INSTITUTE FOR ADVANCED COMPUTING AND SOFTWARE DEVELOPMENT, AKURDI, PUNE

Customer Profile Management for BFS

PG-DBDA March 2024

Submitted By:
Group No: 27

Roll No.	Name.
243527	Moha Nalawade

Mr. Abhijit Nagargoje
Project Guide

Mr. Rohit Puranik
Centre Coordinator

ABSTRACT

This study focuses on developing an advanced forecasting model to predict default risk in financial lending, leveraging several machine learning techniques.

Accurate prediction of default risk is crucial for financial institutions to minimize losses and make informed lending decisions. Traditional risk assessment methods often fall short in capturing complex patterns and interactions within the data.

The purpose of this project is to enhance default risk prediction by employing a range of machine learning models, including Random Forest, Artificial Neural Networks (ANN), and XGBoost. By applying rigorous cross-validation and voting ensemble learning to fine-tune model hyperparameters, the study aims to identify the most effective model for predicting default risk. The dataset, sourced from Lending Club, is meticulously cleaned, normalized, and standardized to ensure optimal model performance. The findings demonstrate that XGBoost outperforms the other models, offering superior predictive accuracy and reliability in assessing default risk. This work contributes to the development of more accurate risk assessment tools in the financial sector.

ACKNOWLEDGEMENT

I would like to express my profound appreciation to the Almighty for His divine blessings and grace, which have played a crucial role in guiding our endeavor to a triumphant conclusion.. I extend my sincere and heartfelt thanks to our esteemed Course Co-Ordinator, Dr. Shantanu Pathak for providing me with the right guidance and advice at the crucial juncture sand for showing me the right way. I extend my sincere thanks to our respected Centre Co-Ordinator Mr. Rohit Puranik, for allowing us to use the facilities available. I would like to thank the other faculty members also, at this occasion. Last but not the least, I would like to thank my friends and family for the support and encouragement they have given me during the course of our work.

Moha Nalawade (240341225026)

Table of Contents

Sr. No	Description	Page No.
1	Introduction	1
2	Requirements	4
3	Dataset	5
4	Design	7
5	Machine Learning Models	12
6	Model Performance	16
7	Conclusion	17
8	References	18

1. INTRODUCTION

1.1 Machine learning and finance:

When referring to the early Artificial Intelligence (AI) systems, which Arthur Samuel introduced in 1959 while working for IBM, the phrase "machine learning" was primarily used to give examples of pattern recognition exercises that have a "learning" element. During that early period of time, wider AI system was the dominant field under which including the idea of Machine Learning systems. After that came the broaden of the range of practical applications of Machine, and jumped out of the limit which defined by former frame of AI.

The earliest instance of machine learning being used to address economic and financial issues dates back to 1974 [1]. While the first paper published not only applied Machine Learning models but also used them exclusively, of which the topic was about economics and finance came out in 1984 [2].

1.2 Lending club & related works:

Lending Club, one of the top peer-to-peer (P2P) lending service providers, is headquartered in San Francisco, California, and is located in the United States. A P2P lending firm registered its loan offers as securities with the Securities and Exchange Commission for the first time at that time (SEC) that made it the pioneer of its peers. Lending Club is also a loan-trading service provider on the secondary market. Its main operation is allowing investors to buy notes secured by loan payments, enabling borrowers to access loans on the basis of an online platform. Given the convenience of searching and browsing each loan that listed on Lending Club website, the investors of loans can also select the loans they prefer to invest after their acknowledge of the information that contributed by the borrower, for example loan amount, the grade of their new loans, and why they apply for loans. The main income of the investors comes from the interest collected from each of these loans. And Lending Club charges both borrowers and investors fees which called origination fee and service fee separately.

P2P investment has its own advantages comparing with traditional banks. From borrower's perspective, the final production of the two organizations is the same which is loan. From lender's perspective, however, there exist difference between the business of the online platforms and the traditional banks. Thanks to explicit and implicit government guarantees which give benefit to banks which would result in rather lower funding-cost, while P2P platforms is a game-changer which in their operation they applied several new technologies that could lower costs of operation and risks in pricing and avoid legacy problems which can be aggravated by crisis. P2P lending platforms' business remove the intermediation of financial institutions [3] thus showing their innovations in both loan and investment field. For those who may not have availability to standard financial intermediaries P2P loans give the borrowers shortcuts to financing [4]. Why P2P lending platforms are able to compete with traditional lending markets as well as attracting so large number of investors, one of the main reasons could be its super ability that meet the demand of nowadays capitals [5].

