Capstone 2: Project Proposal

Loan Prediction

Predicts if a loan will default using Machine learning prediction model. This will predict the chance of default on future loans. Dataset used will be existing loan data available from Kaggle, Lending Club Loan Data Analysis, which consists of the company (Lending Club) borrower's financial data such as credit score, payment history and credit limits.

Data source:

https://www.kaggle.com/datasets/deependraverma13/lending-club-loan-data-analysis-deep-learning/data

Problem statement

Provide Lending Club with a ML model that predict whether or not a loan will default based on their company's lending guidelines which can be used for future borrower risk assessment with at least 85% accuracy.

Context

Companies like Lending Club rely heavily on accurate predictions to assess the risk associated with loans. As borrowers' spending patterns, credit limits and scores change based on their life events, it could impact lending companies positions and ultimately result in default. Having a reliable ML model can ultimately reduce their risk and proactively take actions before reaching a point of borrowers' inability to pay.

Criteria for Success

Create a model that would predict the likelihood of loan default for future cases with at least 85% accuracy. Based on Machine learning industry standards between 70% to 90% is considered acceptable as a realistic and valuable model. However, our target should be 85% accuracy or above since incorrect predictions can result in high amount of profit loss.

Scope of Solution Space

We are unaware of the Lending Club's current risk tolerance and the accuracy targets. Therefore, we are using 85% as our target. The model may only be used to predict based on existing loan types (Credit cards, term loans, mortgages) that are available to Lending Club's

guidelines. Other lending companies lending guidelines and risk tolerance may vary, which will ultimately reflect the borrowers interest rate and ability to pay.

Constraints within solution space

The dataset is highly imbalanced and includes a lot of features. Certain loan types (secured vs unsecured) have different risk factors, interest rates and terms.

Stakeholders to provide key insight.

No current stakeholders. However, it would typically be head of the underwriting department and risk management team. Which would also be approved and reviewed by VP of Lending department depending on the organisational structure of the financial institution.

Key Data Sources

https://www.kaggle.com/datasets/deependraverma13/lending-club-loan-data-analysis-deep-learning/data

Data consists of:

- Borrower personal information, FICO scores and public records pertaining to tax filings.
- Loan data, Credit utilization, Revolving credit balance, days with credit line, interest rate.
- Additional information such as purpose and if customer met underwriting criteria.