horizontal line



**Peer-To-Peer Credit Risk Analysis**

Project Summary

|  |  |
| --- | --- |
| Batch details | April 2022 |
| Team members | 5 |
| Domain of Project | Finance (Risk Management) |
| Proposed project title | Peer-To-Peer Credit Risk Analysis |
| Group Number | Group-4 |
| Team Leader | Mr. Gaurav Aher |
| Mentor Name | Ms. Anjana Agrawal |

Date: 29/July/2022

G.Y. Aher

Signature of the Mentor Signature of the Team Leader

# 

Table of Contents

|  |  |  |
| --- | --- | --- |
| Sl NO | Topic | Page No |
| 1 | Overview | 3 ,4 |
| 2 | Business problem goals | 4 ,5 |
| 3 | Topic survey in depth | 5,6 |
| 4 | Critical assessment of topic survey | 6,7 |
| 5 | Methodology to be followed | 7,8,9 |
| 6 | References | 9,10 |

Project Details

***OVERVIEW (50-100 words)***

* ***What is Loan?***  
  A loan is a sum of money that one or more individuals or companies borrow from banks or other financial institutions so as to financially manage planned or unplanned events. In doing so, the borrower incurs a debt, which he has to pay back with interest and within a given period of time.

**Our dataset is based on Peer-To Peer lending**

* **What is Peer-To-Peer loan?**

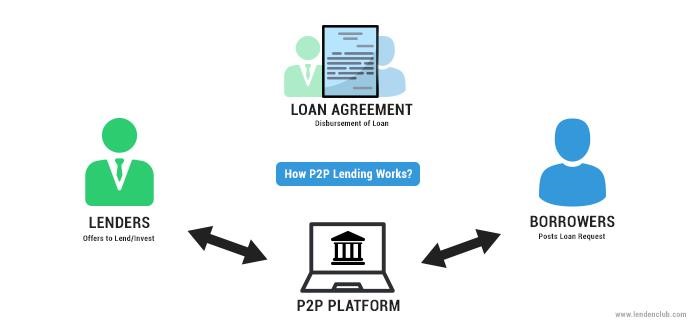
# Peer-to-peer lending, also referred to as P2P lending, is an alternative financing method which allows individuals to avail loans from other individuals through various lending platforms

# Recent peer-to-peer lending trend in the last decade has led to a boom in online lending platforms such as "LendingClub", "Peerform", "Upstart", "Fun ding Circle", "Lendbox", etc.

* Borrowers with low credit scores in traditional financial institutions are

more likely to apply for the same loan in these new platforms, increasing

the potential credit risk of default to the investor.



***Business problem statement (GOALS)***

* ***Business problem statement***
* P2P loans are not completely secure as they involve substantial risk of default (credit risk) and hence require an added effort to identify and determine responsible borrower from a pool of unknown users.
* Prediction of credit default of an individual, is very powerful risk assessment tool to contribute to save the principal and interest amount.
* By understanding and analyzing previous p2p lending records which have either ‘Charged Off’ or ‘Fully-paid’, our main objective is to help predict whether the potential borrower will default or not, based his details.
* ***Business*** ***Objective***
* The main objective of our work is to develop a credit risk analysis prediction model which will help lending club to predict customers who are most likely to default, **Which will help club to reduce losses.**
* ***Approach***

We shall incorporate the methodology of Cross Industry Standard Process for Data Mining (CRISP-DM). Following are the methods to understand the business problem and proceed with our aim:

❏ Understanding the data

❏ Data Cleaning

❏ Data Transformation

❏ Data Discovery and Visualization

❏ Model Building: Regression/Classification/Clustering

❏ Feature Engineering

❏ Model Tuning

* ***Conclusions***

To build a model that predicts loan defaulters by analyzing the customers using the historical data, by keeping track of the condition of customers in real time and assessing the risk factors by implementing machine learning classification models (Logistic Regression, Random Forest etc).