Used historical data comes from the Lending Club website to analyze the operation of P2P company based on. The main risk of the project is default risk. A P2P company can make profit if it clearly knows the percentage of borrowers who will not pay the money back. When receiving a loan application, a decision has to be made by the company whether to approve the new loan or based on the information of loan appliers. Risks can be divided into two types in decision making. For the first one, if the loan is likely to be repaid, then a loss of business would come into exist to the company if the loan is not approved. One the other hand, if the loan is likely not to be repaid, then the borrower likely to default, then approving the loan can be a financial loss to the platform. Applications of ML principles are made in the implementation of a neural network with backpropagation to assess default risk, or the likelihood that a borrower will default [6]. The success rate of loans is inversely correlated with the borrowers' credit ratings [7]. Our classification models are intended to aid in the decision of whether or not to invest in P2P lending to a borrower.

Fig. 1. Lending Club's advertisement poster

1.3 Study overview:

Our aim is to use models to forecast whether an applicant for loan is going to default or not. Dataset is obtained from Lending Club, and divided into training and testing datasets, then apply them to various machine learning models. Finally, performance of the models is evaluated.

2. REQUIREMENTS

2.1 Hardware Specifications

- Machine: Desktop/Laptop
- Operating system: Windows 10 (or above) or Linux 18 LTS (or above) Processors: Intel core i5 or AMD 5 series (minimum)
- Memory: 8 GB RAM or above Hard Disk (SSD): 250 GB or more
- Video Card (optional): Intel Integrated Graphics (suggested – 4 GB graphics card - NVIDIA)
- Network: Ethernet / Wi-Fi with 25 Mbps Speed Connection (UL/DL)

2.2 Software Specifications

- Language: Python 3.7 or above (stable build)Platform: Anaconda Latest stable build
- Notebooks: Jupyter and Google Colab
- Libraries: , Pandas, Numpy, Matplotlib, Seaborn, Keras, Tensorflow, Scikit Learn

3. DATASET

The used dataset is collected from January 2007 to December 2018 which covers a very long time period and make sure our results are universally applicable. The original dataset contains 611780 observations and 27 features. Followed by data preprocessing procedure, the features are reduced to 23.

Figure 3 illustrates the correlation heatmap of the features which are originally numeric and some features such as total payment and loan amount, total payment for investors and funded amount for investors, interest received to date and loan amount, principal received to date and funded amount. This could be because investors require higher payment and interest payment when they invest higher amount of money. Loan statues and issued numbers turn into loan volume of the dataset is in the Table 1.

