Credit Risk Prediction for Loan Borrowers

Kshitij Sankesara
Nitin Nagpal
Siddharth Srivastava
Tavleen Allagh

Part 1:

1.Area:
We will focus on credit modelling, a well-known data science problem that focuses on modelling a borrower's credit risk. Credit has played a key role in the economy for centuries and some form of credit has existed since the beginning of commerce. A credit risk is the risk of default on a debt that may arise from a borrower failing to make required payments.

2. Describe what the problem is:
Lending Club is a marketplace for personal loans that matches borrowers who are seeking a loan with investors looking to lend money and make a return. Each borrower fills out a comprehensive application, providing their past financial history, the reason for the loan, and more. Many loans aren't completely paid off on time, however, and some borrowers default on the loan. So, we need to analyze and predict the credit risk involved with each borrower.

3. Importance of the problem:
Investors are primarily interested in receiving a return on their investments. Approved loans are listed on the Lending Club website, where qualified investors can browse recently approved loans, the borrower's credit score, the purpose of the loan, and other important information from the application. Once they're ready to back a loan, they select the amount of money they want to fund. While Lending Club has to be extremely savvy and rigorous with their credit modelling, investors on Lending Club need to be equally as savvy about determining which loans are more likely to be paid off. While at first, you may wonder why investors would put money into anything but low interest loans. The incentive investors have to back higher interest loans are, well, the higher interest. If investors believe the borrower can pay back the loan, even if he or she has a weak financial history, then investors can make more money through the larger additional amount the borrower has to pay. Most investors use a portfolio strategy to invest small amounts in many loans, with healthy mixes of low, medium, and interest loans. The credit risk prediction will help the investor to know which loan might get paid off so that they can make investment decisions.

4. Description of project’s objective:
The objective of the project is to analyse and predict the credit risk involved with lending money to borrowers. To do that, we need to first understand the features in the dataset, perform exploratory data analysis and then experiment with building machine learning models that reliably predict if a loan will be paid off or not.

5. Expected results:
The objective of our project is to predict the credit risk involved with providing loan to a borrower. This will help investors to make informed decisions about whom to fund and what rate of interest to charge from the borrowers to be generate profits.

6. Statement of relevance:
By accurately predicting the credit risk we can save millions for the investors by providing them with intelligence to be shielded from the bad loans issue.

Part 2:

1.Goal title:
The goal of the project is to analyse and predict the credit risk involved with loan borrowers accurately.

2. Tasks:
The data set is obtained from Lending Club website which has million rows. We will begin with data cleaning. Then we will explore the data using visualizations. Later we will build machine learning models to analyse the data and to predict the credit risk. We will evaluate the data models by comparing their accuracy and efficiency.

3. Expected results:
The expected result of this project is the prediction of the credit risk for loan borrowers. This will help to understand the investments which are more risky. Also the interest rate provided can be altered based on their overall profile. This can help investors save from the issues of bad loan.

4. Expected problems:
As you understand each feature, you want to pay attention to any features that are not formatted poorly and need to be cleaned up, require more data or a lot of processing to turn into a useful feature, contain redundant information, leak information from the future (after the loan has already been funded), features that don't affect a borrower's ability to pay back a loan (e.g. a randomly generated ID value by Lending Club).We need to analyze carefully which features are important for our prediction and remove other features which doesn’t affect credit risk. The data type for values in the columns will have to be made uniform by converting into standard format. We will remove the NA values and rows with missing data or we will use other methods like taking the mean of the remaining data and filling it in the missing data whichever makes more sense.

Part 3:

Dataset: Lending Club Loan Data

Tools: Python, Spark ML, Pandas, Numpy, Matplotlib, Scikit learn

Models: Linear Regression, Logistic Regression, Random Forest

Criteria: Cross Validation, Class Imbalance, Confusion Matrix