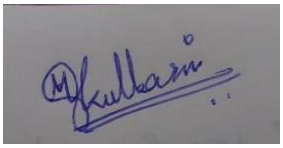
Final Project report on Loan Risk

greatlearning

Batch details	PGPDSE-FT PUNE AUG23
Team members	Sushant Londhe, Mugdha Kulkarni, Mayur Surawat, Amey Shendage, Piyush Kukdiya, Kunal Kadam
Domain of Project	Predictive Analytics
Proposed project title	Loan Risk Analysis
Group Number	5
Team Leader	Mugdha Kulkarni
Mentor Name	Aishwarya Sarda

Date: 06-03-2024

Signature of the Mentor



Signature of the Team Leader

Table of Contents

1) Industry Review

Loans have been a significant aspect of people's lives for quite some time. Everyone has distinct reasons for seeking a loan, whether it is to establish a business, or acquire various products. Even affluent individuals opt for loans over spending their cash to leverage tax benefits and maintain liquidity for unforeseen and unconventional expenses in the future.

Loans hold significance for lenders just as much as they do for borrowers. Nearly all banking institutions derive a major portion of their income from the interest earned on loans. Nevertheless, a crucial point to note is that lenders only reap profits when the loan is successfully repaid. Lending organizations encounter the challenging responsibility of evaluating the risks linked with each client. Hence, it is vital to recognize the potentially hazardous actions of clients and make well-informed decisions.

1.1. Current Practices in the Financial Industry

- Traditional lending institutions typically employ manual underwriting processes based on credit scores, income verification, and debt-to-income ratios. They often use predetermined rules and models to decide on loan approvals.
- However, these methods might not comprehensively capture the shades of borrower behavior and financial status, leading to potential inefficiencies or biases in decision-making.
- Background research indicates a growing reliance on technology and data analytics to enhance risk evaluation processes

1.2. Challenges in Loan Default Prediction

- Borrower behaviour may change over time, and patterns observed in historical data may not always be indicative of future defaults.
- External factors such as inflation rates, interest rates, and housing market conditions can influence loan default rates.
- Economic conditions can change rapidly, affecting borrowers' ability to repay loans. Unforeseen events, such as economic downturns or recessions, may impact borrowers and increase the likelihood of defaults, making it challenging to create robust models.

1.2. Use of Data Science and Machine Learning

- In loan risk analysis, data science is instrumental in developing credit scoring models, employing machine learning algorithms, and utilizing alternative data sources.
- It enables the identification of risk factors, behavioural analytics, fraud detection, and real-time monitoring.
- Data science also contributes to explainable AI, adaptive models, and optimizing loan portfolios, providing lenders with powerful tools for accurate predictions and proactive risk management.

1.3. Relevant Research and Literature

- Ongoing research explores the integration of artificial intelligence (AI) and big data analytics to improve the accuracy and efficiency of loan risk models.
- Ethical considerations and interpretability of AI-driven models are gaining attention as areas of continued exploration.
- Feature engineering, model interpretability, and ensemble methods are key areas of focus in the literature.
- Some research emphasizes the necessity of making machine learning models transparent and explainable, especially to meet regulatory standards.

1.4. Application

- In the banking industry, loan risk analysis applications extend across various types of loans, including personal loans, mortgages, and business loans.
- The adoption of advanced analytics has facilitated more precise predictions of default probabilities and has improved decision-making processes.

2) Dataset and Domain

2.1. Data Dictionary

- The dataset contains information on loans, with each entry uniquely identified by an "id." Key features include borrower details such as employment information and financial metrics.
- The "loan_amnt" represents the applied loan amount, while "funded_amnt" indicates the sanctioned amount. The "term" specifies the loan duration, and "int_rate" signifies the interest rate.
- Features like "annual_inc" provide borrower income, and "dti" offers the debt-to-income ratio. The dataset covers loan status viz our target variable, purpose, credit history, and payment details, making it comprehensive for analysing loan performance and borrower behaviour.