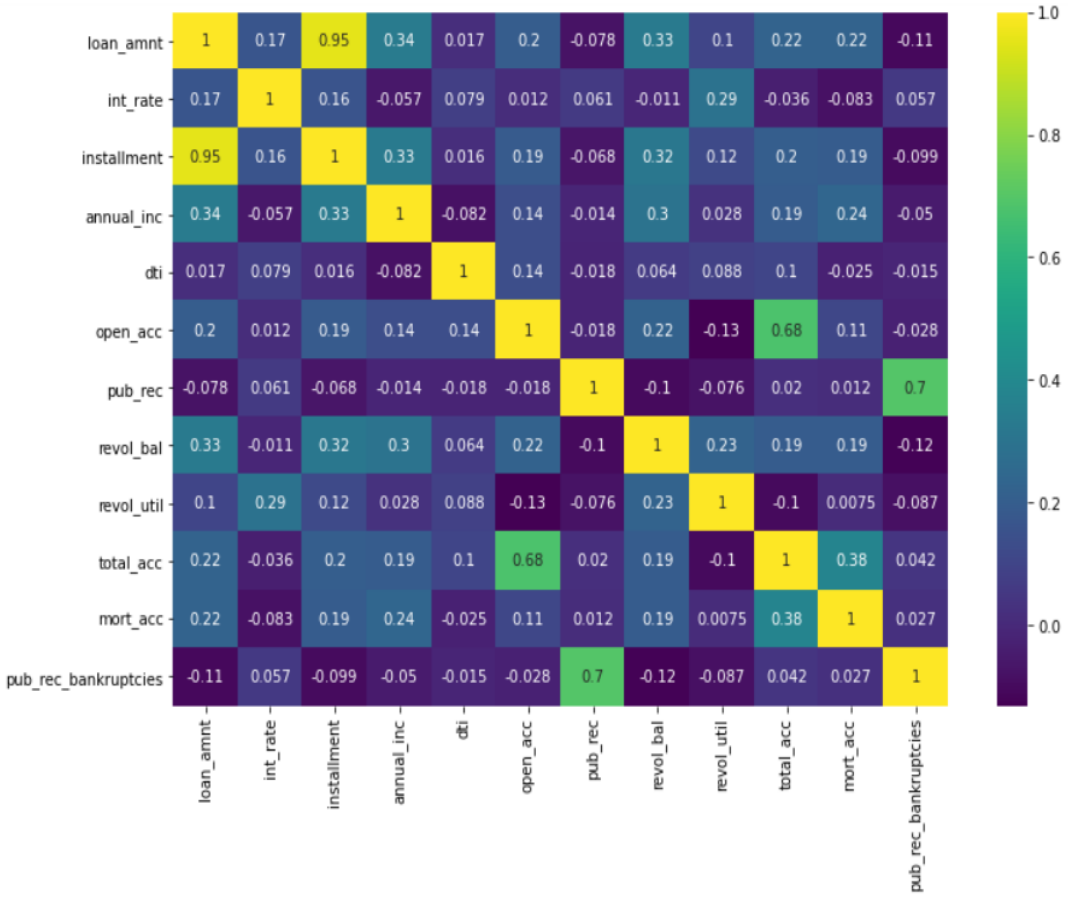

Fig. 3. The correlation matrix of features for the Lending Club dataset

As shown in Table 1, there are instances that borrowers fully paid their loans has significantly greater observation than instances that borrowers charged off. Our categorization models may have issues as a result of the two groups' imbalance because they have a tendency to favor the majority class which is Fully-Paid, and would result in models' overfitting. Table 2 shows the 19 features after reduction and their descriptions.

Table 1. Loan volume broken down by loan status

Loan Status	Loan Volume
Fully-paid	**318357**
Charged-Off	**77673**
Total	396030

Table 2. Features and their descriptions.

Feature Variable	**Description**
Address state	State indicated in the loan application by the borrower
Dti	Total monthly payments made by borrower on outstanding debt obligations
Funded amount	The entire loan amount committed at the time
Funded amnt invt	The entire loan amount committed by investors at the time
Grade	Loan grade assigned by Lending Club
Home ownership	The borrower's declaration of home ownership during registration
Inq last 6mths	The quantity of creditors' queries over the last six months
Installment	The borrower's recurring monthly payment if the loan originates
Interest rate	Rate of interest for the loan
Loan amount	The borrower's indicated loan request's total amount
Open accounts	The quantity of credit-line that are open in the borrower's credit report
Out princpal	Principal that is still owing for the entire amount funded
Out princpal investor	Principal that is still owing for the entire amount funded funded by investors
Public record	Quantity of negative public records
Purpose	A classification of "why" problem of the money borrower for the loans
Revol balance	Total outstanding revolving credit
Revolving utility	Revolving line use percentage
Term	A loan's total number of payments
Verification status	Indicates whether Lending Club has confirmed or not

4. DESIGN

4.1 System Architecture:

Figure 2 illustrates the structure of the methodology. The data comes from Lending Club, and dataset is preprocessed into a desirable one for machine learning models with few steps. Then having the processed data, which is divided into training and testing data of a 75/25 split. The training part is applied to various machine learning models. Performance of models is tested by applying testing data to them and then evaluate the models. Finally, summary of the result and output of data and models is presented.

