# Project Report
# On

## Credit risk minimizing model in Peer to Peer (P2P) lending business.

Submitted in partial fulfillment for the award of
Post Graduate Diploma in Big Data Analytics (PG-DBDA)
From Know-IT(Pune)

**Guided by:**
**Ms. Trupti Joshi**
**Mr. Milind Kapase**

**Submitted By:**

Krishna Pawar (220943025014)
Pratik Chavan (220943025027)
Atharva Shinde (220943025036)
Shubham Sarade (220943025043)

# CERTIFICATE

**<u>TO WHOMSOEVER IT MAY CONCERN</u>**

**This is to certify that**

Krishna Pawar (220943025014)
Pratik Chavan (220943025027)
Atharva Shinde (220943025036)
Shubham Sarade (220943025043)

**Have successfully completed their project on**

## Credit risk minimizing model in Peer to Peer (P2P) lending business.

**Under the guidance of**

Ms. Trupti Joshi

Mr. Milind Kapase

# ACKNOWLEDGEMENT

This project **“Credit risk minimizing model in Peer to Peer (P2P) lending business.”** was a great learning experience for us and we are submitting this work to CDAC Know-IT (Pune).

We all are very glad to mention the name of **Ms. Trupti Joshi** and **Mr. Milind Kapase** for his valuable guidance to work on this project. His guidance and support helped us to overcome various obstacles and intricacies during the course of project work.

We are highly grateful to **Mr. Vaibhav Inamdar** Manager (Know-IT), C-DAC, for his guidance and support whenever necessary while doing this course Post Graduate Diploma in Big Data Analytics (PG-DBDA) through C-DAC ACTS, Pune.

Our most heartfelt thanks goes to **Mr. Shriniwas Jadhav** (Course Coordinator,PG-DBDA) who gave all the required support and kind coordination to provide all the necessities like required hardware, internet facility and extra Lab hours to complete the project and throughout the course up to the last day here in C-DAC Know-IT, Pune.

**From:**

Krishna Pawar (220943025014)<br>
Pratik Chavan (220943025027)<br>
Atharva Shinde (220943025036)<br>
Shubham Sarade (220943025043)

# TABLE OF CONTENTS

# Abstract

Credit risk is a major concern for Peer-to-Peer (P2P) lending platforms. The ability to predict loan defaults and minimize credit risk is crucial for P2P lending businesses to sustain and grow in the market. In this project, we aim to build a credit risk minimizing model for a P2P lending business using machine learning techniques on the Lending Club dataset. We will pre-process the dataset by handling missing values, encoding categorical variables, scaling and splitting the data into training and test sets. We will then use various machine learning algorithms such as Logistic Regression, Random Forest, and Gradient Boosting to predict loan defaults and evaluate the model's performance using metrics such as ROC- AUC score. Finally, we will select the best-performing model and use it to predict loan defaults on new data. The outcome of this project can help P2P lending businesses to make informed decisions and minimize credit risk.

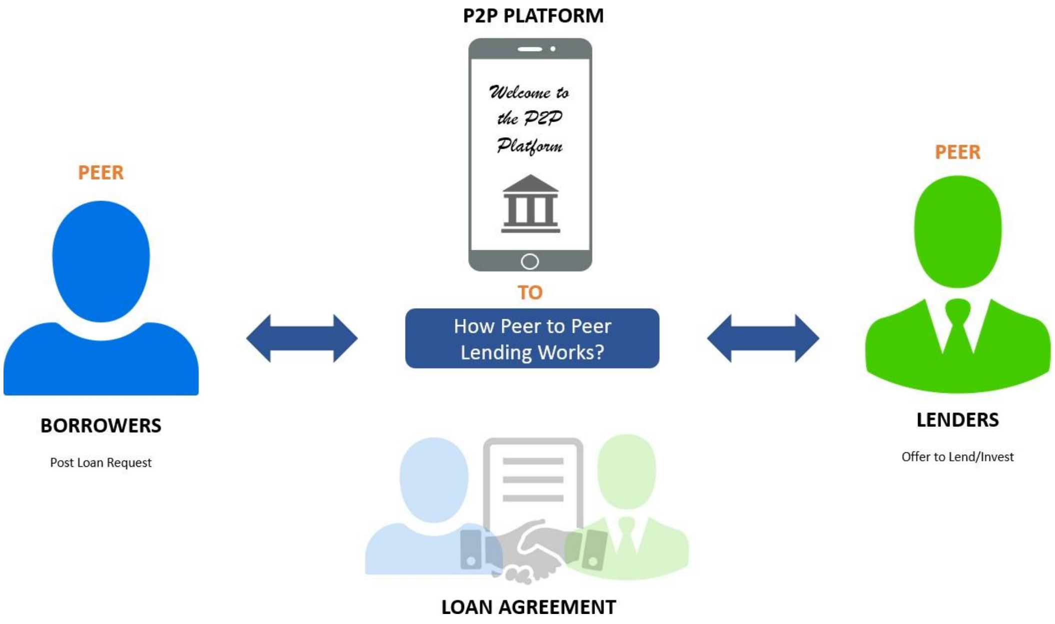
# 1. INTRODUCTION

Peer-to-Peer (P2P) lending business has been gaining a lot of traction in recent years due to its ease of access, low interest rates, and high returns. It is a platform that connects lenders directly with borrowers, removing intermediaries like banks. However, with the rise of P2P lending, there has been an increase in credit risk, leading to a rise in defaults.

