a model with an optimal bias-variance tradeoff. From a model performance standpoint, accuracy does not indicate the true performance of the model in such an imbalanced set, and we would be focusing on criteria such as recall and precision to estimate the ability of our model to successfully capture possible default scenarios. With this effort, LendingClub, operating as the pioneers of P2P lending, would be able to further establish its reliability amongst its clients.

Dataset link: https://www.kaggle.com/wendykan/lending-club-loan-data