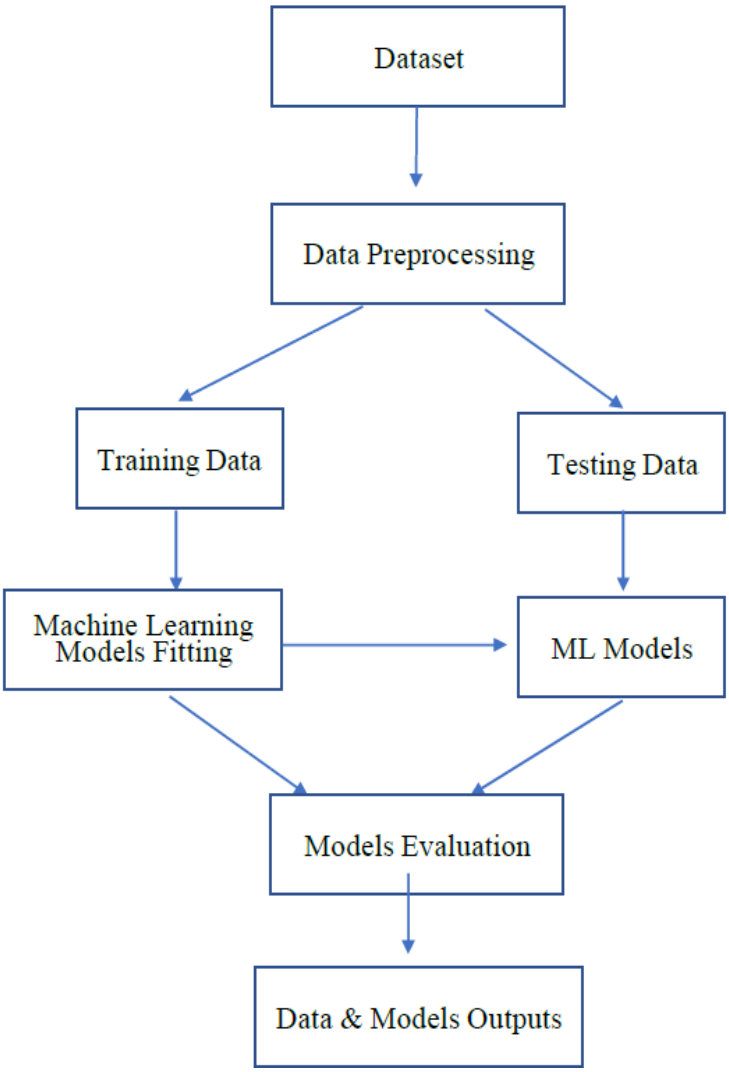

Fig. 2 Structure of the methodology

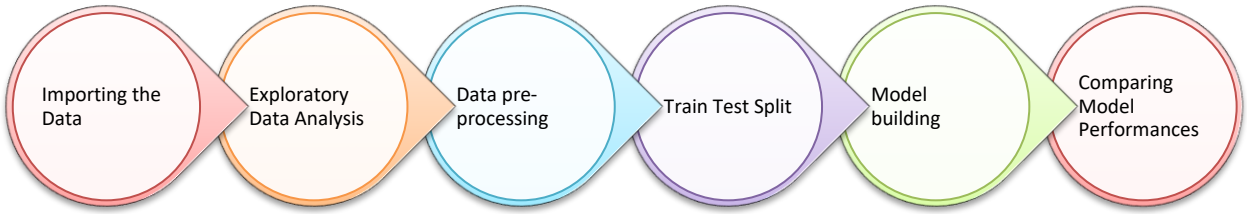

Fig 3. Data Flow Diagram

4.1 Exploratory Data Analysis:

Fig 4 shows that, currently, 24.5% of the total loans have been charged off, indicating a significant portion of loans deemed unlikely to be collected. This percentage is an important metric for assessing the institution's credit risk and overall financial health. Further analysis may be required to understand the underlying causes and implement appropriate risk mitigation strategies.

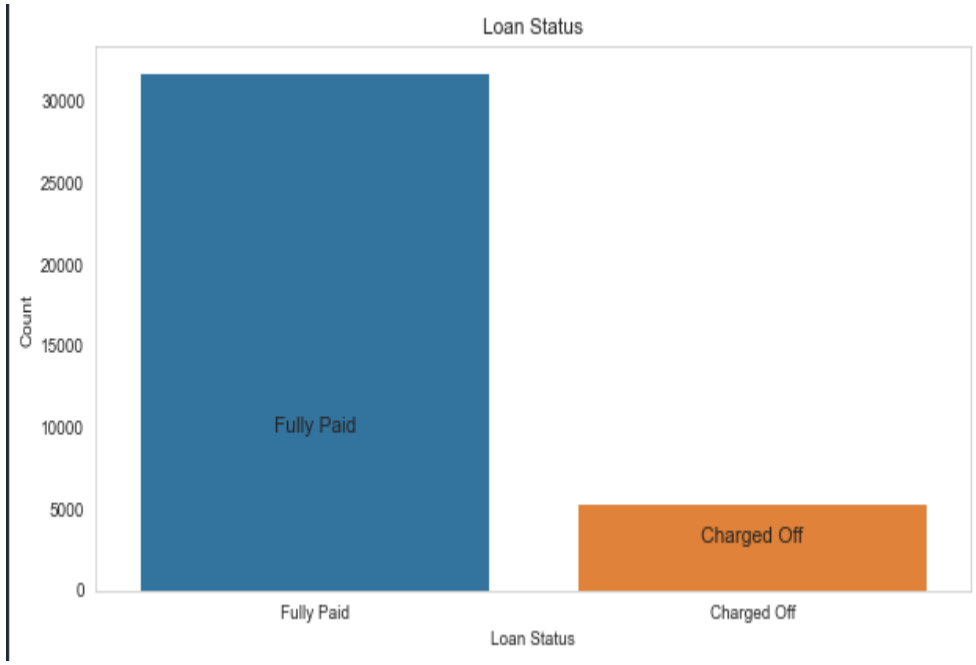

Fig 4. Loan Status: The number of charged off loan is smaller(24.5%) compared to total count

- grade: LC assigned loan grade
- sub_grade: LC assigned loan subgrade

Fig. 5 explore the Grade and SubGrade columns that LendingClub attributes to the loans. It looks like `F` and `G` subgrades don't get paid back that often. Isloate those and recreate the countplot just for those subgrades.