To mitigate credit risk and minimize defaults, machine learning can be used to develop predictive models that analyze borrower profiles and loan characteristics. In this project, we aim to develop a credit risk minimizing model for P2P lending using machine learning on the Lending Club dataset.

The Lending Club dataset contains historical data on loans issued by the Lending Club platform, including borrower information, loan characteristics, and loan status. By analyzing this data and using machine learning algorithms, we can predict the likelihood of a borrower defaulting on a loan.

The goal of this project is to build a robust machine learning model that can accurately predict loan defaults and identify high-risk borrowers. This model can then be used by P2P lending platforms to make informed decisions when issuing loans, ultimately minimizing credit risk and maximizing returns for investors.



# 2. SYSTEM REQUIREMENTS

**Hardware Requirements:**

- Platform – Windows OS
- RAM – 8 GB of RAM,
- Peripheral Devices – Mouse, Keyboard, Monitor
- A network connection for data recovering over network.

**Software Requirements:**

- PySpark
- MLlib
- Tableau
- Google Colab

# 3. FUNCTIONAL REQUIREMENTS

## Apache Spark :-

- Apache Spark is a multi-language engine for executing data engineering, data science, and machine learning on single-node machines or clusters.
- Spark can process data in memory, which can significantly speed up machine learning algorithms.

## PySpark :-

- PySpark is the Python API for Apache Spark, and is widely used in machine learning projects due to its ability to leverage the strengths of both Python and Spark.
- PySpark allows developers to write Spark code in Python, which can be executed on a distributed computing cluster, enabling processing of large datasets in a scalable and efficient way.
- One of the key benefits of using PySpark for machine learning is the ability to leverage Python's rich ecosystem of libraries and tools for data preprocessing, feature engineering, model training, and evaluation.
- Pyspark comes with a number of built-in libraries that can be used for a wide range of tasks, including MLlib, PySpark SQL, GraphX etc.

## MLlib :-

- PySpark MLlib is the machine learning library in PySpark, which provides a set of high-level APIs for building machine learning pipelines in Python.
- MLlib supports common machine learning algorithms, such as classification, regression, clustering, and collaborative filtering.
- It also provides support for feature engineering, model selection, and evaluation.

## Tableau :-

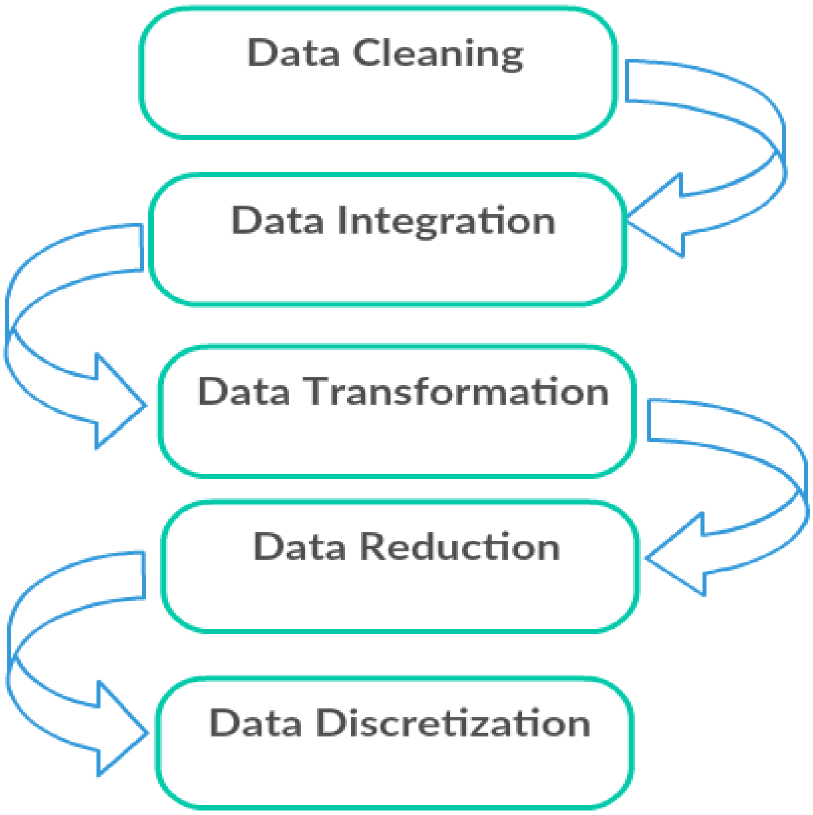
- Data visualization is the graphical representation of information and data.
- Tableau is a powerful data visualization tool that can be used in machine learning projects to help analyze and visualize data, as well as to communicate insights to stakeholders.
- Tableau can connect to various data sources, including databases, spreadsheets, and cloud-based services, making it an ideal tool for working with data generated by machine learning models.

## Google Colab :-

- Google Colab is a cloud-based notebook environment that allows users to write and execute Python code in a Jupyter notebook-like interface.
- It provides a free and easy-to-use platform for running machine learning experiments.
- And also provides access to powerful hardware, including GPUs and TPUs, for running computationally-intensive machine learning tasks.