2.2. Pre-processing Data Analysis

2.2.1. Data type conversion

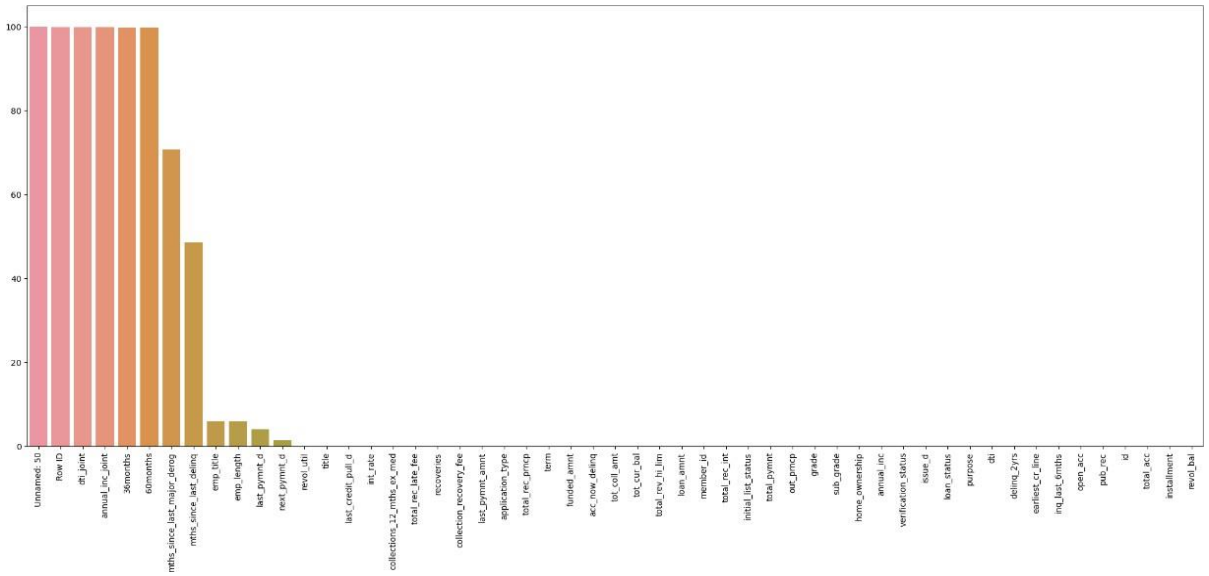
- The date columns in the dataset can be highly valuable for analysing trends, patterns, and time-dependent behaviours in the context of loans.

- There are 212999 rows and 53 columns in this file. The data type of the following 5 variables is specified as categorical and we need to convert them to numeric variables by changing datatype to datetime:

  - ISSUE_D
  - EARLIEST_CR_LINE
  - LAST_PYMNT_D
  - NEXT_PYMNT_D
  - LAST_CREDIT_PULL_D

2.2.2. Check and treat for missing values

There are 53 variables out of which 15 variables have null values in it.



a) If the percentage of missing values in a column exceeds threshold value of 90%, then the column is dropped.

In our case we've 5 columns exceeding threshold value of 90% mentioned below:

UNNAMED: 50	100.00000
ROW ID	99.827229
DTI_JOINT	99.793896
ANNUAL_INC_JOINT	99.792957
36MONTHS	99.760093
60MONTHS	99.760093

b) Variables falls below threshold values are imputed with median or mode depending on the type of variables.

MTHS_SINCE_LAST_MAJOR_DEROG	70.720520
EARLIEST_CR_LINE_MONTH	52.214799
EARLIEST_CR_LINE_YEAR	52.214799
MTHS_SINCE_LAST_DELINQ	48.531214

EMP_TITLE	6.004723
EMP_LENGTH	5.993925
LAST_PYMNT_D_YEAR	3.985465
LAST_PYMNT_D_MONTH	3.985465
NEXT_PYMNT_D_YEAR	1.301884
NEXT_PYMNT_D_MONTH	1.301884
REVOL_UTIL	0.038498
TITLE	0.007042
LAST_CREDIT_PULL_D_YEAR	0.003286
LAST_CREDIT_PULL_D_MONTH	0.003286

Variables with month are categorical so we will impute them with mode, remaining all the variables are imputed with median.