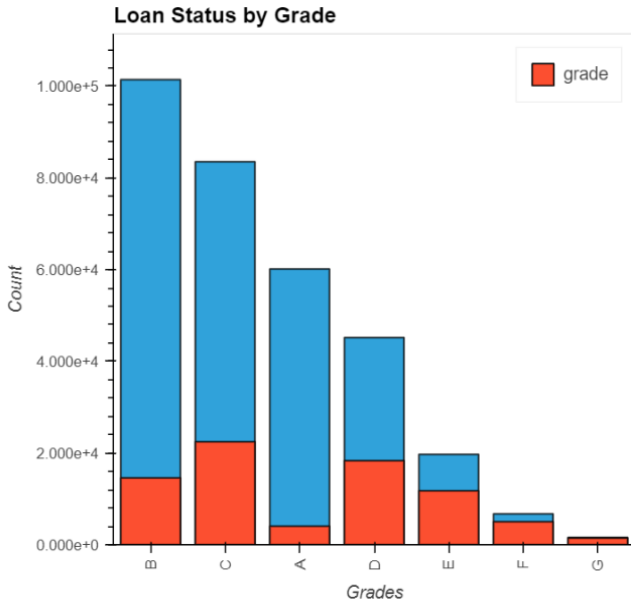

Fig 5.a Loan Status By Grades

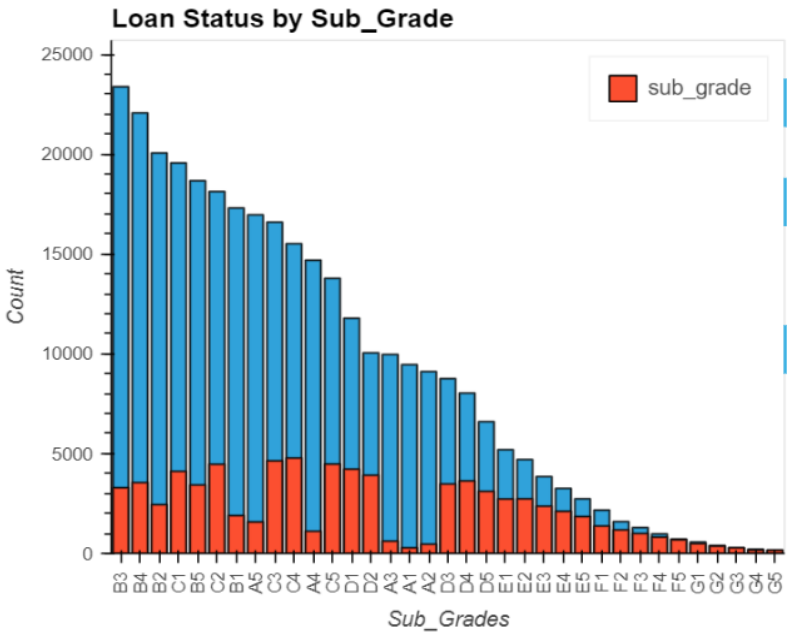

Fig 5 b Loan Status by Subgrades

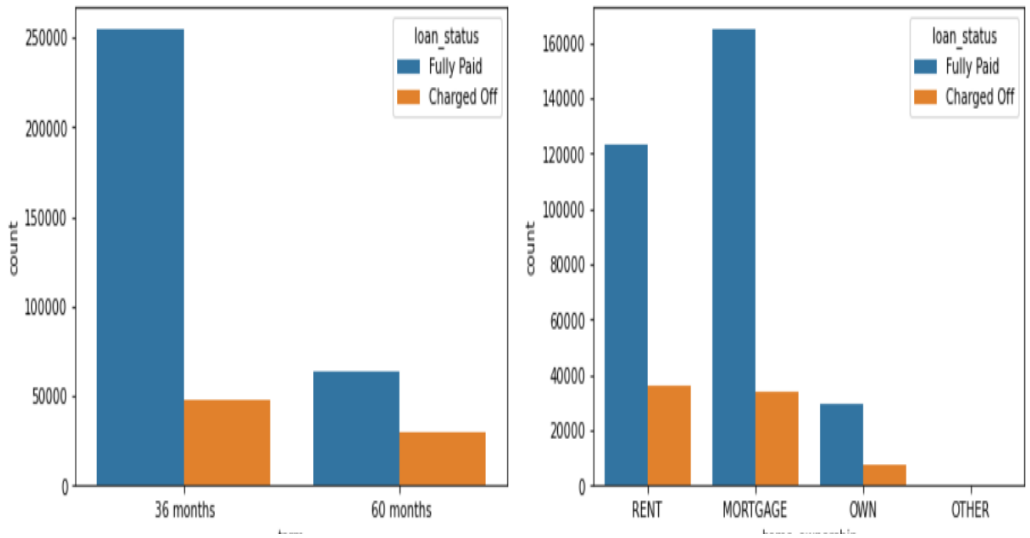

Fig 6 Loan term vs loan status (left) and House Ownership vs loan status (right)

Comparing loan term with loan status could help you understand if the length of the loan term impacts its status. For example, you might find that shorter-term loans are more likely to be paid off successfully, while longer-term loans might have a higher chance of default. We could also identify trends, such as if loans with very long terms are more prone to becoming overdue or defaulting