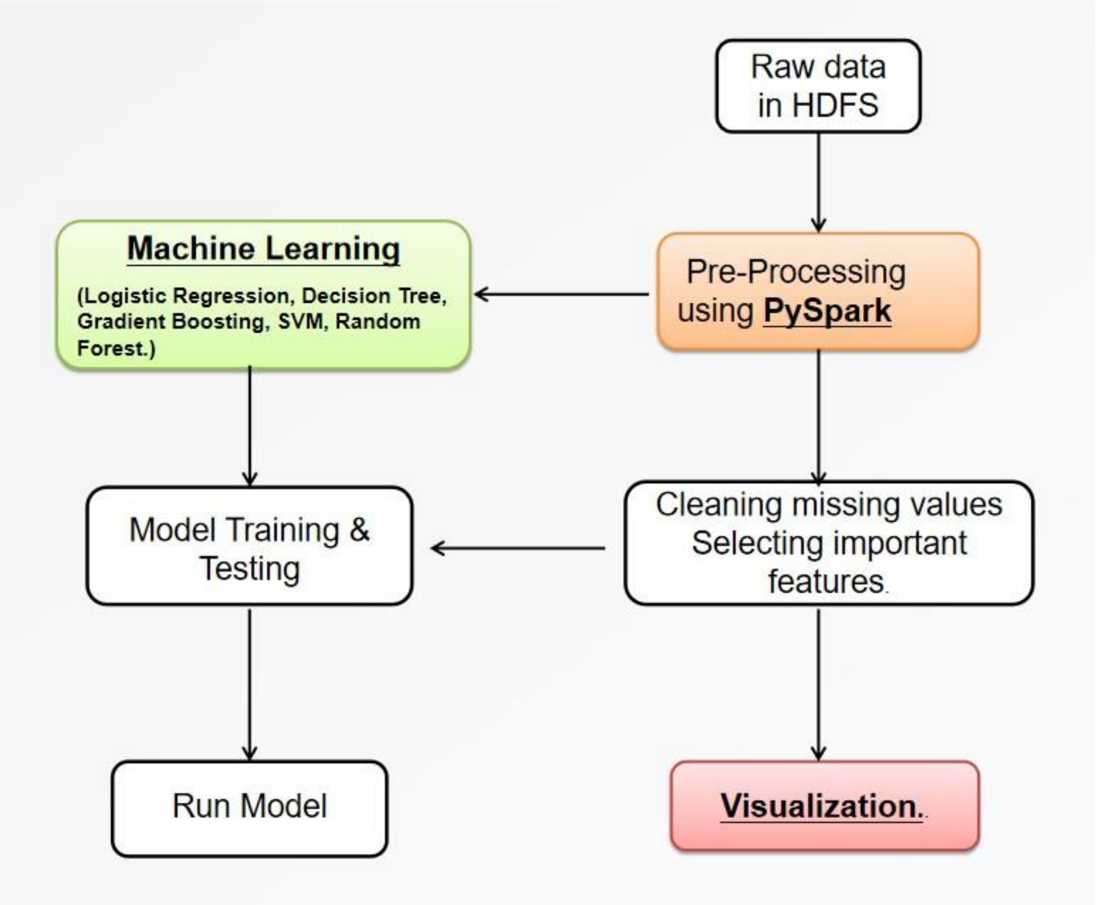
**Data Pre-Processing :-**

Data preprocessing is an essential step in machine learning projects, as it helps to prepare the data for modeling and analysis. The main goal of data preprocessing is to transform the raw data into a format that can be easily used by machine learning algorithms. The following are some common steps in data preprocessing:

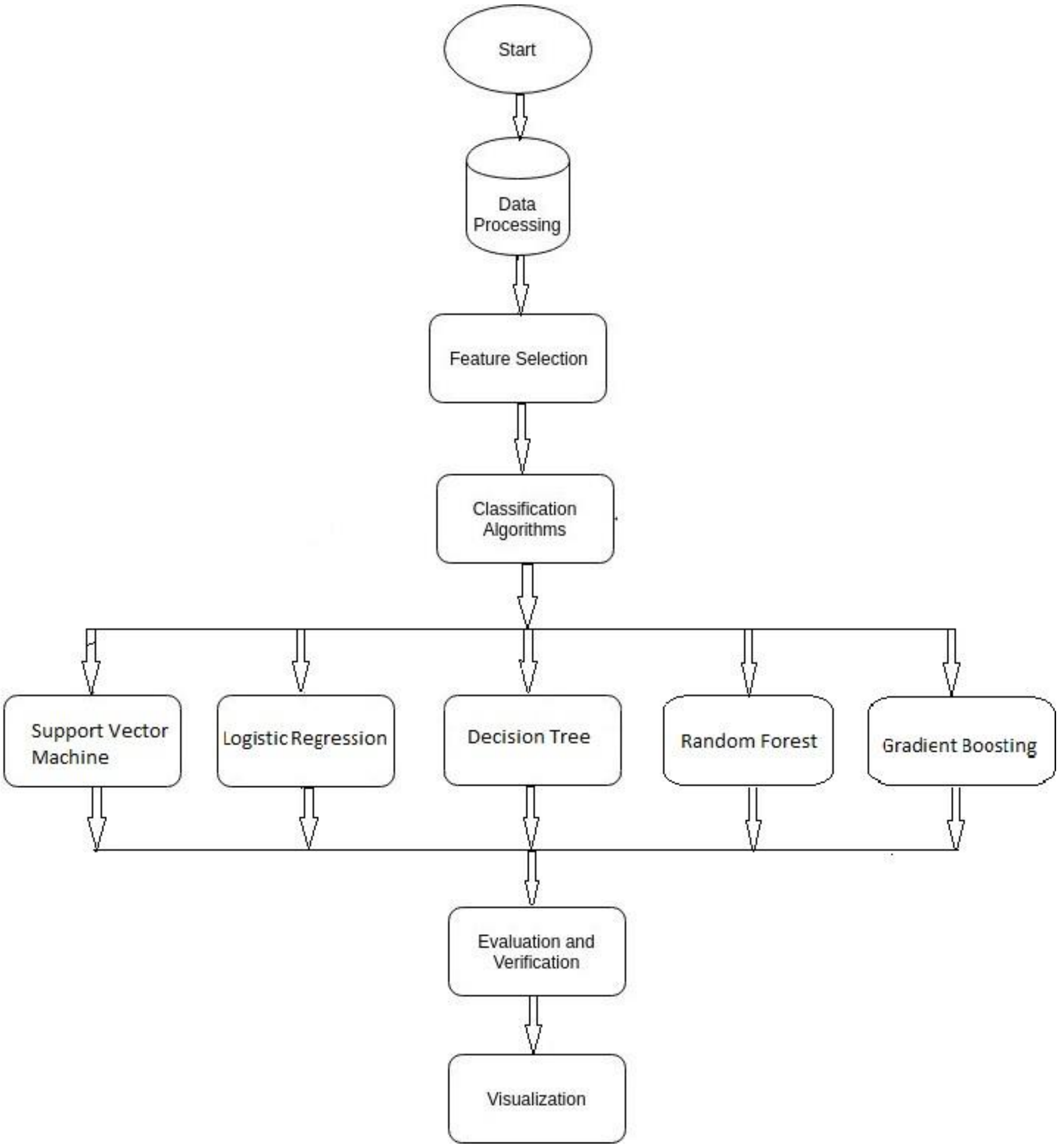


- **Data Cleaning :-** This involves removing or correcting errors, missing values, and inconsistencies in the data. This can be done using techniques like imputation, filtering, or dropping the data points with missing values.
- **Data Integration :-** This involves integrating data from multiple sources, such as combining data from different databases or merging data from different datasets.
- **Data Transformation :-** This involves transforming the data into a more usable format, such as converting categorical variables into numerical variables, scaling the data, or normalizing the data.
- **Data Reduction :-** This involves reducing the size of the dataset by selecting only the most relevant features or by sampling the data.
- **Data Discretization :-** This involves converting continuous data into discrete categories. This can be useful for certain types of machine learning algorithms.

# 4. SYSTEM ARCHITECTURE


Raw data in HDFS
Machine Learning
(Logistic Regression, Decision Tree, Gradient Boosting, SVM, Random Forest.)
Pre-Processing using PySpark
Model Training & Testing
Cleaning missing values Selecting important features.
Run Model
Visualization..

# 5. METHODOLOGY



# 6. MACHINE LEARNING ALGORITHMS

In this project we apply various different types of Classification and Regression Algorithms such as Logistic Regression, Decision Tree, Random Forest, Gradient Boosting and Support Vector Machine (SVM). During the implementationwe analyze the accuracy of all the algorithms.

Machine learning is the research that explores the development of algorithms that can learn from data and provide predictions based on it. Works that study flight systems are increasing the usage of machine learning methods. They were mainly used for classification and prediction. In this project we use various machine learning algorithms which are as follows:

## Logistic Regression :

- Logistic regression uses a logistic function (also called sigmoid function) to model the relationship between the predictor variables and the probability of the binary outcome.
- The logistic function has an S-shaped curve that maps any real-valued input to a value between 0 and 1, which represents the estimated probability of the binary outcome.

**Pros:**

- Simplicity :- It is a simple and fast algorithm that can handle large datasets with many input variables.
- Robustness :- It is less sensitive to outliers and noise than other machine learning models making it more robust to data quality issue.

**Limitations**

- Linearity :- It assumes a linear relationship between the input variables and the output variables which may not hold in same cases.
- Over fitting :- It may over fit the training data if the model is too complex or the dataset is too small.

## Decision Tree :

- Decision Tree is a visual tool used to make decisions by breaking down complex problem into smaller and more manageable parts.
- To create a decision tree, start by selecting the most important variable and splitting the data until you have a tree that accurately predicts the outcomes.
- Decision Tree can be pruned to reduce overfitting and improve accuracy.

**Pros**

- Non-Linearity :- Decision trees can capture non-linear relationships between the input variables and the output variable which may be important for predicting credit risk accurately.
- Feature Selection :- It can automatically select the most informative input variables for the classification task, reducing the need for manual feature engineering.

**Limitations**

- Over fitting :- Decision trees are prone to over fitting the training data if the model is too complex or the dataset is to small.
- Bias :- Decision trees can introduce bias into the model if the training data is unrepresentative or biased towards certain classes.

## Random Forest :

- In Random Forest algorithm, it combines multiple decision trees to improve accuracy.
- Each decision tree is trained on a subset of the data and a random subset of variable.
- It predicts outcomes based on the combined results of all the decision trees.

**Pros**

- Non-Linearity :- Random forests can capture complex non-linear relationships between the input variables and the output variable, which can be important for accurately predicting credit risk.
- Ensemble Learning :- Random forest use an ensemble of decision trees to make predictions, which can improve the accuracy and stability of the model.

