Educational Administration: Theory and Practice
2024, 30(5), 12975-12989
ISSN: 2148-2403
https://kuey.net/

Research Article

Machine Learning-Based Loan Default Prediction: Models, Insights, And Performance Evaluation In Peer-To-Peer Lending Platforms

E.Srinivas Jayaram[1], Dr.G.Balachandar[2], Dr. KompalliSasi Kumar[3]

[1*]Research Scholar, Department of Business Administration, Annamalai University, Tamilnadu, es.jairam@gmail.com, 7093025442
[2]Assistant Professor, Department of Business Administration, Govt. Arts and Science College for Women, Alangulam, Tenkasi, Tamilnadu, aubalachandar@gmail.com, 9952636466
[3]Associate Professor, GITAM School of Business, Hyderabad, GITAM University, skompall@gitam.edu, 9848192864

Citation: E.Srinivas Jayaram, et al. (2024), Machine Learning-Based Loan Default Prediction: Models, Insights, And Performance Evaluation In Peer-To-Peer Lending Platforms, Educational Administration: Theory and Practice, 30(5), 12975-12989
Doi: 10.53555/kuey.v30i5.5637

ARTICLE INFO

ABSTRACT

In financial institutions, mitigating capital loss is paramount, especially when considering the risks of extending loans. It's crucial to analyze potential dangers and thoroughly assess the likelihood of default to address these risks. Despite possessing extensive customer behavior data, financial institutions often need help accurately predict the loan default probabilities. Data mining, a rapidly advancing field in data analysis, offers promising solutions by extracting valuable insights from complex datasets. This research aims to develop and prototype a classification model based on deep learning algorithms, leveraging tools provided in the statistical tool Python. We preprocess the raw dataset to remove unimportant dimensions, detect outliers, remove them, input missing values, and normalize data to enhance prediction accuracy. Once we develop the model, we implement it to predict outcomes using a test dataset. Experimental findings confirm its accuracy in forecasting loan defaults.

Keywords: Loan Default Prediction, Machine Learning Algorithms, Peer-to-Peer Lending, Predictive Analytics, Default Risk Modeling, Default Forecasting

1. Introduction:

Crowdfunding and peer-to-peer (P2P) lending have existed for over ten years. Today, numerous P2P lending platforms exist in emerging and developed nations. A more thorough comprehension is required, not merely understanding the complexities of successful peer-to-peer lending and how P2P lending mechanisms are used and distributed. Loans made using peer-to-peer (P2P) systems include default risks, and this research looks into such aspects. Increased interest rates and inflation are associated with a heightened probability of default in these lending markets. The findings of this research apply to the regional and across the nation of P2P lending as is.

Studies on these P2P lending practices highlight the various opportunities for sharing efficient procedures while encouraging new financial lending methods. Researchers are looking directly at these peer-to-peer lending platforms for gaining knowledge, but most are theoretical rather than empirical. Prior studies have neglected the trends and essential variables responsible primarily for the growth of P2P lending platforms.

The decade that followed the financial crisis saw an increase in other lending practices. Among these, Entrepreneurial start-ups now have a new way to get the funding they need: peer-to-peer lending (Oren, 2013; Zhang, Baeck, Ziegler, Bone, & Garvey, 2016). In this context, investor capital and other traditional types of venture capital do not constitute access to cash. Individual lenders can combine their resources to support financing requests from individuals and businesses using online peer-to-peer lending platforms. Eliminating intermediaries distinguishes P2P lending and has contributed to its widespread popularity. As with every different lending contract, a lender is liable for providing required finance over the loan's term, while the borrower agrees to repay the funds with interest as specified in the loan agreement.

However, these peer-to-peer lending platforms serve as intermediaries, verifying borrowers' creditworthiness and facilitating connections with potential lenders. Consequently, peer-to-peer (P2P) lending constitutes a direct arrangement for loans between the lender and borrower, with these lending platforms serving solely as facilitators. As a result, these marketplaces have evolved as novel sources of financing due to their ability to eliminate mediators from conventional lending practices. Due to the attractive and consistent returns offered, these platforms attracted investors eager for higher yields, aiming to diversify their portfolios through alternative investment opportunities.

There are a considerable number of P2P lending platforms operating around the world. Since 2010, seventy-nineP2P European Union platforms have raised approximately Seventeen billion pounds (ORCAMoney, 2017). The US lending market was worth over two billion dollars in 2018, an increase of forty percent over 2017 (CCAF, 2020). China accounts for the majority of the market share in Asia-Pacific. Before 2010, the Chinese alternative finance market was almost nonexistent, but by 2015, it had grown to phenomenal lending, amounting to more than a hundred billion dollars (Zhang et al., 2016). However, the Asia-Pacific market, including Australia, grew more than three hundred percent between 2014 and 2015 (Zhang et al., 2016).

Despite experiencing rapid expansion in recent years, P2P still constitutes a minuscule portion of the total lending. According to Milne and Parboteeah (2016), peer-to-peer lending accounts for less than one percent of all unsecured consumer finance and half a percent of all borrowing from small and medium-sized enterprises (SMEs) in nations like the UK, where it is expanding at the highest rate. Furthermore, the rapid expansion of this market points to a favorable business environment as opposed to the robustness of P2P lending markets. Since the Global Financial Crisis, the financial services sector has been sailing in safe waters, with interest rates that are low and mostly stable. Despite the impressive growth rates, the industry's future is uncertain due to its reliance on complex and linked macroeconomic indices(J. Li, Hsu, Chen, & Chen, 2016). According to Tomlinson et al. (2016), the interest rate alone may result in a divergence in P2P lending penetration in the United Kingdom's marketplace between £0.5 billion and £35.5 billion by 2025.

Section One of this study touches upon P2P lending and the concept of risk rating methods while explaining the significance of the subject matter. Section Two comprehensively analyzes pertinent literature, highlighting critical studies and presenting comparable research findings. The third portion commences with a comprehensive elucidation of the algorithms employed in this study and assessing their prediction efficacy, followed by an elucidation of the experimental design and an exhaustive depiction of the data, including the procedures undertaken during the data visualization and preliminary investigation, cleansing, wrangling, and model development. In the study, section Four will address the performance evaluation measures. Section five presents the practical outcomes, and inferences from these outcomes are derived. Finally, we will analyze the documentation of the constraints of this study and propose suggestions for future endeavors. We conclude the study by providing the Citations at the end.

2. Literature Review

The risk of default is an inherent component of lending. Financial institutions prioritize minimizing information asymmetry between lenders and borrowers to reduce the number of problematic loans. In their study, Lahsasna et al., 2010 argued that the profitability of financial organizations relies heavily on credit risk decisions, as making incorrect decisions can result in significant expenses. Wu et al., 2010 stated that borrower risk evaluations form the core of the default risk analysis on which the investor will decide to fund the loan application. Given the significance of risk analysis in lending, experts propose employing various tools to enhance risk assessment and boost the accuracy of credit risk prediction. Credit scoring aims to categorize consumers into two groups: good customers, who have successfully repaid their loans, and poor customers, who have defaulted on their loans. Researchers employed machine learning and deep learning techniques like decision trees, ensemble models, and neural networks to compute credit risk.

An effective credit scoring methodology ensures financial institutions' sustained performance, particularly these P2P lending platforms. Leyla Mammadova (2021) investigated default risk, considering both macroeconomic factors and conventional drivers, including credit scoring, debt-to-income ratio, employment prospects, grade, loan amount, annual income, and borrower's creditworthiness. Hand and Henley (1997) and Abdou and Pointon (2011) argue that the effectiveness of categorization systems is contingent upon certain factors. These factors could serve as variables inside a dataset or as the target for categorization. Abdou and Pointon (2011) analyzed 214 publications and books that focused on using credit scoring systems in different business domains. There is no single universally superior categorization strategy for developing credit score models. While Abdou and Pointon (2011) argued that there may not be a single ideal credit scoring technique, a substantial quantity of literature compares several classification algorithms. The majority of this research, carried out by Yeh& Lien (2009), Tsai et al. (2009), and Akkoc (2012), presents novel classification methodologies and compares these classifiers with a set of models, including logistic regression.

Lessmann et al. (2015) contended that comparing a new classification approach, often refined without previous preconceptions, and demonstrating a higher efficiency over Logistic regression may not indicate methodological progress. Carlos Serrano-Cinca et al. (2015) intend to uncover the primary factors contributing to delinquency by examining a dataset that includes borrower profiles and loan performance data. The research uses various statistical techniques to identify the correlation between the borrower's characteristics and the default probability. The result of the study will help us gain knowledge about the importance of creditworthiness and give us the right direction about the risk management features in the P2P lending platforms. Researchers widely recognize logistic regression as the prevailing method for credit risk assessment (Abbod, 2015).

Efstathios Polyzos et al. (2023) examined the possible economic consequences of utilizing peer-to-peer lenders to fund financial projects. As the number of peer-to-peer lenders increased, they observed a rise in economic instability, a decrease in GDP, and an increase in unemployment. Conversely, peer-to-peer lending boosts the overall amount of loans disbursed. Still, it exhibits a bias towards consumer loans rather than business loans, which has a detrimental impact in the long term. In their research, Zhengwei Ma et al. (2021) "initially examined the fundamental progression of China's lending marketplace and the credit hazards borrowers face in the sector. Derived from the characteristics of P2P lending and prior studies, they devised twenty-nine indicators to evaluate borrowers' credit risk in P2P lending".