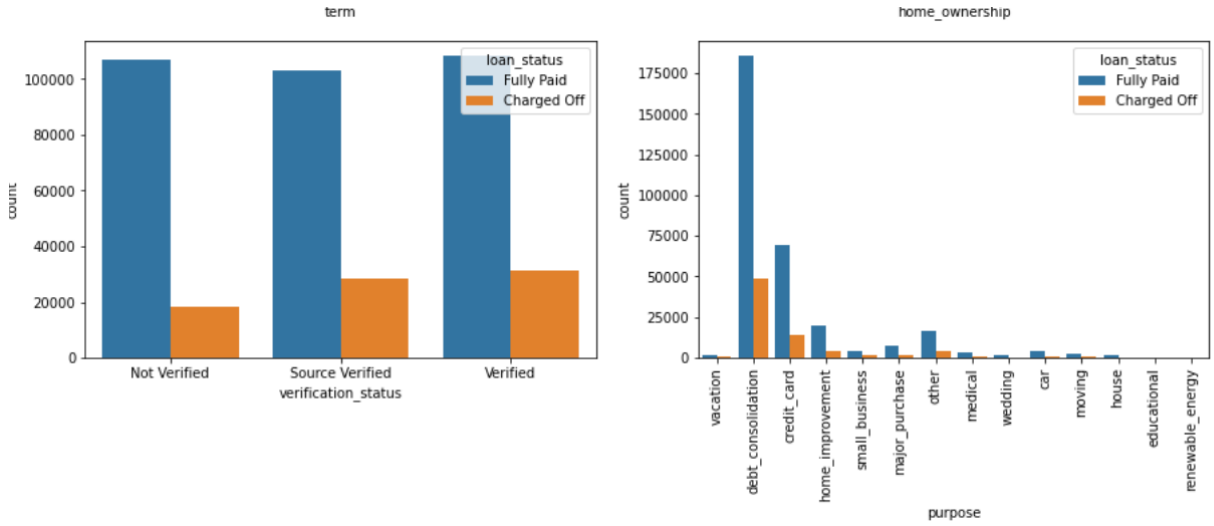

Fig 7 Verification Status vs loan status (left) and Purpose of loan vs loan status (right)

Verification Status vs. Loan Status: This analysis can reveal if proper verification is crucial for loan management and performance. Purpose of Loan vs. Loan Status: This comparison helps understand how different reasons for borrowing impact loan outcomes, which can be useful for tailoring loan products or assessing risk.

4.1.1. Recommendations:

Major factor which can be used to predict the chance of defaulting and avoiding Credit Loss:

1. DTI
2. Grades
3. Verification Status
4. Annual income
5. Pub_rec_bankruptcies

Other considerations for 'defaults' :

1. Burrowers having annual income in the range 50000-100000.
2. Burrowers having Public Recorded Bankruptcy.
3. Burrowers with least grades like E,F,G which indicates high risk.
4. Burrowers with very high Debt to Income value.
5. Burrowers with working experience 10+ years

4.2 . Data preprocessing

For those features which are object, need be turned into numeric with mapping function, counting the unique values of each feature and then map them with integer starting from 0. Features including term, loan grade, home ownership, verification status, purpose, address state and loan status went through this procedure. For the high correlation features, we need to drop them which includes total payment, total payment for investors, interest received to date and principal received to date. We also need to drop NA. Now, the mean and scaling to unit variance were removed from all of the numerical variables by using normalizing and standardizing. After that the dataset is divided into training and testing data which is 75/25. Our y is loan status which has binary values: ‘0’ as Fully Paid and ‘1’ as Charge Off, and X is the rest other features.

The loan status of the dataset is shown in Figure 4, and loan status distributions about other six features including term, credit grade, employment length, home ownership, verification status and loan purpose are shown respectively in Figure 5. For example, a greater proportion of 36-month loan in Fully Paid group is obtained, grade A, B, C are the majority in Fully Paid group, etc.

5. MACHINE LEARNING MODELS

The project uses a variety of machine learning models, such as logistic regression, random forest, gaussian naive bayes, and artificial neural networks, SVM, XGBoost. These are the typical supervised learning models. Cross validation and voting ensemble are also introduced.

In machine learning, a confusion matrix is among the best performance measures. The confusion matrix structure for binary classification is shown in Figure 8.

		Actual Values	
		Positive	**Negative**
Predicted Values	**Positive**	**True Positive (TP)**	**False Positive (FP)**
	Negative	**False Negative (FN)**	**False Negative (FN)**

Fig 8. Actual and predicted confusion-matrix

In terms of performance evaluation, accuracy, recall, precision, F1 score , and calculation formulas for the four parameters are listed as follow, from equation (1) to equation (4).

