**Capstone Project**

**(Credit Risk Analysis Model)**

**Document Skeleton**

# **Problem statement**

* **What is the problem or the opportunity that the project is investigating?**

The problem is Lender fails to detect the credit risk in advance, it exposes them to the risk of default and loss of funds.

We would like to investigate the probability of default of the borrower before providing a new lone, then investor can decide whether or not to extend credit to the borrower and the credit to be charged.

Maximize the Return of Investment and Reduce Risk of Default.

* **Why is this problem valuable to address?**

Lending Club issued $10.9 billion USD in new loans in 2018.

Net Revenue in 2018: $694.8 million USD

If the Default rate reduced by 1% will increase revenue approximately by $3.5M

* **What is the current state (e.g. unsatisfied customers, lost revenue)?**

Lending club lose a significant amount of their potential earnings to loan defaults.

* **What is the desired state?**

Maximize the Return of Investment and Reduce Risk of Default.

* **Has this problem been addressed by other research projects? What were the outcomes?**

Yes

# **Industry/ domain**

* **What is the industry/ domain?**

Financial Institutions and Banks

* **What is the current state of this industry?** (e.g. challenges from start-ups)

Lending Club is to empower our members on their path to financial health. As one of the original fin tech innovators with a 14-year track record of improving customers’ financial health, they have helped millions of members save money by reducing their high-interest credit card debt through a lower interest personal loan. Today, they are taking the next step in Lending Club’s journey and are poised to reimagine banking with the acquisition of Radius Bank.

* **What is the overall industry value-chain?**

****

**Borrowers** get funded

**Investors** build portfolio

**Borrowers** repay automatically

**Investor** earn & reinvest

**Borrowers** apply for Loans

**Investor** Open an account

**Borrowers** repay automatically

**Investor** earn & reinvest

Lending Club with the idea that bringing borrowers and investors together can help everybody succeed.

* **What are the key concepts in the industry?**

Lending club Marketplace Platform helps borrowers take control of their debt and empowers everyone to reach their financial goals.

* **Is the project relevant to other industries?**

Yes

# **Stakeholders**

* **Who are the stakeholders? (be as specific as possible)**

Annie Armstrong

Chief Risk Officer, Lending Club

* **Why do they care about this problem?**

In 2018, after all expenses, lending club lost approximately 128M as consolidated defaulters loan amount was 30.2%.

* **What are the stakeholders’ expectations?**

Increase company Annual Revenue

# **Business question**

* **What is the main business question that needs to be answered?**

Can we predict if the customer will default on their loan or not so that we can increase the company revenue by reducing the default rate ?

* **What is the business value of answering this question?** (quantify value and make necessary assumptions)

If the Default rate reduced by 1% will increase revenue approximately by $3.5M

* **What is the required accuracy? What are the implications of false positives or false negatives?**

More than 90% accuracy

# **Data question**

* **What is the data question that needs to be answered?**

Using historical data, Can we predict if the customer will default on their loan or not before provide a new loan?

* **What is the data required to answer the question?**

Historical data of Lending club

# **Data**

* **Where was the data sourced?**

Lending club, Data. World

* **What is the volume and attributes of the data?**

342,540 Rows, 152 Features

* **How reliable is the data?**

Data is reliable, it is from Lending Club online platform, and they shared some sample data.

* **What is the quality of the raw data?**

Quality of raw data is acceptable and has some missing values need to clean**.**

* **How was this data generated?**

Generated by Lending club, when borrower filled online form while applying for a Loan.

* **Is this data available on an ongoing basis?**

Yes

# **Data science process**

## **Data analysis**

* **What data pipeline was to wrangle the raw data?**

Data is the distribution of the loan by the year 2018.

Data has 152 columns with 32 columns containing high percentage of missing data.

* **What are the highlights of the Exploratory Data Analysis (EDA)?**
  + This is an imbalance data, because we have a lot more entries of people that currently reply their loans than people that did not pay back.
  + In the loan amount distribution we can see spikes in even ten thousand dollar, so this is indicating that there are certain amounts that are basically standard loans.
  + 52% of the people borrow the loan for debt consolidation and credit card repayments.
  + The mean interest rate is c.13% and the rate spread is very large from 5%-31%.
* **Is the pipeline reusable? (for example, to process future data?)**

Yes

* **What are the intermediary data structures used (if any)?**

## **Modelling**

* **What are the main features used?**

Target variable is defaulter\_probability

140 Features (Including dummy variable)

Loan amount, term, interest rate, grade, income, emp\_length, fico\_range, loan status etc…

* **Did you find any interesting interactions between features?**
  + The interest rate has essentially the highest negative correlation with whether or not someone is going to repay their loan.
  + Looks like regardless of what actual employment length you have if you were to pick someone, about 20% of them are going to have not paid back their loans.
* **Is there a subset of features that would get a significant portion of your final performance? Which features?**

Loan Amount

Term

Interest Rate

Installment

Grade

Sub Grade

Employee Length

Home Ownership

Annual Income

Verification Status

Loan Status

Purpose

Addr State

Dti

Fico Range Low

Fico Range High

Open Acc

Pub Rec

Revol Bal

Revol Util

Total Acc

Application Type

Mortagage Account

Pub Rec Bankruptcies

* **How did you select features?**

Select the features with are most relevant to predict default

* **What feature engineering techniques used?**
  + transfer term into a numeric variable with unit years
  + map the grade feature to numeric values
  + convert Verification status, application type, initial\_list\_status columns into dummy variables and concatenate them with the original data frame
* **What are the models used?**

Logistic Regression, Decision Tree, Bagging-K-NN, Random forest, K-NN, Bagging Tree, AdaBoot, XGBoost, Gradient Boost, Baseline model with grade and credi

* **How long does it take to train your model?**

1 hour (approximately)

* **What are the tools used? (cloud platform, for example)**
* **What are the model performance metrics?**

\*\*\*\*\*\*\*\*\*\*\*\*

XGBoost

\*\*\*\*\*\*\*\*\*\*\*\*

Accuracy : 0.9464 [TP / N] Proportion of predicted labels that match the true labels.

Precision: 0.9170 [TP / (TP + FP)] Not to label a negative sample as positive.

Recall : 0.9816 [TP / (TP + FN)] Find all the positive samples

ROC AUC : 0.9464

* **Which model was selected?**

XGBoost

## 

## **Outcomes**

* **What are the main findings and conclusions of the data science process?**

Credit score and Grade are great for rate definition for prediction we need more features

## **Implementation**

* **What are the considerations for implementing the model in production?**

# Data answer

* **Was the data question answered satisfactorily?**

Yes

* **What is the confidence level in the data answer?**

94%

# Business answer

* **Was the business question answered satisfactorily?**

Yes

* **What is the confidence level in the business answer?**

High

# Response to stakeholders

* **What are the overall message and recommendations to the stakeholders?**

Execute the model every time before providing a new loan

Review and refine the model with updated data

# End-to-end solution

* **What is the overall end-to-end solution to use the model developed in the project?**

Data Answer:

Yes, we can predict if customer will default on their loan before providing a new loan using historical data with 94% accuracy.

Business Answer:

Yes, we can predict if the customer will default or nor to increase company revenue. If the Default rate reduced by 1% will increase revenue approximately by $3.5M.

Model Selected: XGBoost

# References

* **Where are the data and code used in the project? (show a simplified list of main items: notebooks, datasets, exported models)**

Lending club data 2018

Notebook: EDA, Modelling with grade and without Grade

* **What are the resources used in the project? (libraries, algorithms, etc)**

https://www.lendingclub.com/company/about-us