

THINKFUL – DATA SCIENCE LENDING CLUB LOAN PREDICTION PROJECT SCOPE

February 23, 2016

OVERVIEW

1. Project Background and Description

Lending Club is the world's largest online p2p marketplace connecting borrowers and investors. Every year, there are hundred thousand people loan requests through Lending Club. While Lending Club assigned a “grade” to each requested loan to help individual investors understand the risk; some loans still inevitably go bad regardless what the grades are. The project sets out to try to predict whether a loan will likely to go bad utilizing the information provided by the borrower and the grades assigned by Lending Club to help individual investors improve their portfolio return.

2. Project Scope

We'll be using the loan data provided by Lending Club from 2007 to 2013 that already have the results indicating whether a loan is good or bad – this will consist of over 200,000 data points. The dataset we'll be using is loans that was already vetting and approved by Lending Club, and we will not bother with the rejected loans. We'll break this dataset into training, validation, and testing sets to fairly access model accuracy. In addition, we'll integrate the zip code median income statistics provided by Population Studies Center into our dataset to make a few more predictors in addition to the one provided by Lending Club.

3. High-Level Requirements

- Clean up datasets provided by Lending Club and extract only the necessary fields
- Join Lending Tree dataset with Zip Code Income Median Dataset.
- Break the dataset into training, validation, and testing sets.
- Run a predictive model and assess its accuracy improvement against existing bad loan rate.

4. Deliverables

The final project report.