2.3. Check for duplicates and Redundant variables.

- We do not have any duplicate rows in the dataset.
- We observed that the 'Title' and 'Purpose' columns in the dataset provide redundant information. As a result, we have decided to drop the 'Purpose' column, considering that the 'Title' column already conveys the necessary information.
- Additionally, we identified that the 'Sub_grade' column is not necessary as we already have the 'Grade' column, which encompasses the grade categories. Therefore, we will remove the 'Sub_grade' column to streamline the dataset.
- 'Issue_year' and 'next_payment' year contains only 1 unique value i.e. 2015 and 2016 respectively. Hence, we will drop these features.

2.4. Alternate sources of data that can supplement the core dataset

- Supplementing a loan risk analysis dataset can be achieved by incorporating credit score data, employment market statistics, and industry-specific economic indicators.
- Including credit scores offers standardized insights into borrowers' creditworthiness, while employment market data provides context on economic conditions affecting loan repayment.
- Industry-specific indicators add nuance to risk assessment. External debt-to-income ratios verify self-reported values, enhancing model accuracy. Social media data and utility payment histories provide behavioural insights, adding depth to borrower profiles.
- Ensuring compliance with ethical and legal standards is crucial in integrating diverse data sources for a comprehensive loan risk analysis.

2.5. Problem Justification

2.5.1 Problem Statement

- Optimizing Loan Approval and Minimizing Default Risk
- The lending institution aims to enhance its loan approval process by leveraging historical loan data. The goal is to develop a predictive model that accurately assesses the risk associated with loan applicants and improves decision-making in approving or denying loan requests.