In their paper, Kamil˙e Taujanskaite and Eugenijus Milˇcius (2022) "examine the development and factors contributing to the fast expansion of Lithuania's lending marketplace and its effects on the retail credit market, emphasizing sustainability concerns. An analysis is conducted on the statutory differences separating these online lending and traditional banking segments, highlighting their contribution to developing these sectors.

In their study, Thiti Promsungwong and Tanpat Kraiwanit(2021) "explored the comprehension of peer-to-peer lending among Thailanders and the various factors that impact their decision-making. Their analysis revealed that factors such as the usage of these platforms, the transaction frequency, and the borrowers' age play a significant role in comprehending these lending platforms".

After thoroughly analyzing the literature, it can be inferred that there is no universally superior strategy for constructing risk-evaluation models. Lee et al., 2002 and West, 2000 stated that credit rating techniques assess new consumer loans; these models facilitate cost reduction in credit analysis, expedite credit decision-making, and mitigate potential risks. Kočenda and Vojtek (2011) asserted that models employing logistic regression and regression trees exhibit similar levels of efficiency.

Emekter et al. (2014) found that the absence of face-to-face interaction among investors and borrowers in online P2P lending creates a problem of unequal access to information. Therefore, it is crucial to have a proficient and precise credit risk assessment technique that will decrease the investment uncertainty devoid of human involvement to ensure the continuous growth of the P2P lending sector. Khan et al. (2023) studied the impact of P2P platform factors like simplicity of use, utility, competence, and reliability on the intent of users to utilize online P2P platforms. Over the past few decades, technology has permeated every facet of human existence. There must be facilities that can accommodate everyone's demands because people are busier, more mobilized, and more in need of assistance all the time (Thacker et al., 2019).

Yan Feng et al. (2015) "analyzed different features and concluded with the features they thought were most crucial for successful funding. They discovered that interest rates and higher loan amounts play a significant role in securing lending from investors. In contrast, the loan duration and the size of the loan did not attract the investors".

Kah Boon Lim et al. (2023) "assert that ease of use, utility, and openness in information sharing is critical to adopting these online lending platforms."TaufikFaturohman et al. (2020) "devised a plan to create credit rating models utilizing social media data on loan applicants to lower the likelihood of default among borrowers. Credit scoring becomes a laborious process because the data on these lending platforms is composed of textual information and complicated data kinds". Cuiqing Jiang et al. (2018) "used qualitative data to develop a credit model after extracting significant information from the loan application." In their finding, Mauro Aliano et al. (2023) "revealed that features like the loan amount, higher interest rates, and more extended repayment periods coupled with lower education levels and gender of the borrower were the significant causes of loan default."

As previously stated, the COVID-19 pandemic has presented numerous extra obstacles for risk managers. To minimize the adverse consequences on their operations, organizations must comprehend the repercussions of the pandemic on investments. According to Arroyo et al. (2020), implementing a redesigned underwriting process by utilizing the latest techniques in the field of analytics can enhance efficiency and effectiveness, resulting in a significant improvement in predicting accuracy. Default is the term used to describe the inability to meet financial obligations, particularly concerning a loan (Anderson, 2007).

It is crucial to quantify the likelihood of default, particularly concerning the series of defaults that impacted the Chinese market in 2018. The main reason for the slowdown of the Chinese P2P platforms is the funds' withdrawal by the lenders. This withdrawal resulted in a loss of over USD 115 billion, as Zhu et al. (2020) stated.

Research Objective

The research objective of the study is to calculate the predicted accuracy of default prediction frameworks in P2P platforms, specifically utilizing LendingClub's data, accessible through the company's website. We evaluate the models' performance using statistical metrics, such as the Kolmogorov-Smirnov (KS) test and Area Under the Curve (AUC), to create a novel quantitative model employing a machine learning technique, an artificial neural network. The building of this model will aid the P2P platforms in predicting loan defaults, thereby identifying high-risk loan applicants, decreasing the prevalence of such loans, and mitigating credit losses.

3. Research Methodology

3.1 Data Description

The rationale behind our study on financial technology consumer lending centers around LendingClub is due to two key factors. Firstly, the organization stands out as one of the few numbers of lenders that have chosen to provide their data to the public. Furthermore, it is worth noting that LendingClub is the leading fintech provider in the personal loan sector. As a result, we expect the findings presented in this context to have broader applicability. It provides comprehensive data on every loan proposal accepted or rejected since its establishment in 2007. We compile diverse information regarding the borrowers and funded loans, encompassing borrower details such as FICO scores, employment tenure, debt-to-income ratio, home ownership status, and zip code. Additionally, we gather loan details, including the interest rate, duration until maturity, date of origination, necessity for verification, and the purpose of the loan. Additionally, we track each loan's monthly payment and performance.

Data is an essential and fundamental requirement for any high-quality study. The primary focus of data collection is on various aspects of borrowers. We sourced the dataset included in this paper from the website kaggle.com, and it consists of 890,000 observations of the loan organization spanning from 2007 to 2015. Nevertheless, not all features possess significant value and importance in this research. We exclude some aspects from the analysis, such as the borrower's pin code, membership ID, and fields with values that are not present. The paper primarily focuses on the loan position, reasons for the loan and the borrowed quantity, and the borrower's creditworthiness as its primary elements. The data type of this set is a combination of qualitative and quantitative variables, encompassing discrete, continuous, ordinal, and nominal values.

3.2 Methods Employed

Software development tools: The investigation utilized Anaconda Navigator and the Python programming language. We implemented the suggested model using the Jupyter Notebook, the scikit-learn framework, and the data visualization library Seaborn in Python. The analysis included descriptive statistics and graphical tools to summarize and analyze the data. The Lending Club data analysis employs descriptive statistics to identify minimum, maximum, mean, median, and first and third-quartile values. We visualize data distribution using graphical approaches like scatter plots, bar graphs, tables, and histograms.

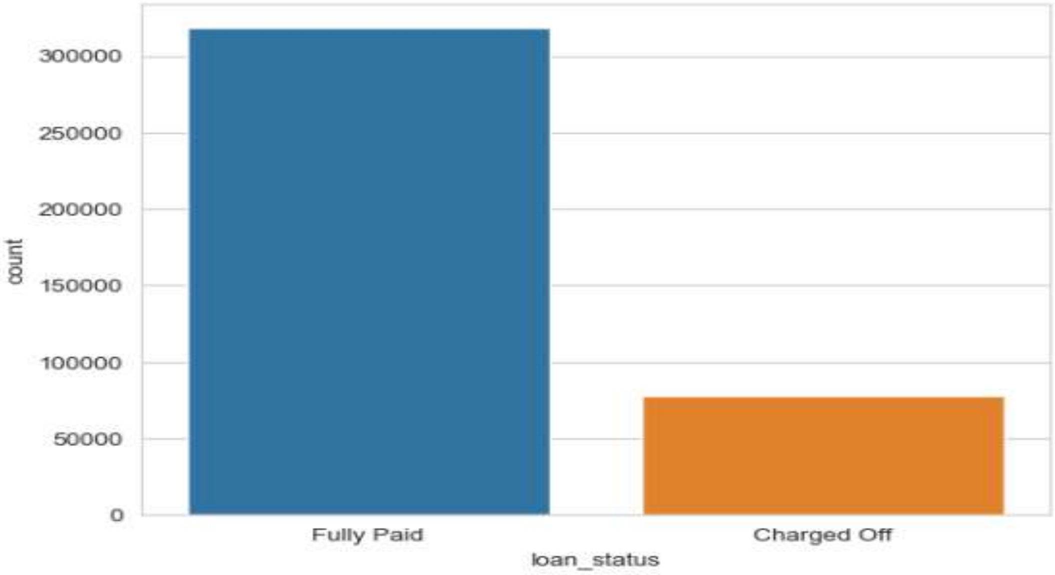
We imported two Python libraries, NumPy as np and Pandas as PD, and used the pandas read.csv method to load the raw data into Python. The pandas' method pd.read_csv instantly reads data from pandas data frames and objects, making it ideal for storing data for various manipulations and analyses.

4. Analysis and Interpretation

We analyze the data to examine summary statistics, visualize it, and understand highly significant features.