**Limitations**

- Complexity :- Random forest can be more complex and computationally intensive than some other machine learning models especially for large datasets.
- Hyper parameter Tuning :- It require careful tuning of hyper parameters, such as the number of trees and the maximum depth of each tree, to achieve goo performance.

## Gradient Boosting :

- It is a type of ensemble learning that combines multiple weak models to create a strong predictive model.
- Gradient Boosting is an iterative process that starts with a single weak model and adds more models sequentially, each one trying to correct the errors of the previous model.

**Pros**

- Non-linearity :- Gradient boosting is a non-linear model, which means it can capture complex relationships between features and the target variables. This is particularly important in credit risk modelling, where the relationship between different variables can be complex and non-linear.
- Ensemble Learning :- It is an ensemble learning method which means it combines the predictions of multiple weaker models to create a stronger final model. This can improve the accuracy and robustness of the model, particularly when dealing with noisy or incomplete data.

## Support Vector Machine (SVM) :

- SVM is a supervised learning algorithm that finds the best hyperplane to separate data points into different classes.
- The hyperplane is determined by maximizing the margin between the closest points from each class, which are called support vectors.

**Pros**

- Non-linearity :- Like gradient boosting, SVMs can handle non-linear relationships between the features and the target variable, using techniques such as kernel methods to map the data into a higher-dimensional space.
- Tunable Parameters :- SVMs have several hyperparameters that can be tuned to optimize performance on a given dataset, such as the choice of kernel functions and the regularization parameters. This flexibility can help improve the model's accuracy and generalizability.
- Binary Classification :- SVMs are well-suited for binary classification tasks such as credit risk modeling, where the goal is to classify loans as either good or bad.

**Limitations**

- Computational Complexity :- SVMs can be computationally expensive, particularly when dealing with large datasets or high-dimensional feature spaces. This can make training and evaluating the model more time consuming and resource intensive.

# 7. DATA VISUALIZATION AND REPRESENTATION

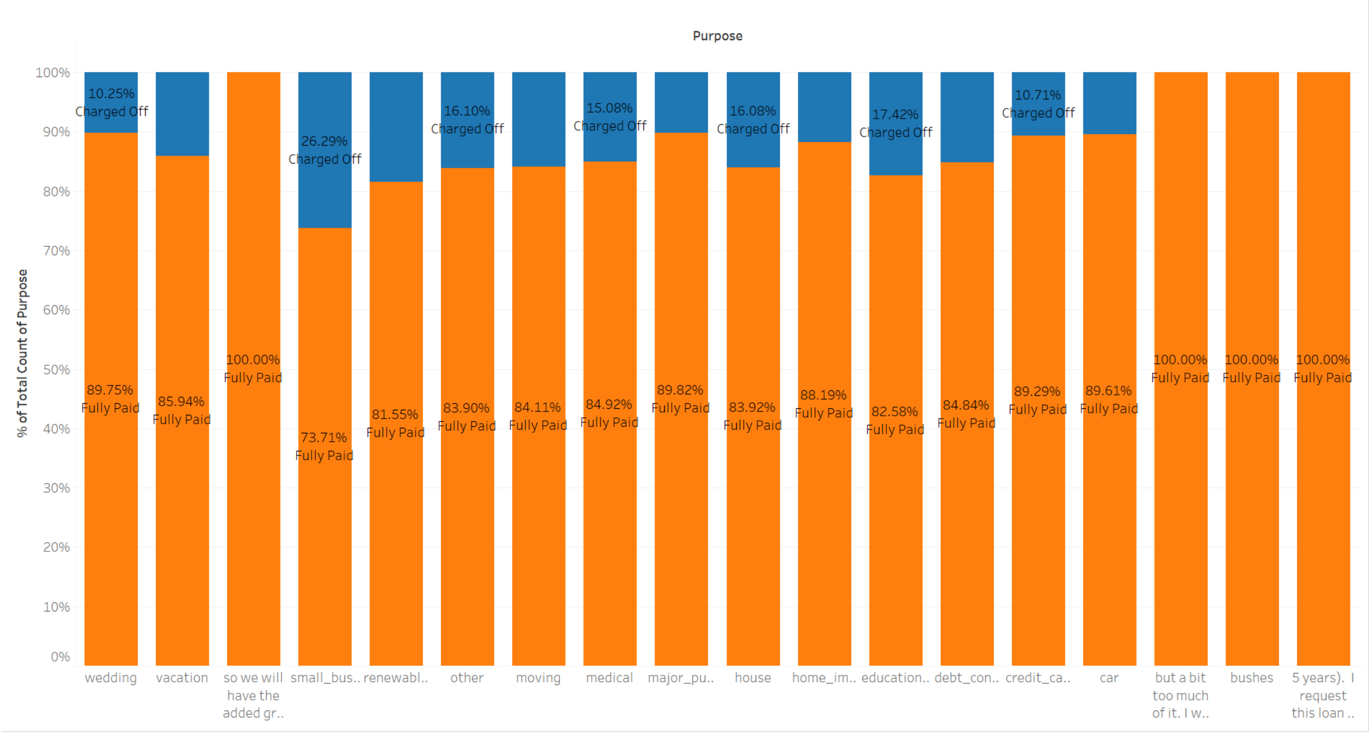


Fig :- The most risky loan purpose among all is small business type borrower’s.