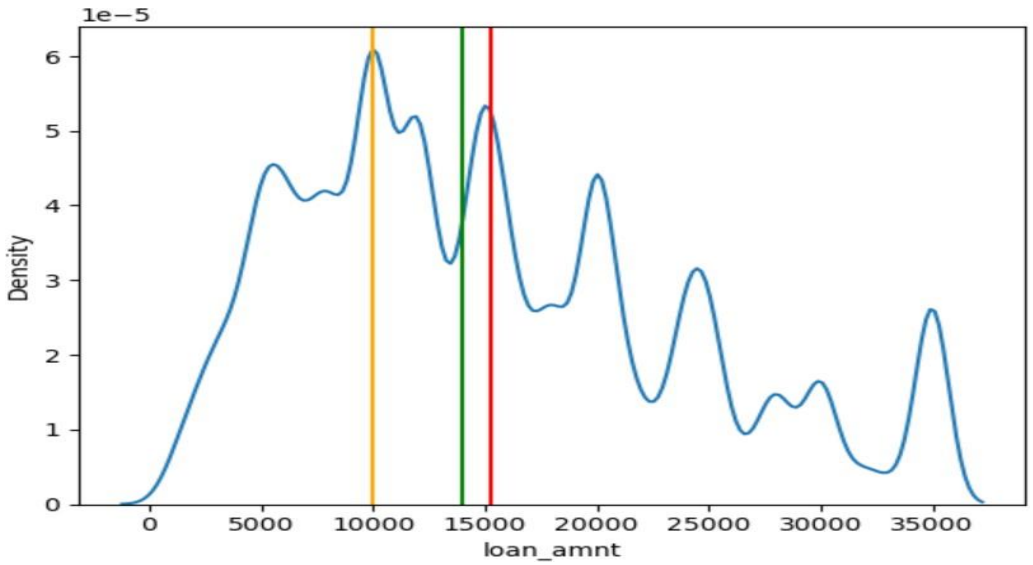
2.5.2 Project Outcome

- For businesses and financial institutions, the predictive model developed for loan default using advanced machine learning techniques can significantly enhance risk management practices. By accurately assessing the likelihood of loan defaults, these organizations can make more informed lending decisions, reducing financial risks and optimizing their loan portfolios.
- The project contributes to the academic community by showcasing the practical application of data science and machine learning in the financial sector. Researchers and students can draw insights from the methodology employed, exploring new avenues for advancing predictive modelling techniques in loan risk analysis.
- Socially, the project's impact is felt in responsible lending practices. By developing a model that accurately predicts loan default, there is potential to mitigate financial hardships for borrowers. The model can contribute to fair and transparent lending, ensuring that individuals receive loans based on their true creditworthiness.

3) Data Exploration (EDA)

3.1. Univariate Analysis

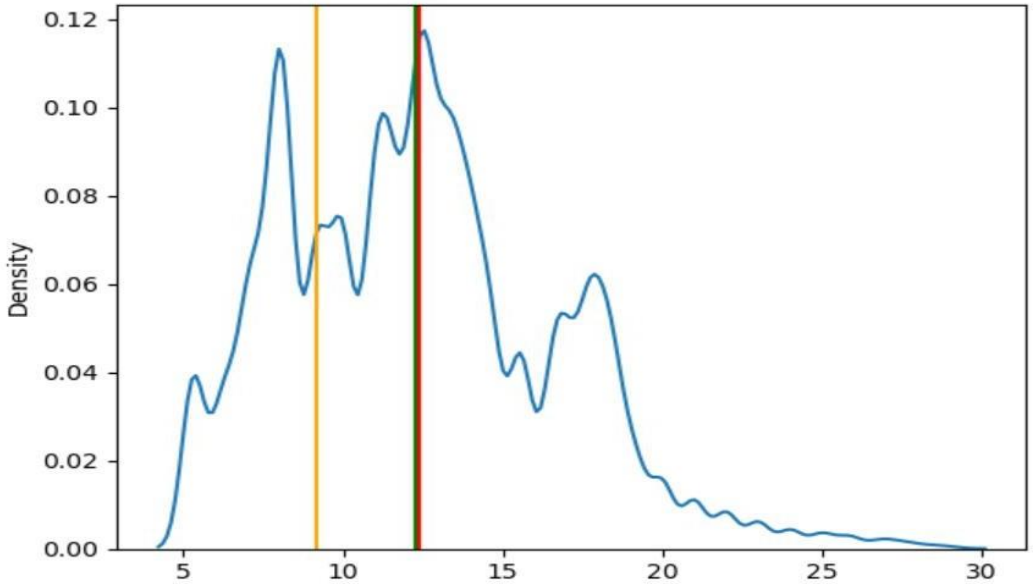
- **loan_amnt**:



Inferences:

- The mean (average) loan amount is approximately $15,257.97.
- The standard deviation is approximately $8,611.71. This measures the variability or spread of the loan amounts around the mean. A higher standard deviation indicates a wider range of loan amounts.
- 25% of the loans have an amount less than or equal to $8,500.
- Half of the loans have amounts below $14,000, and half have amounts above.
- 75% of the loans have an amount less than or equal to $20,000.