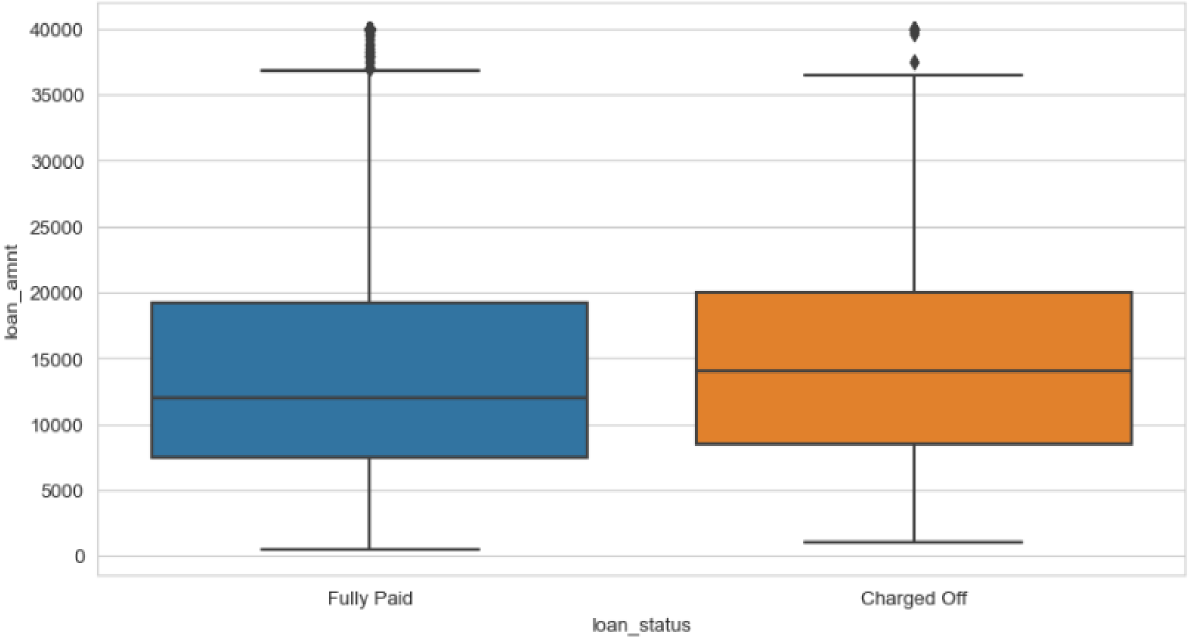
4.1 Exploratory Data Analysis

4.1.1 Fully paid vs. Charged-off Loans



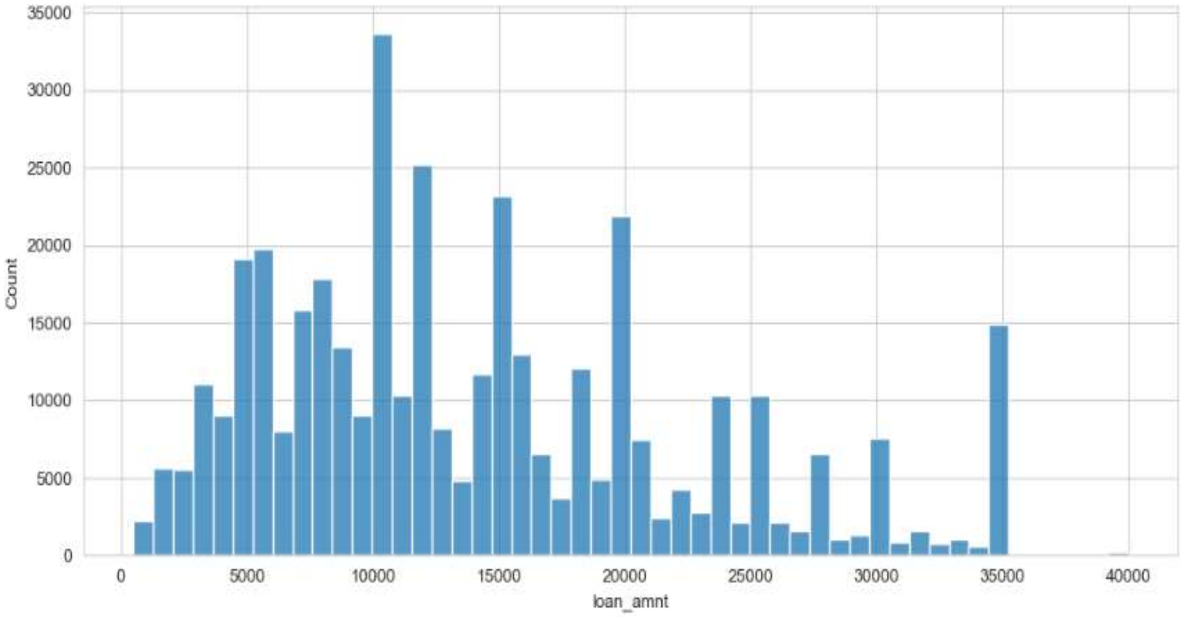
The graph conveys the loan statuses, comparing the fully paid to the charged-off loans. In simple terms, the graph shows how many borrowers are non-defaulters and how many defaulted and got their loans charged-off. The number of non-defaulted or fully paid loans exceeds those of defaulted or charged-off borrowers. Although compared to the fully paid, the number of charged-off loans is less, they still represent a substantial number; this implies that although most loans get repaid, a significant proportion still fails to do so.

4.1.2 Distribution of loan amounts between two different loan statuses



The boxplot compares the distribution of loan amounts between two different loan statuses, fully paid and charged-off. The interquartile range for both statuses is similar, indicating a consistency in the variability of the two loan amounts. The median line is identical for both categories, indicating that the loan amounts do not vastly differ, and the outliers in the charged-off status show that few big-ticket loans defaulted.

4.1.3 Frequency of Loan Amount



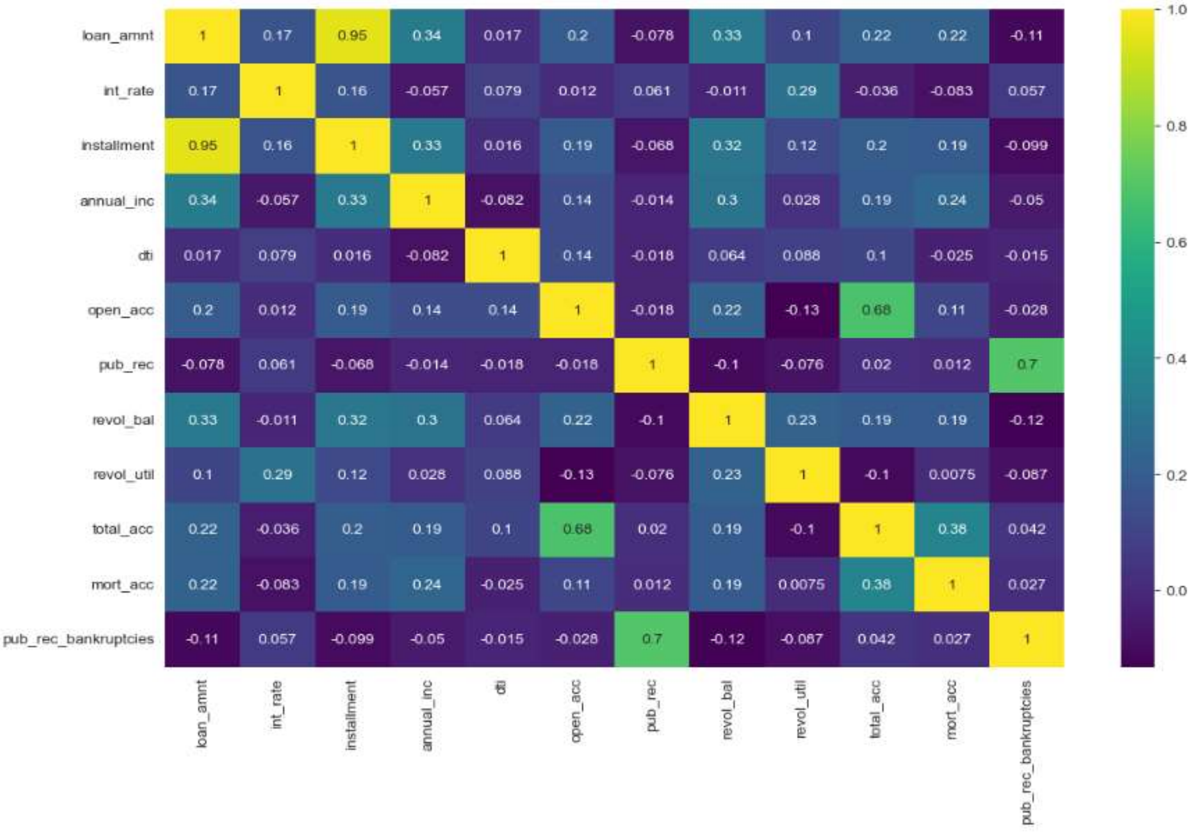
The graph is a histogram showing the distribution of the loan amount. The graph displays values ranging from zero to forty thousand, with regular spikes emphasizing popular loan amounts. The most frequent loan amounts are 5000, 10000, 15000, 20000, and 25000, which states that borrowers prefer the loans in rounded amounts. The graph's right-skewness states that most of these loans fall from zero to twenty thousand, and few exceed twenty thousand.