***TOPIC SURVEY IN BRIEF***

* ***Problem understanding***

The reason behind describing customer credit risk analytics in the preceding paragraphs is because, the goal of our next Machine Learning project is to develop a model that can predict customers who are most likely to default. With the help of various visualization libraries that are at our disposal, we will be able to figure out possible parameters that govern a customer’s decision to default.

* ***Current solution to the problem***

1. A strong data infrastructure is vital as it underpins a lender’s ability to understand the risks it is facing. Data should be readily available and easily aggregated
2. Borrowers in distress should be identified early so that viable borrowers can be provided with sustainable solutions in a timely manner.
3. Create a clear plan for payment reminders at every stage

* ***Proposed solution to the problem***

We’ll use Exploratory Data analysis, Data Visualization, Machine Learning Algorithm drive predictive systems to build various machine learning models

* **EDA**

1. We’ll perform all possible combinations of univariate, bivariate and multivariate analysis and will visualize the results to get insights from the data.
2. We’ll clean the data for model building.

* **Model Building**

We’ll use :-

1. **Decision Tree Classification**
2. **Random Forest Classification**
3. **Logistics Regression**

* ***Reference to the problem***

We are going to apply our knowledge as a group and take guidance from the mentor, and plenty of references from the Internet to back up our stance and objective.

***CRITICAL ASSESSMENT OF TOPIC SURVEY (50-100 words)***

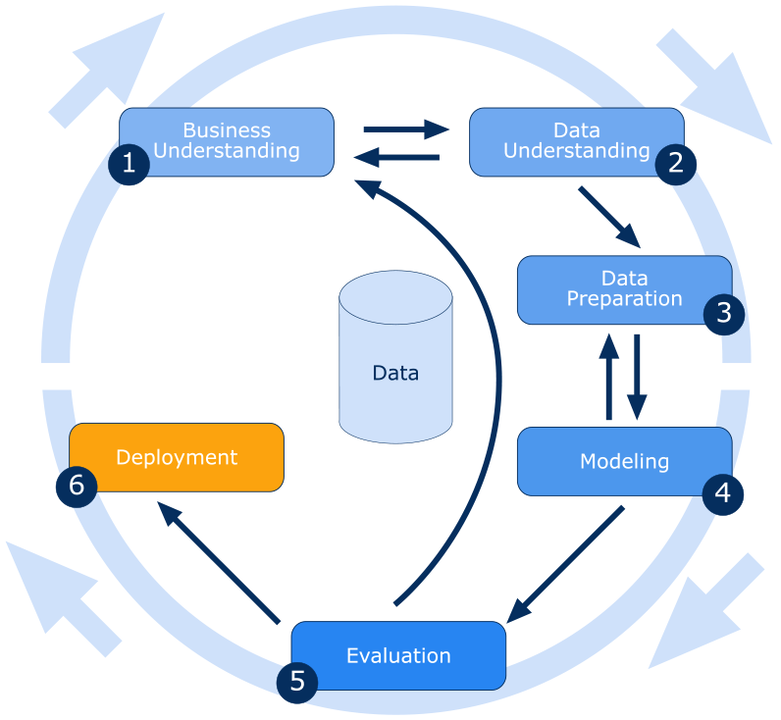
* ***Find the key area, gaps identified in the topic survey where the project can add value to the customers and business?***

1. Getting started with peer-to-peer lending can be a bit daunting for new investors. Lending Club offer hundreds of loans to choose from and a myriad of ways to invest money. Some people that are eager to get started jump out of the gate without thinking through their options and end up disappointed with their investment returns.
2. Investors have a choice to invest in A grade loan which provides them with mid-single digit interest to loans of over 20% interest. Sure, you might think, a 20% return sounds much better. But keep in mind these are higher risk borrowers that historically have had a higher default rate. So, a portfolio of loans that are earning over 20% could easily end up with a real world return under 10% once all the defaults have been taken into consideration. Using the ML model, we could achieve more favorable outcomes of a credit default and the grade of loan which investors should provide to the borrowers.

* ***What key gaps are you trying to solve ?***

Our main objective is to predict credit risk of lenders , so that there will be minimum loss.