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (1)$$

$$Recall = \frac{TP}{TP + FN} \quad (2)$$

$$Precision = \frac{TP}{TP + FP} \quad (3)$$

$$F1 = \frac{2}{\frac{1}{Precision} + \frac{1}{Recall}} \quad (4)$$

5.1 Random Forest

Random Forest is a robust ensemble learning method that builds upon the foundation of decision trees. Specifically, it utilizes a collection of decision trees, each constructed through a process known as bagging (Bootstrap Aggregating). This technique involves training each tree on a randomly sampled subset of the training data to enhance model performance and stability.

The primary objective of each decision tree within a Random Forest is to identify the optimal splits in the data, which helps in classifying or predicting outcomes. These trees are typically built using the Classification and Regression Tree (CART) algorithm. CART recursively partitions the data by selecting splits that best separate the target classes or outcomes.

To evaluate the effectiveness of these splits, Random Forest employs metrics such as Gini impurity. Gini impurity measures the degree of impurity or disorder in a dataset; a lower Gini impurity indicates a purer split, which is desirable for better classification performance.

By aggregating the predictions from multiple decision trees, Random Forest achieves improved accuracy and robustness, reducing the risk of overfitting and enhancing overall model reliability.

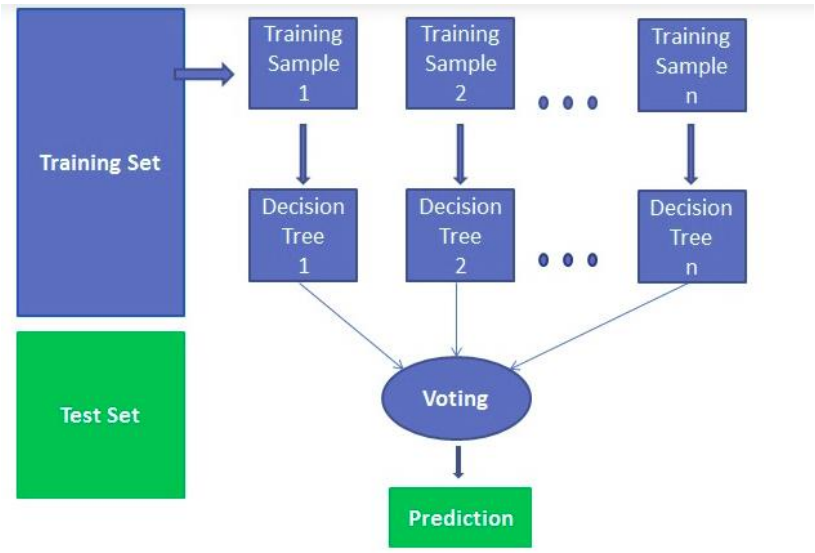

Fig. 9 Random forest Classifier

5.2 Artificial Neural Network:

To explain the foundation of an Artificial Neural Network (ANN). It is a group of interconnected artificial neurons. A synthetic neuron receives impulses, analyses them, and then sends messages to neurons that are connected to it. The message delivered at such connection is always a real-number. The output of each neuron is determined by a nonlinear-function of the inputs' totals. Edges are the names for the connections between the layers. As learning occurs, the machine always changes the weight of neurons and edges each time. The purpose of wight is to modify whether to strengthen or weaken one of the many signals at a connection in the layers. The widely accepted industrial standard, logistic regression, is outperformed by artificial neural networks [8, 9].

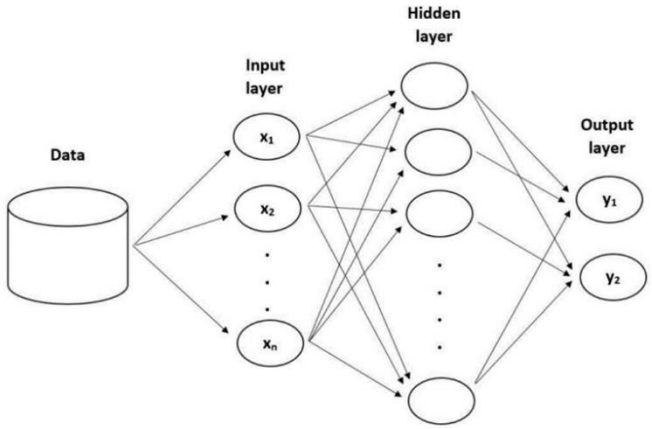

Fig. 10 ANN's multi-layer structure

The ANN algorithm was developed with inspiration from the human brain [10]. In an ANN, primarily, there are always three different kinds of layers. For the first layer which is so-called input- layer. Every input variable in this dataset will be represented by every neuron in the input-layer. Following that, the neurons in a new layer receive a modification that has an activation function from these changed neurons. There may be more than one Hidden layer, which is the name of the second- layer(s). The machine processed neurons from the hidden layer. After that, these final results are transferred to the top-layer. The output layer, the last layer, is made up of neurons that either belong to the final classes or carry prediction information, such probability predictions. Figure 5 presents a simple structure of ANN.