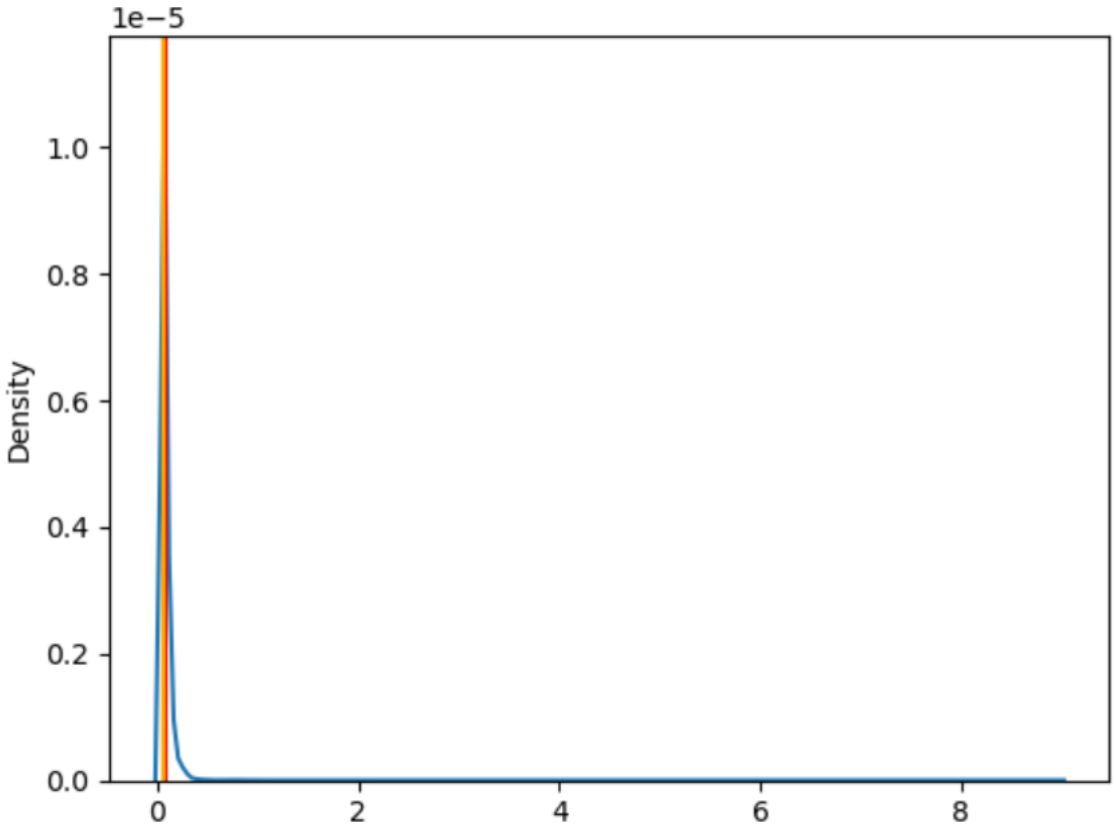
- **Int_rate**:



Inferences-

- The mean interest rate is approximately 12.40%.
- The minimum interest rate is 5.32%, while the maximum is 28.99%.
- 50 % of the interest rates are 12.29%.
- 75% of the interest rates are less than or equal to 14.65%.
- The distribution of interest rates is right-skewed, as the mean is greater than the median, and the maximum value is considerably higher than the mean.
- Most interest rates fall within the range of 9.17% to 14.65%, with the majority concentrated around the median value of 12.29%.
- There is variability in interest rates, as indicated by the standard deviation.

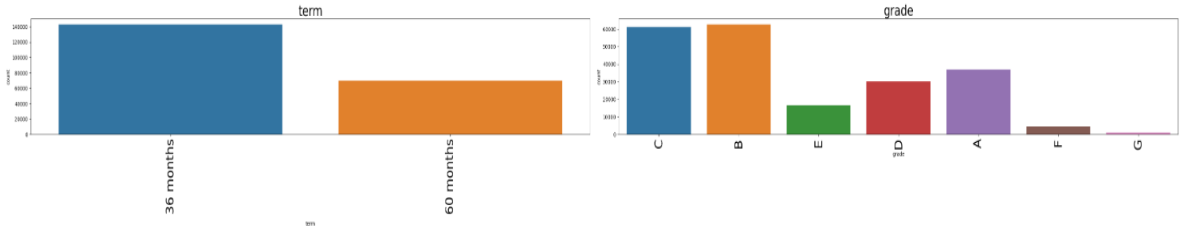
➢ **annual_inc**:



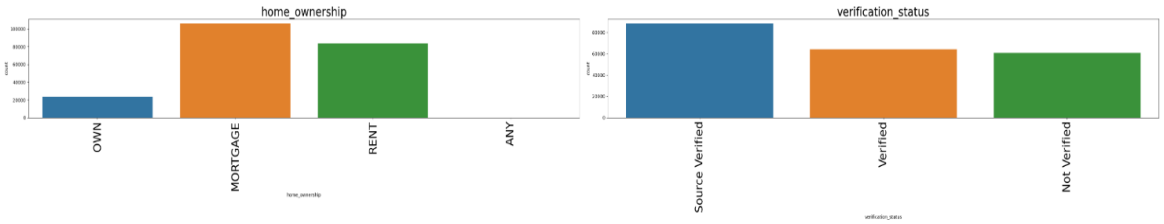
Inferences:

- The distribution of annual incomes is right-skewed, as the mean is greater than the median, and the maximum value is considerably higher than the mean.
- Most annual incomes fall within the range of $47,000 to $92,500, with the majority concentrated around the median value of $65,000.
- There is a wide variability in annual incomes, as indicated by the high standard deviation.

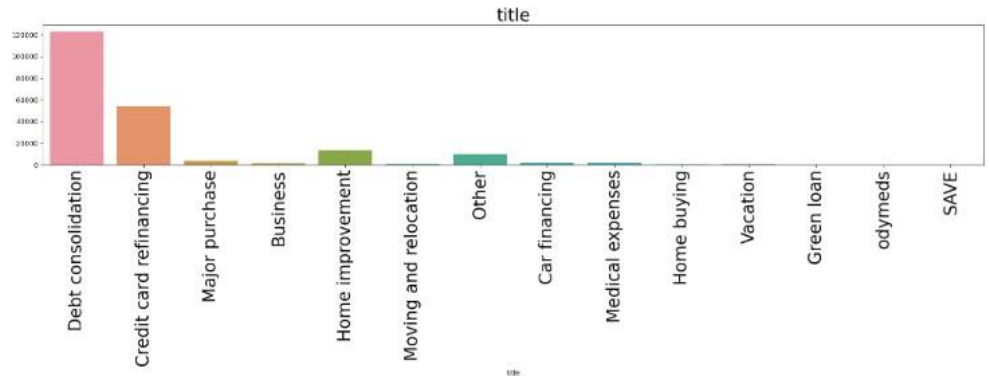
Univariate Analysis of categoric variables:



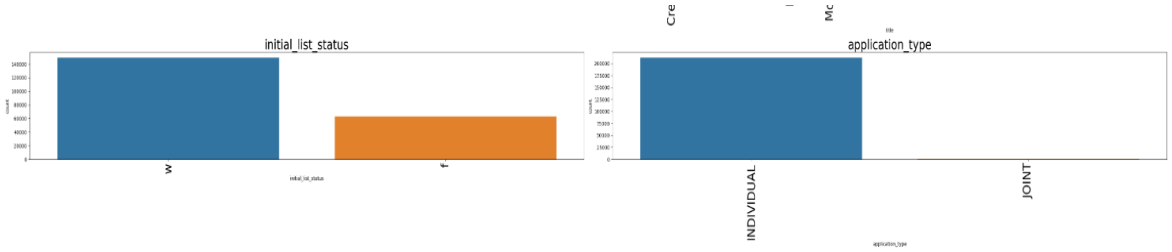
- 36 months term duration is more common than 60 months
- Higher grades such as B, C, A, D are more in numbers



- Borrowers who are having ownership status as Mortgage and Rent are more in number
- Verified financial income is more than not verified borrowers