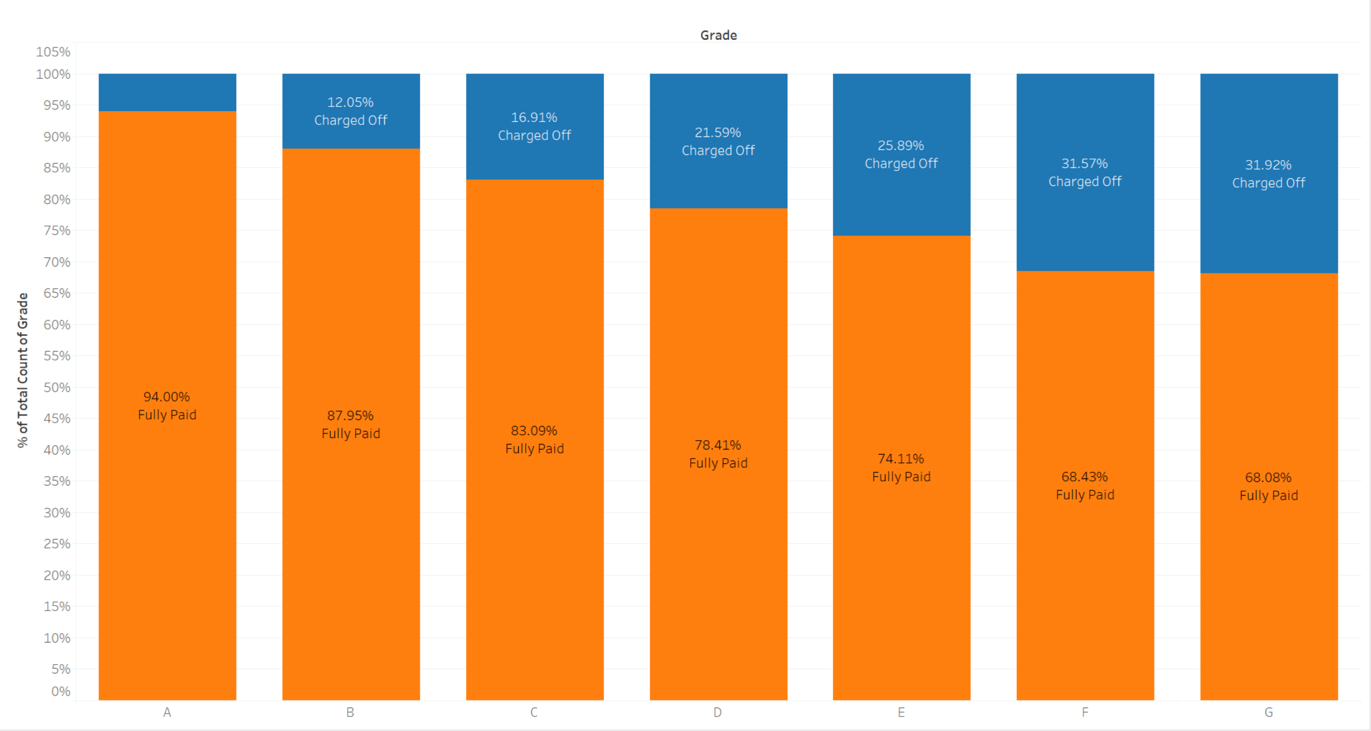


Fig :-.Grading system of company providing good results by correctly segregating Full Paid and Charged Off borrower’s into different categories for easy processing.