4.1.4 Exploring correlation

A correlation matrix is a tool used in statistical analysis for finding the linear relationship between variables. The correlation value lies between -1 to 1. If the value is closer to 1, it reflects that the two variables are positively correlated and vice-versa. When the correlation value is zero, there is no relationship between the two variables.

	loan_amnt	int_rate	instalment	annual_inc	dti	open_acc	pub_rec	revol_bal	revol_util	total_acc	mort_acc	pub_rec_bankruptcies
loan_amnt	1	0.168	0.953	0.336	0.017	0.198	-0.07	0.328	0.099	0.223	0.222	-0.106
int_rate	0.168	1	0.162	-0.056	0.079	0.011	0.060	-0.01	0.293	-0.03	-0.082	0.057
installment	0.953	0.162	1	0.330	0.016	0.188	-0.06	0.316	0.123	0.202	0.193	-0.098
annual_inc	0.336	-0.05	0.330	1	-0.08	0.136	-0.01	0.299	0.027	0.193	0.236	-0.050
dti	0.016	0.079	0.015	-0.081	1	0.136	-0.01	0.063	0.088	0.102	-0.02	-0.014
open_acc	0.198	0.011	0.188	0.136	0.136	1	-0.01	0.221	-0.131	0.680	0.109	-0.027
pub_rec	-0.077	0.060	-0.062	-0.013	-0.02	-0.01	1	-0.10	-0.075	0.019	0.011	0.699
revol_bal	0.328	-0.01	0.316	0.299	0.064	0.221	-0.10	1	0.226	0.191	0.194	-0.124
revol_util	0.099	0.293	0.123	0.027	0.088	-0.13	-0.07	0.226	1	-0.10	0.007	-0.086
total_acc	0.223	-0.03	0.202	0.193	0.102	0.680	0.012	0.191	-0.1043	1	0.381	0.042
mort_acc	0.222	-0.08	0.193	0.236	-0.03	0.109	0.011	0.194	0.0075	0.381	1	0.027
pub_rec_bankruptcies	-0.10	0.057	-0.09	-0.050	-0.01	-0.02	0.699	-0.12	-0.0868	0.042	0.027	1

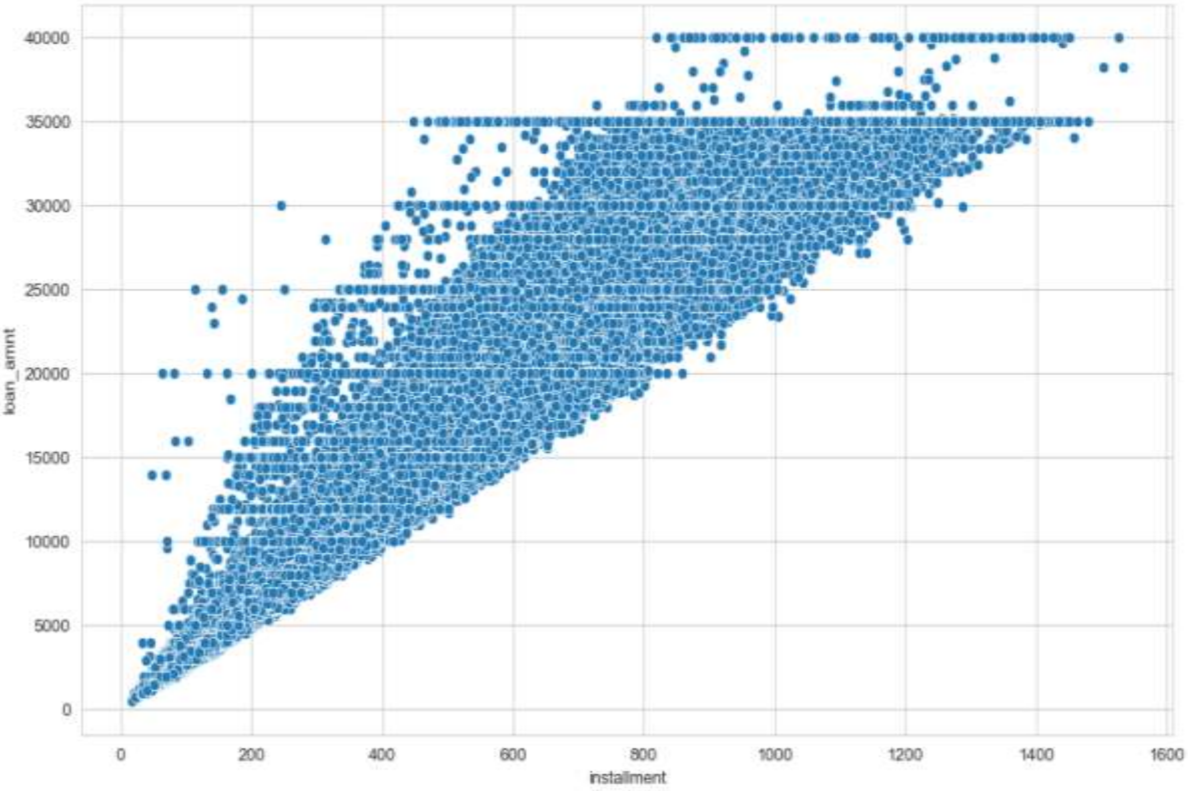
We primarily apply correlation between the continuous variables in the form of a heat map.



The graph represents a correlation between various financial variables related to loans. We observed a favorable correlation between the loan amount, which signifies the sum requested by the borrower, and the installment, indicating the higher the loan amount, the higher the monthly payments to repay the loan. There's a positive relationship between these two continuous variables. At the same time, the graph shows a moderate positive relationship between installment and annual income, suggesting that higher yearly incomes are related to higher installments, and a weak relationship exists between the loan amount and the interest rate, indicating that the loan amount does not influence the interest rates.

4.1.5 Relationship between Loan Amount and Installment Amount

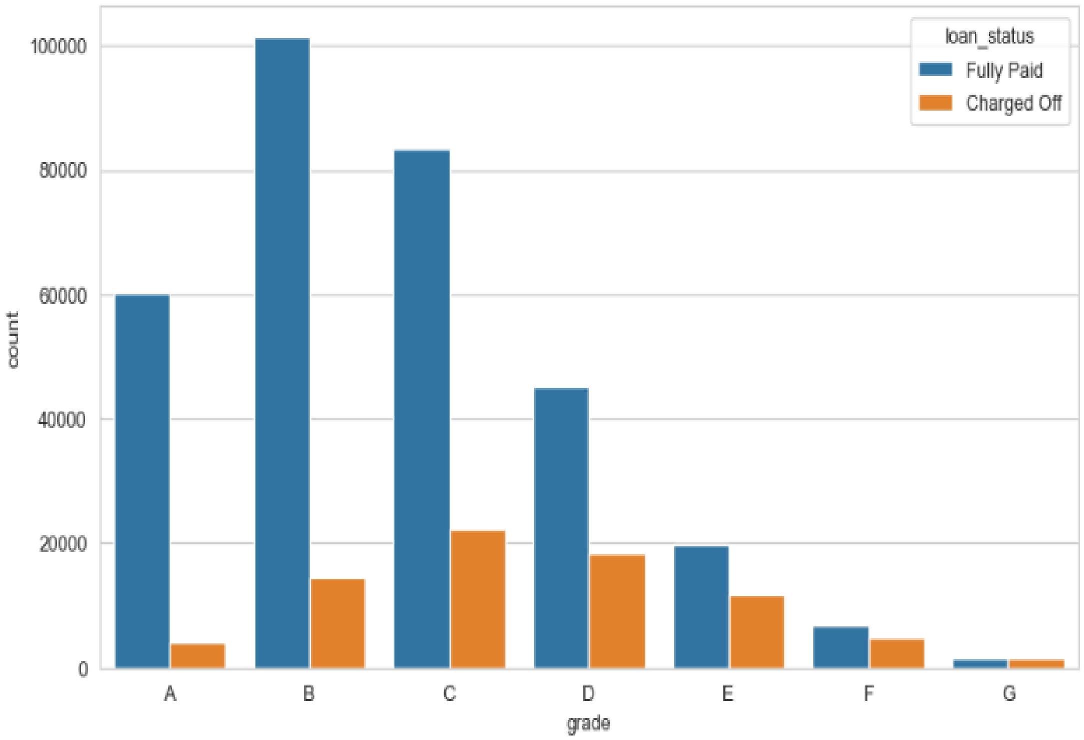
To clarify the above statement about the loan amount and the installment, we plot a scatter plot, which clears our interpretation. A scatter plot is plotted between two variables; each point on the scatter plot represents an observation determined by the values of the two variables.



The above scatter plot supports our intuition that the loan amount and installment paid depend on each other. We drop the installment variable from the dataset. The company may use an established algorithm to determine the payments following the loan obtained by the borrower.