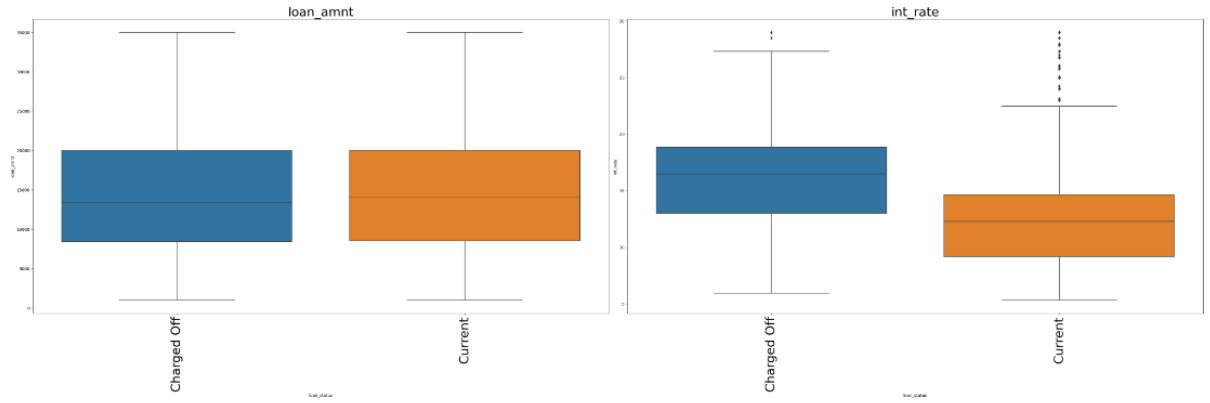


- Most of the people take loans for Debt consolidation, credit card refinancing, home improvement

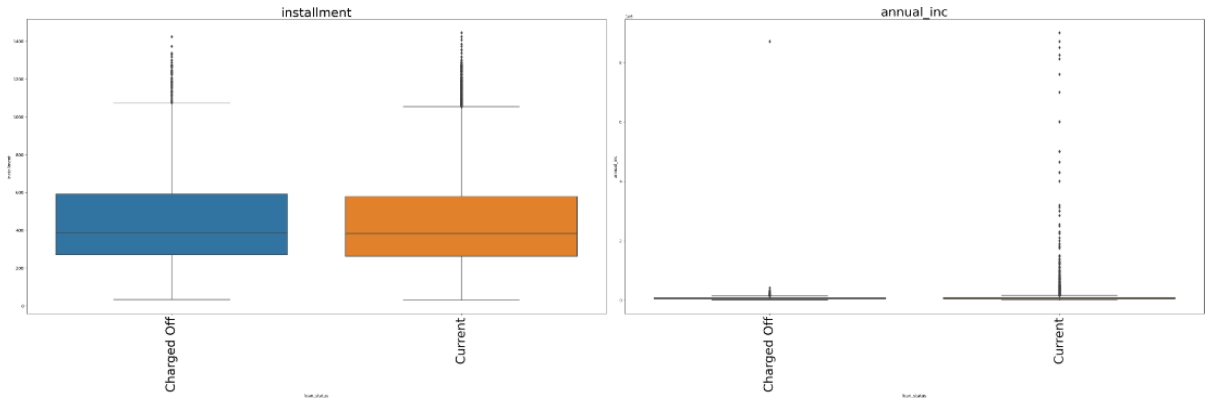


- Whole initial status is higher compared to fractional
- Individual loans are higher compared to joint loansf