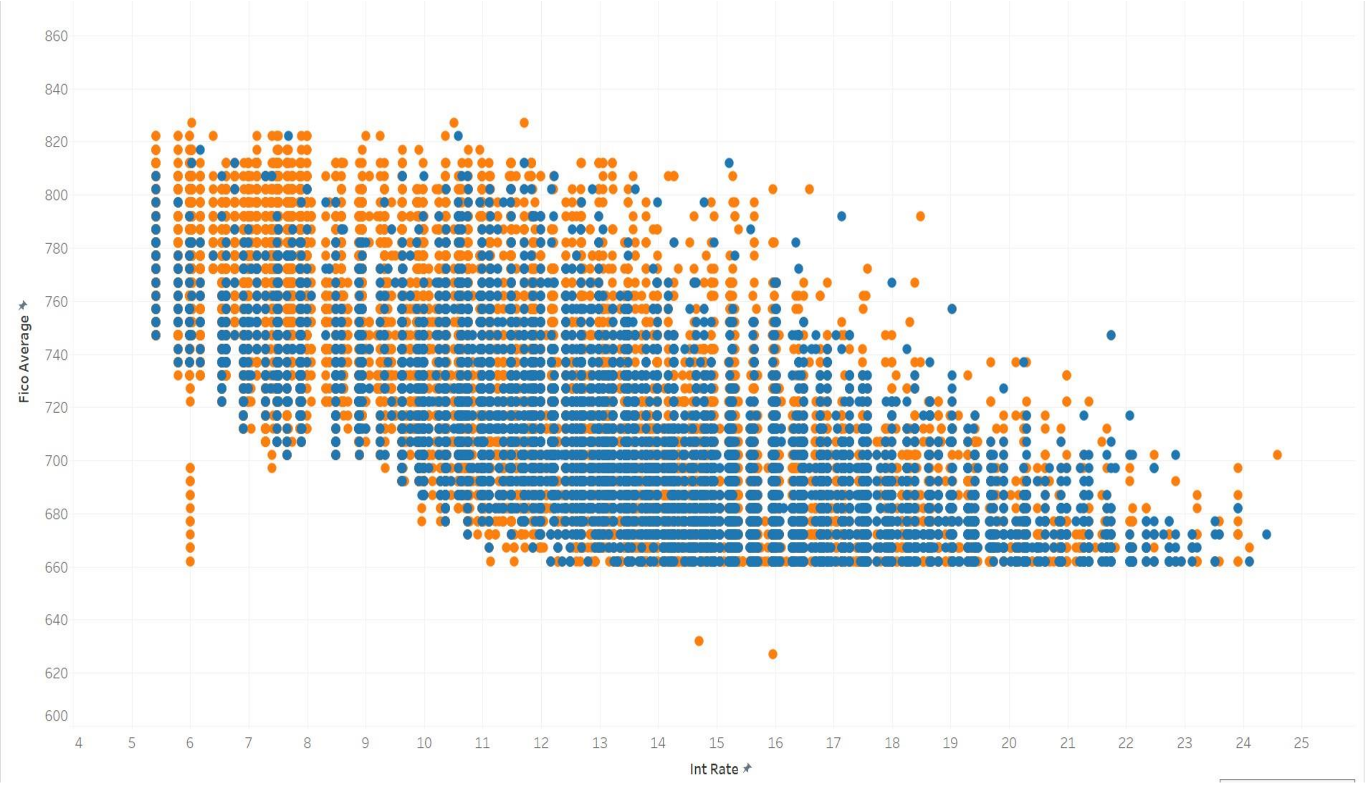


Fig :- Borrower’s with higher interest rate and low FICO score have higher chance of being defaulter.

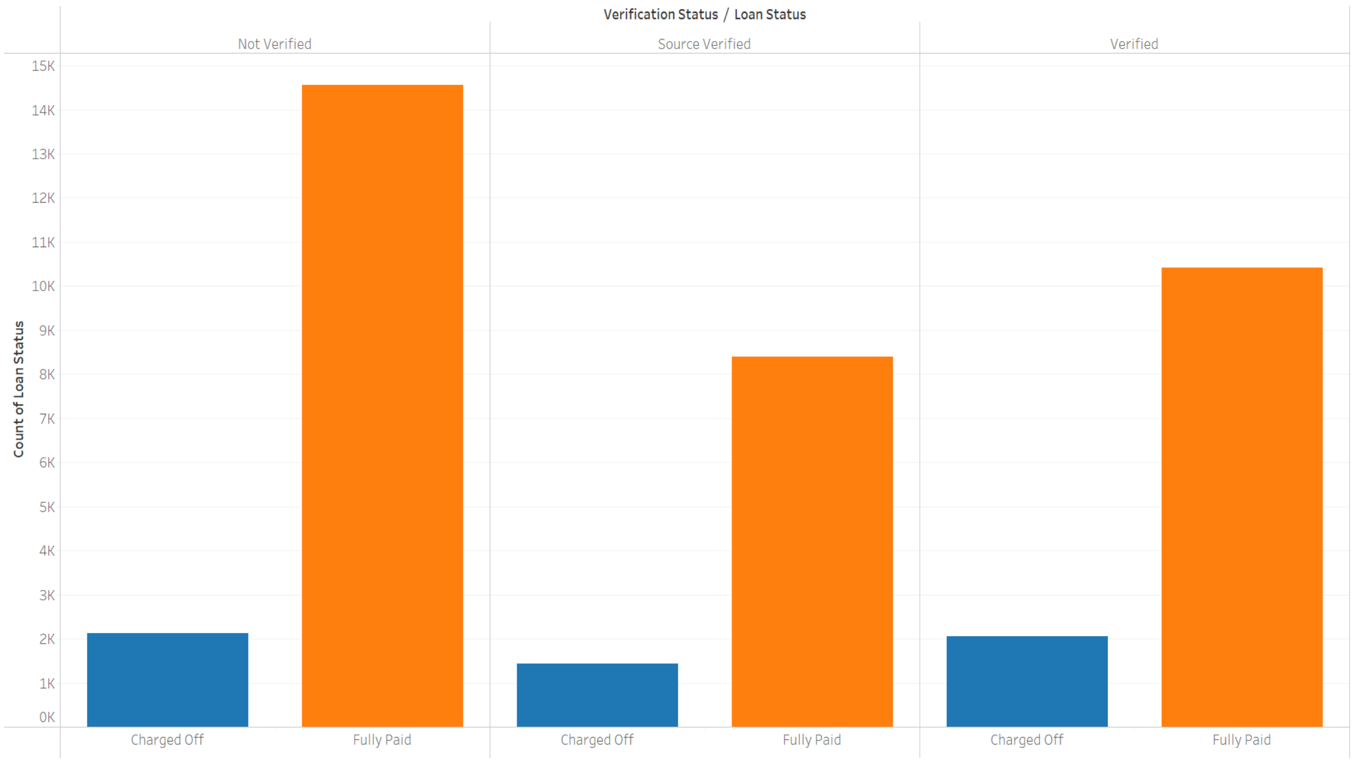


Fig. - The verification process can be improved based on the number of not verified and verified borrower’s.