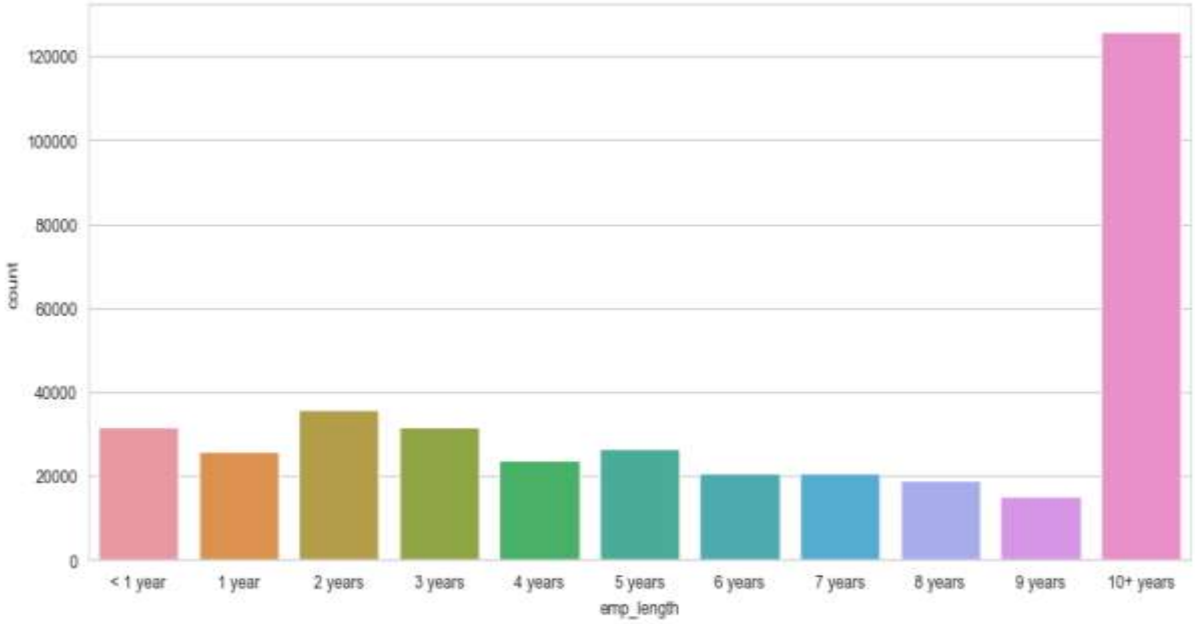
4.1.6 Loan Grade

We'll now investigate the ratings assigned to the loans based on grade and subgrade.



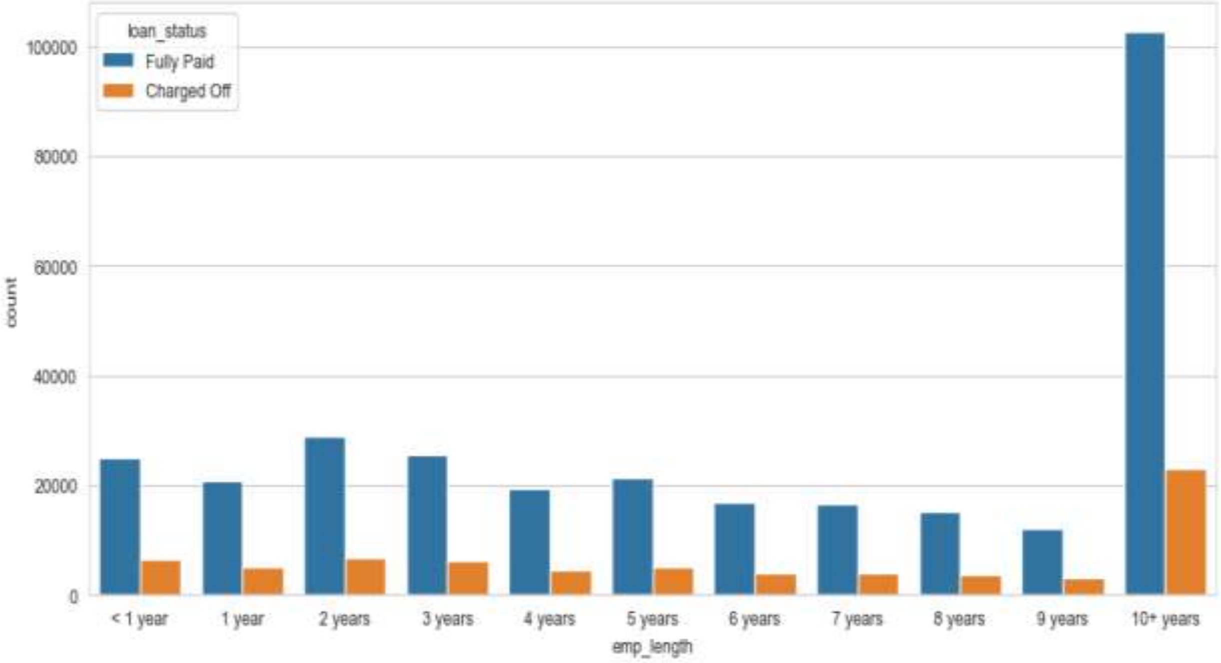
We produced a frequency plot to analyze the distribution of loan status across various loan grades to identify potential differences. Most loans fall into the A, B, C, and grades, and fewer grades fall into the F and G categories. We observe a clear trend where the likelihood of default rises as the loan grade decreases. Notably, for loan grades F and G, the default probability is notably higher than grades A and B.

4.1.7 Employment Length



The above graph illustrates the frequency of loan applicants by their employment length. The x-axis represents the loan applicants, describing their experience from one year to ten years and above. We observe that applicants with ten and above years of experience have applied for loans in more numbers than the other categories.

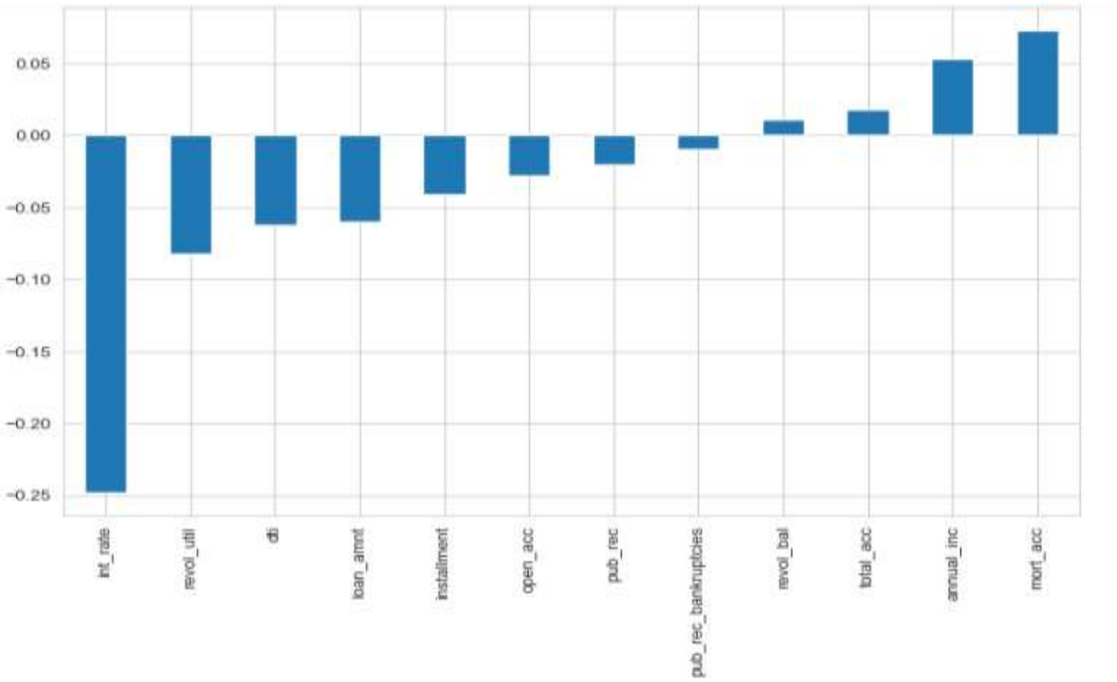
4.1.8 Employment Length vs. Loan Status



The above graphs depict the loan status distribution by the applicants' employment length. The loan status is divided into two categories: fully paid and charged-off and the frequency on the y-axis. We observe that the fully paid loans for all the employment length categories are consistent and lower than ten and above years; even the number of charged-off loans for all the categories is consistent and lower than ten and above years. A positive correlation exists between the employment length of ten and above years and the likelihood of repaying the loans. Overall, applicants with longer years of employment tend to repay the loans due to their higher income and maturity levels.

4.1.9 Interest Rates

We'll now examine the association between interest rates and loan repayment status.



The graph facilitates the identification of the variables that exhibit the highest correlation with the probability of loan repayment. The variables on the left side of the graph negatively correlate with the variable loan repaid, suggesting that the likelihood of repaying a loan decreases as these variable values increase. The most visible variables are interest rate, debt-to-income ratio, and revolving line utilization. Similarly, the variables on the right side of the graph are positively correlated, indicating that borrowers with higher income and more mortgage accounts are more likely to repay loans. The interest rate demonstrates the most significant negative correlation with loan repayment, as higher interest rates may result in substantial financial stress on the borrowers.