***METHODOLOGY TO BE FOLLOWED (Explain each steps from 1-5)***



# *\* If deployment is out of scope to the team or not advised by the mentor, please opt to leave out the step no 6.*

# *Reference documents of CRISP-DM*

1. <https://paginas.fe.up.pt/~ec/files_0405/slides/02%20CRISP.pdf>
2. <https://en.wikipedia.org/wiki/Cross-industry_standard_process_for_data_mining>

***Business Understanding***

**What does the business need?**

This phase consists of a very precise specification of the problem together with methods of evaluating the achievement of the goal so that we can have a clear understanding of the business problems:

* Classify whether the borrower will default or not
* Classify the grade(risk) of the loan.

***Data understanding***

This phase involves understanding the data considered for finding the solution:

* Dataset consists of 396030 observations and 27 features.
* Summary of missing values:

|  |  |
| --- | --- |
| **Feature Name** | **% Of missing values** |
| emp\_title | 5.79 |
| emp\_length | 4.62 |
| title | 0.44 |
| revol\_util | 0.06 |
| mort\_acc | 9.54 |
| pub\_rec\_bankruptcies | 0.13 |

* For missing values - As we can infer that all missing values are less than 10%, so we’ll impute null values with **mean, median and mode** **based on** **the skewness** of the features.
* If we find skewness in data we can use various transformation methods such as log transformation, square root transformation, box cox transformation,etc.
* For outliers – First we’ll visualize the outliers by using boxplot, then we’ll treat outliers using **IQR** method or **Capping** method.
* Dataset is imbalanced having 318357 ‘Fully Paid’ observations and 77673 ‘Charged Off’ observations

***Modelling:***

This phase involves finding the model that captures the solution to the business problem using available data. We may have to try multiple models and go back and forth between data preparation and modelling to choose the correct model.

***Evaluation:***

Once the model is built, we will check how good the model performs on unseen data by using measures like AUROC, F1 ration, Precision & Recall for the minority class.

* **Deployment**:

If we are satisfied with the performance of the model from the previous phase, we deploy it in the deployment phase.

***REFERENCES***

The references can be blogs, articles or even social media news relevant to explain the importance of the projects.

1. Why credit risk analysis is important- ([link](https://www.researchgate.net/publication/312038529_Credit_Risk_Analysis_in_Peer-to-Peer_Lending_System" \l ":~:text=The%20P2P%20system%20allows%20investors,and%20interest%20not%20being%20rep))
2. Credit risk analysis with machine learning techniques in peer-to-peer lending market-([link](https://www.diva-portal.org/smash/get/diva2:1375762/FULLTEXT01.pdf))
3. Logistic Regression.

Retrieved from [https://www.sciencedirect.com/topics/computer-science/logistic-](https://www.sciencedirect.com/topics/computer-science/logistic-regression) [regression](https://www.sciencedirect.com/topics/computer-science/logistic-regression)

1. Decision Tree Algorithm, explained.

Retrieved from [https://www.kdnuggets.com/2020/01/decision-tree-algorithm-](https://www.kdnuggets.com/2020/01/decision-tree-algorithm-explained.html) [explained.html](https://www.kdnuggets.com/2020/01/decision-tree-algorithm-explained.html)

1. Gradient Boosting from scratch.

Retrieved from [https://blog.mlreview.com/gradient-boosting-from-scratch-](https://blog.mlreview.com/gradient-boosting-from-scratch-1e317ae4587d) [1e317ae4587d](https://blog.mlreview.com/gradient-boosting-from-scratch-1e317ae4587d)

**Notes For Project Team**

*Sample Reference for Datasets (to be filled by team and mentor)*

|  |  |
| --- | --- |
| Original owner of data | **Lending Club** |
| Data set information | **Finance Domain, Shape (**396030,27**)** |
| Any past relevant articles using the dataset | NA |
| Reference | KAGGLE |
| Link to web page | ([link](https://www.kaggle.com/code/rhesamulyadi/prediction-model-with-ml-and-dl-tf-and-keras/data)) |

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