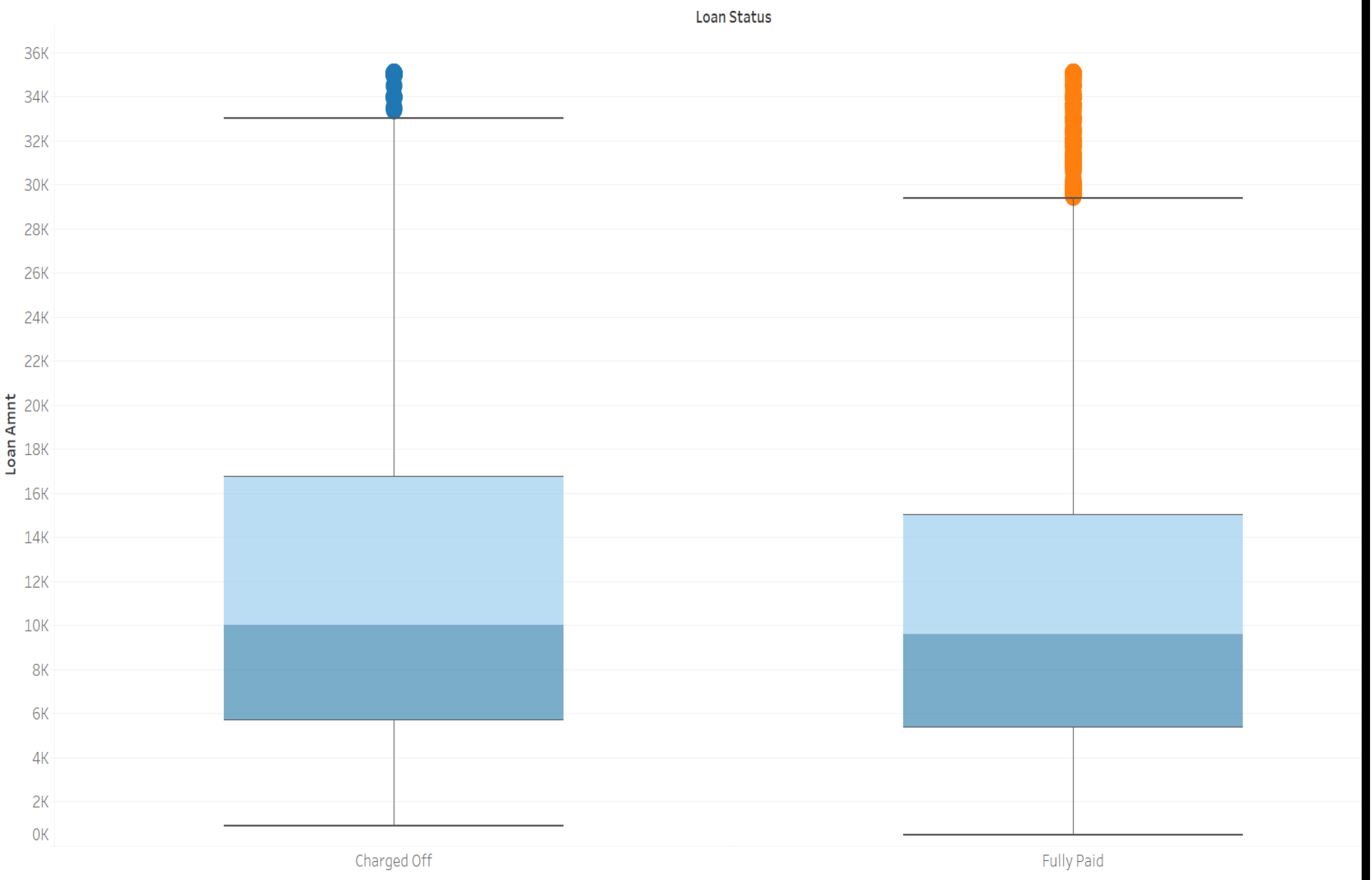


Fig. - There is no significant impact of loan amount on loan status.

# 8. CONCLUSION AND FUTURE SCOPE

After analyzing and modelling the Lending Club dataset, we can conclude that machine learning can be effectively used in building a credit risk minimizing model for P2P lending businesses.

We have demonstrated how to pre-process the data by handling missing values, removing outliers, scaling numerical features, and encoding categorical variables using one-hot encoding. Furthermore, we have used techniques such as under sampling and hyperparameter tuning to improve the performance of the machine learning models.

After implementing different model, we came to the conclusion that Gradient Boosting model is best fit for predicting the customer will be a Fully Paid or Charged Off. Moreover, we have used evaluation metrics such as ROC AUC score to determine the performance of our models.

Overall, we can conclude that using machine learning algorithms can help lenders make informed decisions by predicting the credit risk of potential borrowers. This can lead to a reduction in default rates and an increase in profitability for P2P lending businesses.

Potential future directions for improving the credit risk minimizing model in P2P lending business using machine learning on Lending Club dataset:

Real-time monitoring: Developing a monitoring system to track the performance of the model in real-time and to provide updates to the stakeholders on the model's effectiveness.

Explaining model's prediction: The model's prediction can be made more interpretable to help the stakeholders understand the factors contributing to the borrower's credit risk.

# References

1. https://data.world/jaypeedevlin/lending-club-loan-data-2007-11

2. https://www.lendingclub.com/investing/peer-to-peer

3. https://spark.apache.org/mllib/

4. https://spark.apache.org/docs/latest/api/python/

5. https://www.nature.com/articles/s41598-021-98361-6

6. https://files.eric.ed.gov/fulltext/EJ1279989.pdf