4.2 Data Pre-processing Techniques

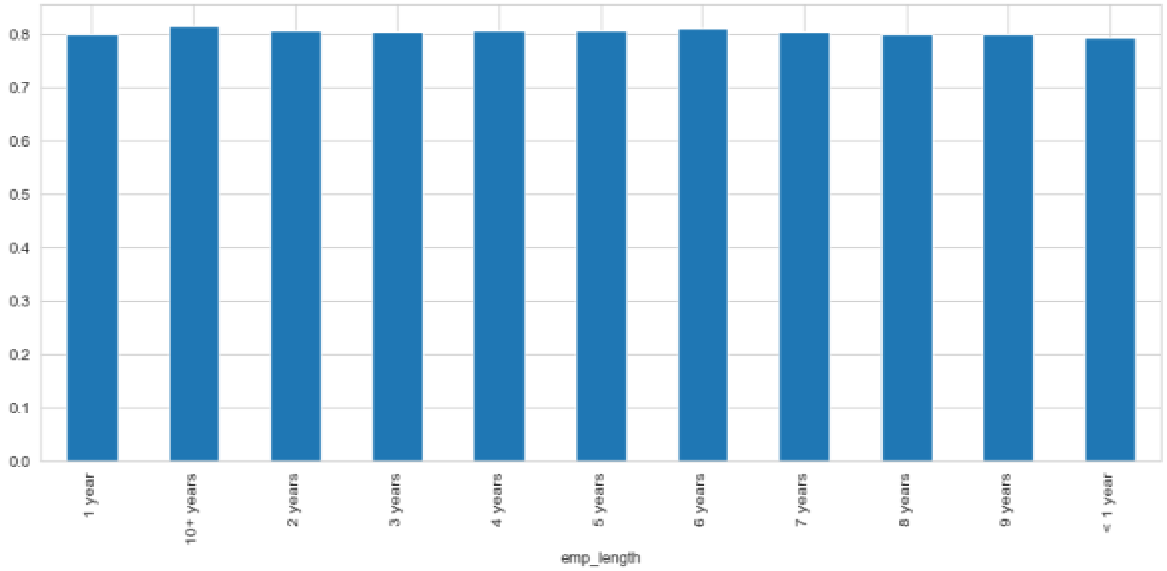
After performing the preliminary visualizations, we proceeded to pre-process our data. To prepare the data set for predictive modeling, we eliminate or plug in any missing information, reduce irrelevant and duplicate attributes, and transform them. Different data pre-processing was employed to prepare the data set. Data preprocessing methods include cleansing, merging, standardizing, and normalizing data. Additionally, the dimensionality of the dataset was reduced to seventy-six dummy variables utilizing dimensionality reduction techniques, specifically focusing on the missing value ratio.

4.2.1 Handling Missing Values

The `isnull()` method determines each variable's total count of null values. A few columns, like employee title, length of employment, other mortgages, title, and revolving utilization have missing values. Furthermore, we determine each variable's total and the average of missing values.

Attribute	**Missing values**	**Percentage**
Employee title	22925	0.057892
Employment length	18300	0.046211
Mort_acc	37795	0.095435
pub_rec_bankruptcies	535	0.001351
title	1756	0.004434
revolving utilization	276	0.000697

It appears that the "Mort_acc" variable contains the most number of null values, closely followed by the "emp_title" and "emp_length" variables. The other variables have fewer missing values, and we can delete the rows for these variables. Data imputation techniques may be necessary to address these missing values before continuing with additional analysis or modeling. Now, we'll handle the missing values present in the "Mort_acc," "emp_title," and "emp_length" variables. In the "emp_title" variable, we've around 173,105 unique employment titles, and converting these many unique values into a dummy variable is difficult, so we decided to drop this variable. In the "emp_length" variable, we'll find the association between borrowers' employment length and the status of their loan repayment, i.e., whether paid off or charged off. We plot the proportion of defaulted loans across different employment length categories.



We do not find any variation across different employment length categories. The charge-off rates across the categories are almost the same, and we found no incentive to keep this variable, so we decided to drop it. The title and purpose variables provided similar information when observed, so we dropped the title variable.

The mort_acc variable gives us information about how many mortgage accounts each borrower has. The mort_acc variable has close to ten percent missing values, the highest among all the variables. Initially, we thought of dropping the variable, which is vital for finding the loan status. Second, ten percent of missing values in the variable can be imputed with the rest of the data. We assume that the total_acc variable is closely related to mort_acc. We averaged the borrower values across all categories using all the accounts and interpolated that number into the mort_acc variable.

4.2.2 Handling Categorical Data

In data pre-processing, we handled the correlation between variables and missing values by either dropping the variables or imputing the values. Now, we'll handle the categorical variables and transform them into dummy variables. We began exploring the 'term' variable and found that it indicates the number of loan payments, presented as string values such as '36 months' or '60 months'. Because this data primarily consists of numerical values, we converted it accordingly: '36 months' was changed to 36, and '60 months' was encoded as 60. We wrote a user-defined function and applied it to the 'term' variable using the apply method to achieve this.

The following step was to take care of the 'grade' and 'sub_grade' variables. Our previous graphical analysis showed that the variable 'grade' redundantly contains data already captured in the variable 'sub_grade.' Hence, we discarded the 'grade' variable and retained solely the 'sub_grade' variable. We generated dummy variables for the 'sub_grade' variable and subsequently merged these dummy variable columns with the original data frame and eliminated the original categorical data for the sub_grade.

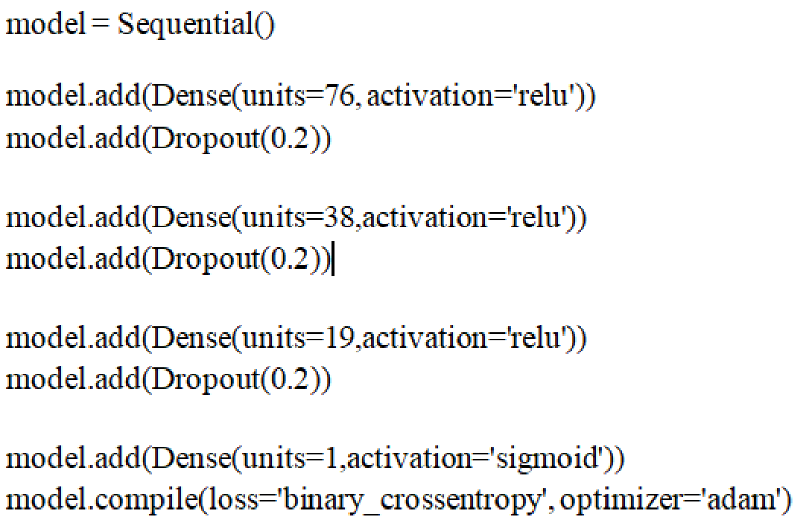
At this point, we evaluated the rest of the categorical variables. We observed that 'application_type,' 'verification_status,' 'purpose,' and 'initial_list_status' were well-suited for creating dummy variables, given their limited range of unique values, including some binary categories. As a result, we generated dummy variables based on these values, integrated them into our dataset, and eliminated the original categorical columns.

We found that the 'home_ownership' variable had a few unique values, making it suitable for conversion into a dummy variable. However, we had to combine a few categories before transforming them. Three main categories stood out among the six categories: 'MORTGAGE,' 'RENT,' and 'OWN.' The remaining three categories, namely 'OTHER,' 'NONE,' and 'ANY,' were deemed suitable for consolidation. We applied the replace method to merge the remaining categories. As a result, we generated dummy variables based on these values, integrated them into our dataset, and eliminated the original categorical column.

The variable 'earliest_cr_line' denotes when the borrower first opened their reported credit line. The information in the earliest_cr_line variable is a combination of values that includes the month and year. We thought the most effective thing to do was to separate the year, transform it into a numeric value, and then eliminate the 'earliest_cr_line' from our dataset. As the loan_status variable is precisely similar to the loan_repaid variable, we can drop this variable from the model.

4.3 Model Building and Evaluation

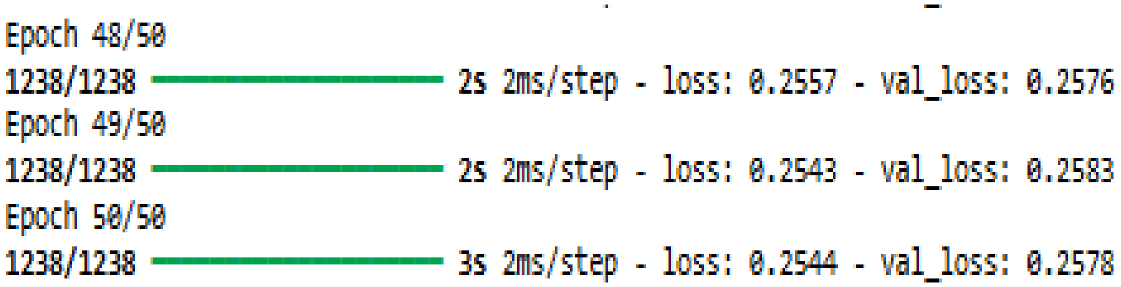
Following the completion of exploratory data analysis and implementation of data pre-processing techniques, we proceed to construct the model. The initial model development phase involves dividing the dataset into training and testing subsets. To split the data, we use the train_test_split method. For model building, we use train data (eighty percent) and test data (twenty percent) to test the model's accuracy. As we predominantly use numeric data, we need to normalize it; for this purpose, we apply the MinMaxScalar method.



```
model = Sequential()

model.add(Dense(units=76, activation='relu'))
model.add(Dropout(0.2))

model.add(Dense(units=38,activation='relu'))
model.add(Dropout(0.2))

model.add(Dense(units=19,activation='relu'))
model.add(Dropout(0.2))

model.add(Dense(units=1,activation='sigmoid'))
model.compile(loss='binary_crossentropy', optimizer='adam')
```