5.3 XGBoost

A scalable and dispersed Gradient Boosting Decision Tree (GBDT) ML system is called Extreme Gradient Boosting (XGBoost). By examining a tree of "if-then-else" T/F feature questions, then calculating and finding out the bare minimum of number of these questions that needed to determine the chance of selecting the right answer. In such procedures, decision trees create a model which can predicts labels. Figure 8 shows the structure of a simple XGBoost model**.**

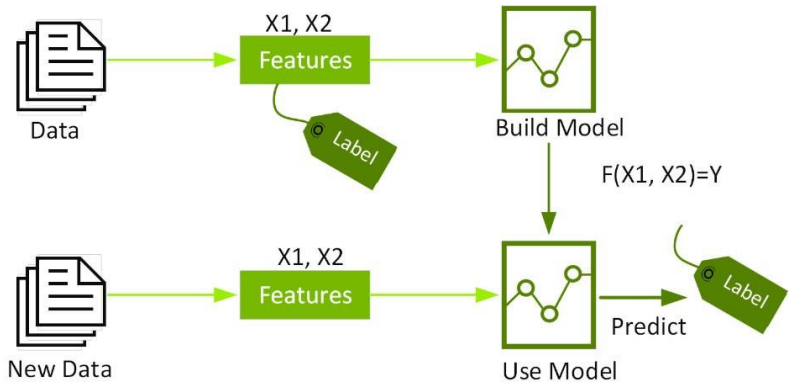

Fig. 11. XGBoost Structure [https://www.nvidia.com/en-us/glossary/data-science/xgboost/]

6. MODEL PERFORMANCE

Accuracy, recall, precision and F1 score are considered for each algorithm using the matching optimal hyperparameter set in order to evaluate the prediction models for default. Performance of the models are shown in Table 3 and Figure 8. For neural network, the hidden layer was set 150. The best performing model is XGBoost which gives the highest predicting accuracy. The three models were also subjected to a three-fold cross validation, the results of which are shown in Table 4.

Table 3. Model performance

Algorithms	Accuracy	Recall	Precision	F1 score
Neural Network	**88.86%**	**46%**	**93%**	**62%**
Random Forest	**88.93%**	**46%**	**95%**	**62%**
XGBoost	**88.94%**	**48%**	**91%**	**63%**

Table 4. Cross validation performance.

Algorithms	Accuracy
Random_Forest	**88.93%**
Neural_Network	**88.86%**
XGBoost	**88.94%**

Fig 12. Comparison for ROC Scores of ML Models

7. CONCLUSION

Conclusion

The project tried several machine learning classifiers as well as cross validation and voting ensemble in the classification case about credit risk within P2P market taking Lending Club as example. The main contribution of our study is finding the best model predicting whether a loan will default or not and ranking of different classification models and methods, which displayed in Tables 3, 4 and 5, based on our Lending Club dataset. We tried to find out the best performing model thus with given data of money borrower we can forecast whether he/she will default in the future or not which is very important in risk management of P2P loaning project.

Random Forest, Artificial Neural Networks, XGBoost models are used in our study, which also runs a three-fold cross validation method and a voting ensemble. The results show that among the models and methods, XGBoost is the best performing model as it gives the highest accuracy in our evaluation testing. But it is also shown that the accuracy of the six models is very close to each other so that the advantage of XGBoost is not very obvious.

8. REFFERENCES

1. Lee, Samuel C., and Edward T. Lee. "Fuzzy sets and neural networks." (1974): 83-103.

2. Wang, H., Li, C., Gu, B., & Min, W. (1984). “Does AI-based credit scoring improve financial inclusion? Evidence from online payday lending”. In 40th international conference on information systems. ICIS 2019.

3. Chen, D., & Han, C. (2012). A comparative study of online P2P lending in the USA and China. Journal of Internet Banking and Commerce, 17(2), 1.

4. Reddy, S., & Gopalaraman, K. (2016). Peer to peer lending, default prediction-evidence from lending club. The Journal of Internet Banking and Commerce, 21(3).

5. Mills KG, McCarthy B (2016) The State of Small Business Lending: Innovation and Technology and the Implications for Regulation. HBS Working Paper No. 17-042.

6. Zhang, Z., Gao, G., & Shi, Y. (2014). Credit risk evaluation using multi-criteria optimization classifier with kernel, fuzzification and penalty factors. European Journal of Operational Research, 237(1), 335- 348.

7. Freedman, S., & Jin, G. Z. (2017). The information value of online social networks: Lessons from peer- to-peer lending. International Journal of Industrial Organization, 51, 185-222.

8. Lessmann, S., B. Baesens, H. V. Seow, & L. C. Thomas (2015): “Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research.” European Journal of Operational Research 247(1): pp.124–136