Fig 1: Neural Network Model for Binary Classification using Keras

After standardizing the dataset, we employed the Keras library with TensorFlow.The initial layer comprises seventy-six neurons to match the complexity and features of the dataset; the activation function of ReLU was employed to induce non-linearity and intricate patterns present within the dataset. A dropout layer was added to avoid overfitting. Subsequently, we established two hidden layers, each halving the neuron count compared to the previous layer; this will help the model learn the high-level features. Due to binary classification, a sigmoid function is applied to confine output values in the range of zero and one for the output layer, consisting of a single neuron. Finally, as the problem is a binary classification, the model is configured with binary cross-entropy as the loss function and 'Adam' as the optimizer. We now train the model that we developed by inputting the independent and dependent train variables. As the amount of train data is vast and may overfit, we cut the data into a batch size 256 and pass these batches into the model.



```
Epoch 48/50
1238/1238 ━━━━━━━━━━━━━━━━━━━━ 2s 2ms/step - loss: 0.2557 - val_loss: 0.2576
Epoch 49/50
1238/1238 ━━━━━━━━━━━━━━━━━━━━ 2s 2ms/step - loss: 0.2543 - val_loss: 0.2583
Epoch 50/50
1238/1238 ━━━━━━━━━━━━━━━━━━━━ 3s 2ms/step - loss: 0.2544 - val_loss: 0.2578
```

Fig 2: Training Log for the Neural Network Model

The output states that the training process has completed 50 epochs with 1238 batches of data processing in each epoch. The training loss is 0.2544, and the validation loss is 0.2578, with approximately three seconds per epoch. Training loss indicates how well the model learns the training data, and validation loss indicates how well it generalizes the test data. We then plot a graph using training and validation loss data to know how well our model can learn and predict the test data. The ideal scenario is the training and validation losses decreasing over time with each passing epoch. Compared to the training loss, the lower validation loss is always preferred. If validation loss exceeds training loss, our model cannot learn and give the required accuracy, thus ending as an overfitting model.

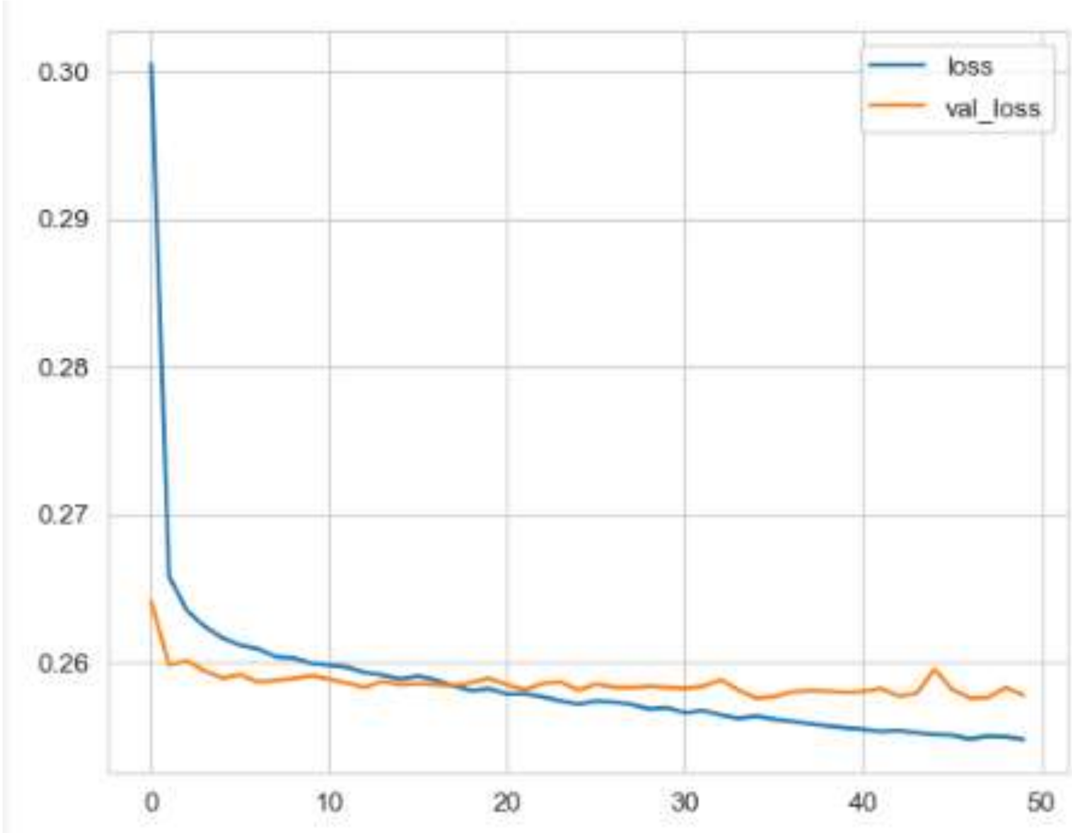


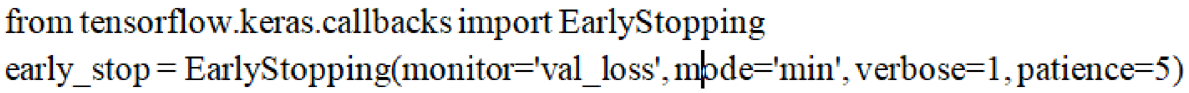
Fig 3: Training and Validation Loss Over Epochs

The graph above illustrates that the validation loss surpasses the training loss, indicating the need to fine-tune the parameters to address this issue. The classification table presents the precision, recall, and f1-score for two classes: class 0, which represents applicants who will not default, and class 1, which represents applicants who default. Here, accuracy is the proportion of correct predictions made by our model. We've achieved an accuracy of around eighty-nine percent.

	0	1	Accuracy	Macro Average	Weighted Average
Precision	0.94	0.88	0.89	0.91	0.89
Recall	0.46	0.99	0.89	0.73	0.89
f 1-score	0.62	0.94	0.89	0.78	0.87

Table 1: Performance Metrics for the Neural Network Model

To address the concerns of validation loss being more than the training loss and to increase the model accuracy, we incorporated the 'EarlyStopping' callback function into our training process.



```
from tensorflow.keras.callbacks import EarlyStopping
early_stop = EarlyStopping(monitor='val_loss', mode='min', verbose=1, patience=5)
```

Fig 4: Implementing Early Stopping Callback for Neural Network Training

This function monitors the training and validation loss with a patience value of five, meaning that if the validation loss fails to decrease for five consecutive epochs, the model's training halts automatically. This proactive measure helped to prevent overfitting and ensured a slight increase in the model's generalization ability.

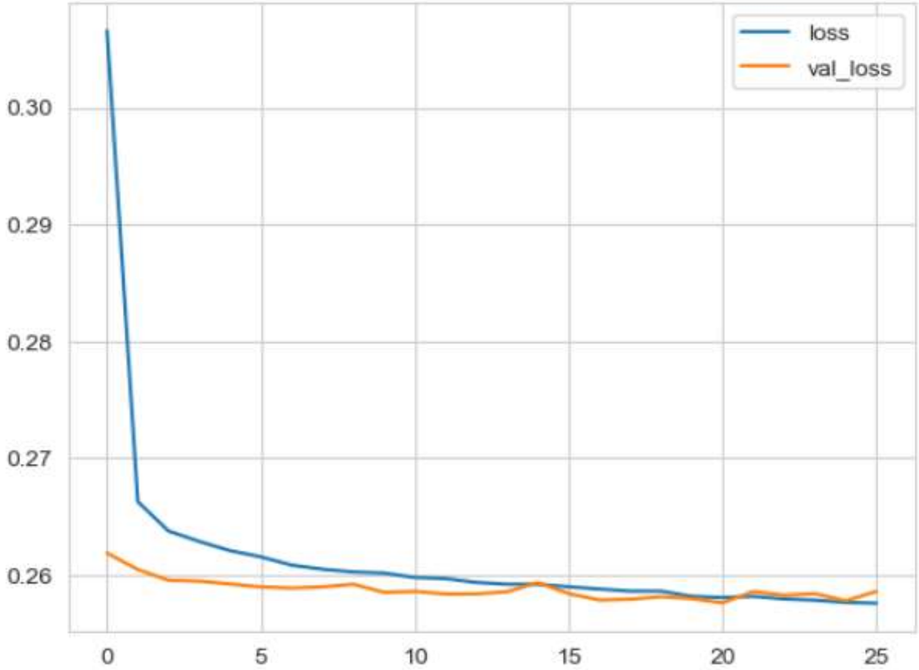


Fig 5: Training and Validation Loss Over Epochs After Implementing Early Stopping Callback

The graph above shows that the gap between the validation and training loss has decreased, indicating a considerable increase in the model's learning. Similarly, the classification table also shows an increase in the accuracy levels.

	0	1	Accuracy	Macro Average	Weighted Average
Precision	0.99	0.88	0.89	0.93	0.90
Recall	0.44	1.00	0.89	0.72	0.89
f 1-score	0.61	0.94	0.89	0.77	0.87

Table 1: Performance Metrics for the Neural Network Model After Implementing Early Stopping Callback

Unfortunately, this dataset is skewed, with many more fully paid loans than charged-off ones. Here, the f1-score, recall, and precision are the metrics to watch. We've achieved an accuracy of around ninety percent. Our predictive model achieves a forty-four percent accuracy rate in determining when a debt will be charged off.

5. Limitations

Adjusting a few of our model's parameters would yield better results. Experimenting with the number of neurons and hidden layers could be one option. We could also consider additional feature engineering. Second, we did not utilize the employment title column. If we could classify job titles according to an anticipated average wage, it might be feasible to introduce dummy variables, but it would be labor-expensive. However, given the time and effort involved, it can be taken up in the coming studies.

Conclusion

Efficient credit scoring models have essential consequences for Lending Club's company's operations, such as improving loan approval processes, lowering default risk, and increasing investor confidence. By accurately determining the borrower's creditworthiness, the Lending Club may draw new lenders and investors while maintaining a solid loan portfolio.

Regular evaluation and improvement are needed to adapt to changing marketplace dynamics, regulatory changes, and borrower behavior.

We found vital variables significantly impacting credit scoring outcomes using exploratory data analysis (EDA) and feature selection methods. The variables in question comprise borrower characteristics such as employment duration, earnings, the ratio of debt to income, inst_rate, total_acc, and mort_acc.

By examining financial indicators such as the rate of default for various borrower categories and mortgage scores, we gained insight into borrowers' credit risk. It allowed us to gain a more nuanced understanding of the factors that contribute to loan defaults and evaluate borrowers' overall creditworthiness.

We used a deep learning algorithm to create predictive models for credit scoring. Mathematical algorithms have been crafted based on previous data to anticipate the probability of default in fresh loan submissions, allowing for aggressive risk-control strategies. A comprehensive assessment of the model's outcome utilizing measures that include precision, recall, precision, and training and validation plot revealed information about the efficacy of the models. Cross-validation techniques helped ensure the models were robust and applicable across different datasets.

To summarize, calculating a credit score on Lending Club entails using data-driven approaches to evaluate borrower creditworthiness, reduce the likelihood of default, and improve lending decisions. Using modern machine learning and statistical analysis techniques, Lending Club can increase its competitiveness, promote financial empowerment, and drive sustainable growth within the financing market.

References

1. Abdou, H. A., &Pointon, J. (2011, April 1). *CREDIT SCORING, STATISTICAL TECHNIQUES AND EVALUATION CRITERIA: A REVIEW OF THE LITERATURE*. Intelligent Systems in Accounting, Finance, and Management.
 https://doi.org/10.1002/isaf.325
2. Emekter, R., Tu, Y., Jirasakuldech, B., & Lu, M. (2015). Evaluating credit risk and loan performance in online Peer-to-Peer (P2P) lending. *Applied Economics*, *47*(1), 54–70.
https://doi.org/10.1080/00036846.2014.962222
3. Mammadova, Leyla (2021). Peer-to-peer (P2P) lending: default, default dependency, and industry potential. Loughborough University. Thesis.
https://doi.org/10.26174/thesis.lboro.14544420.v1
4. Lahsasna, Ainon, R., &Wah, T. (2008, September 25). *Credit Scoring Models Using Soft Computing Methods: A Survey* . https://www.ccis2k.org/.
5. Wu, T. P., Wu, H. C., Chen, B., Lin, Q., & Zou, T. (2019, October 1). *Does P2P Lending Affect Bank Lending? Evidence from China. | Journal of Applied Economics &amp; Business Research | EBSCOhost*.
https://openurl.ebsco.com/EPDB%3Agcd%3A1%3A10212561/detailv2?sid=ebsco%3Aplink%3Ascholar&id=ebsco%3Agcd%3A140997784&crl=c
6. Lessmann, S., Baesens, B., Seow, H., & Thomas, L. C. (2015, November 1). *Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research*. European Journal of Operational Research. https://doi.org/10.1016/j.ejor.2015.05.030
7. Serrano-Cinca C, Gutiérrez-Nieto B, López-Palacios L (2015) Determinants of Default in P2P Lending. PLoS ONE 10(10): e0139427.
https://doi.org/10.1371/journal.pone.0139427
8. Jiang, C., Wang, Z., Wang, R. et al. Loan default prediction by combining soft information extracted from descriptive text in online peer-to-peer lending. Ann Oper Res 266, 511–529 (2018). https://doi.org/10.1007/s10479-017-2668-z
9. Aliano M., Alnabulsi K., Cestari G., Ragni S. (2023). Peer-to-Peer (P2P) Lending in Europe: Evaluating the Default Risk of Borrowers in the Context of Gender and Education. European Scientific Journal, ESJ, 19 (7), 60.
https://doi.org/10.19044/esj.2023.v19n7p60
10. Polyzos, S., Samitas, A., &Rubbaniy, G. (2023, May 29). *The perfect bail-in: Financing without banks using peer-to-peer lending*. International Journal of Finance & Economics/ International Journal of Finance and Economics. https://doi.org/10.1002/ijfe.2838
11. Citation: Ma Z, Hou W, Zhang D (2021) A credit risk assessment model of borrowers in P2P lending based on BP neural network. PLoS ONE 16(8): e0255216. https://doi.org/10.1371/journal.pone.0255216
12. Taujanskaitė, K., &Milčius, E. (2022). *Accelerated Growth of Peer-to-Peer Lending and Its Impact on the Consumer Credit Market: Evidence from Lithuania*.
https://econpapers.repec.org/article/gamjecomi/v_3a10_3ay_3a2022_3ai_3a9_3ap_3a210-_3ad_3a904658.htm
13. Akkoç, S. (2012). An empirical comparison of conventional techniques, neural networks and the three-stage hybrid Adaptive Neuro-Fuzzy Inference System (ANFIS) model for credit scoring analysis: The case of Turkish credit card data. European Journal of Operational Research, 222(1), 168–178. https://doi.org/10.1016/j.ejor.2012.04.009
14. Tsai, C. H. (2018). To regulate or not to regulate: A comparison of government responses to peer-to-peer lending among the United States, China, and Taiwan. U. Cin. L. Rev., 87, 1077.
15. Yeh, I.C. and Lien, C.H. (2009) The Comparisons of Data Mining Techniques for the Predictive Accuracy of Probability of Default of Credit Card Clients. Expert Systems with Applications, 36, 2473-2480. https://doi.org/10.1016/j.eswa.2007.12.020

16. Feng, Y., Fan, X., & Yoon, Y. (2015). LENDERS AND BORROWERS'STRATEGIES IN ONLINE PEER-TO-PEER LENDING MARKET: AN EMPIRICAL ANALYSIS OF PPDAI. COM. Journal of Electronic Commerce Research, 16(3), 242.
17. Ala'raj, M., &Abbod, M.F. (2015). A systematic credit scoring model based on heterogeneous classifier ensembles. *2015 International Symposium on Innovations in Intelligent SysTems and Applications (INISTA)*, 1-7.
18. Boon, K., Lim, Lo, C.Z., Yeo, S.F., Ling, C., & Tan Understanding of Peer-to-Peer Lending Platform Intention: Evidence among Millennials.
19. Taufik Faturohman; Muhammad Abdullah Hamzah Syaiful Mukminin; Sudarso Kaderi Wiryono; Gun Indrayana; Raden Aswin Rahadi; Kurnia Fajar Afgani, International Journal of Monetary Economics and Finance (IJMEF), Vol. 16, No. 3/4, 2023
20. Promsungwong, &Kraiwanit. (2021). *Factors affecting knowledge and understanding of P2P lending In Thailand*. https://rsuir-library.rsu.ac.th/handle/123456789/1554.
21. Khan, M. T. I., Yee, G. H., &Gan, G. G. G. (2023). Antecedents of Intention to Use Online Peer-to-Peer Platforms in Malaysia. Vision, 27(5), 680-694.
https://doi.org/10.1177/09722629211039051
22. Bofondi and Gobbi (2018, January 8). *The Big Promise of Fintech - European Economy*. European Economy. https://european-economy.eu/2017-2/the-big-promise-of-fintech/
23. Bachmann, Becker, Buerckner, Hilker, & Funk. (2011, August). *Online Peer-to-Peer Lending – A Literature Review*. https://www.arraydev.com/commerce/jibc/. Retrieved May 18, 2024, from https://www.researchgate.net/profile/Burkhardt-Funk/publication/ 236735575_Online_Peer-to-Peer_Lending--A_Literature/links/54d9fb820cf24647581ff432/ Online-Peer-to-Peer-Lending--A-Literature.pdf
24. Baldassarre, B., Calabretta, G., Bocken, N., & Jaskiewicz, T. (2017, March 1). *Bridging sustainable business model innovation and user-driven innovation: A process for sustainable value proposition design*. Journal of Cleaner Production. https://doi.org/10.1016/j.jclepro.2017.01.081
25. *The State of Small Business Lending: Credit Access During the Recovery and How Technology May Change the Game - Working Paper - Faculty & Research - Harvard Business School*. (n.d.). https://www.hbs.edu/faculty/Pages/item.aspx?num=47695
26. *Wang, H., & Greiner, M. E. (2011). Prosper—The eBay for Money in Lending 2.0. Communications of the Association for Information Systems, 29, pp-pp. https://doi.org/10.17705/1CAIS.02913*
27. M. J. Ariza-Garzón, J. Arroyo, A. Caparrini and M. -J. Segovia-Vargas, "Explainability of a Machine Learning Granting Scoring Model in Peer-to-Peer Lending," in IEEE Access, vol. 8, pp. 64873-64890, 2020, Doi: 10.1109/ ACCESS.2020.2984412.