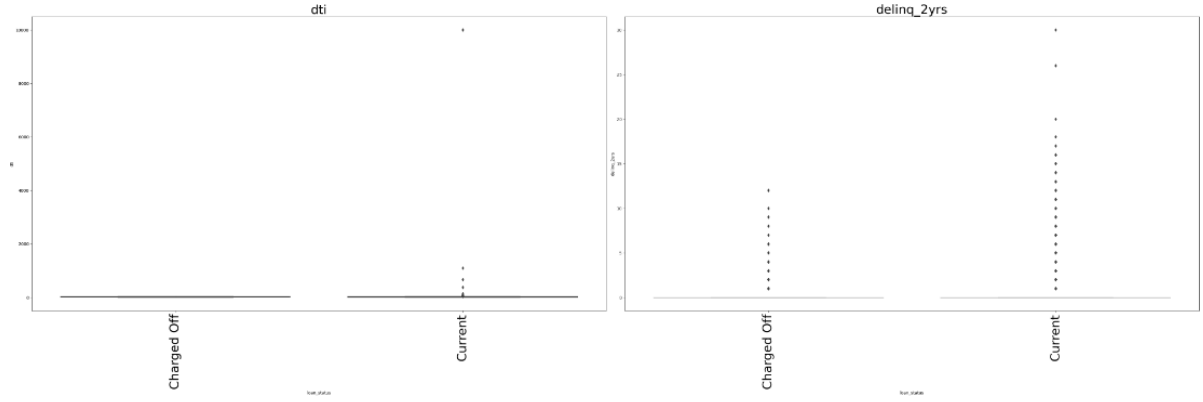
3.2. Relationship with the target variable



- No outliers present in Loan amount or Funded amount column indicates so there are no higher or lower loan amounts in any category than majority.
- There are only 2 people in Charged off category having higher interest rate. People who are in current have more interest rates.

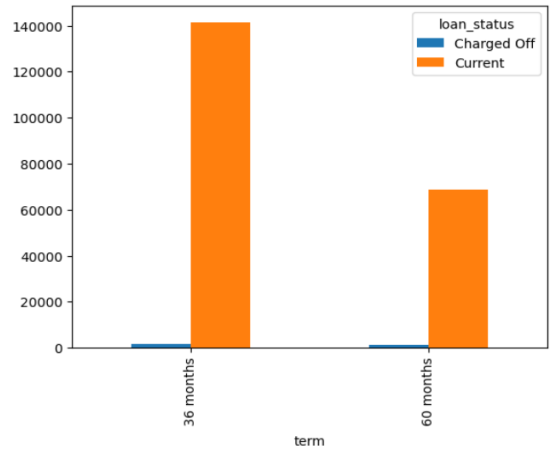


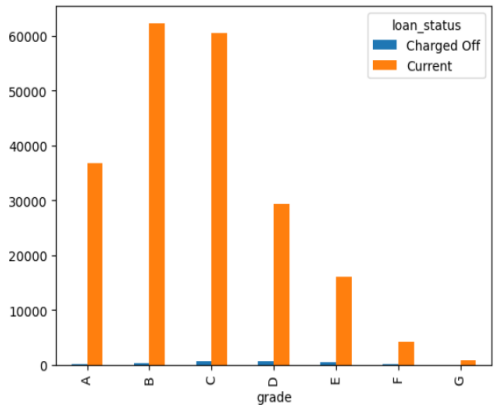
- Both charged off and current are having higher instalments
- Annual income in charged off category is significantly lower than current category. It means that people who have lower annual income tend to be Charged off

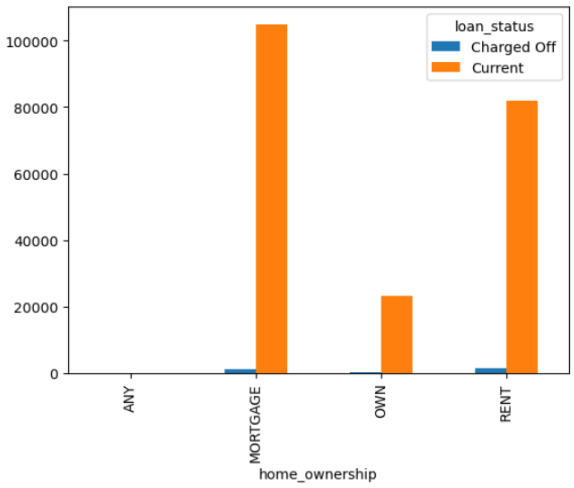


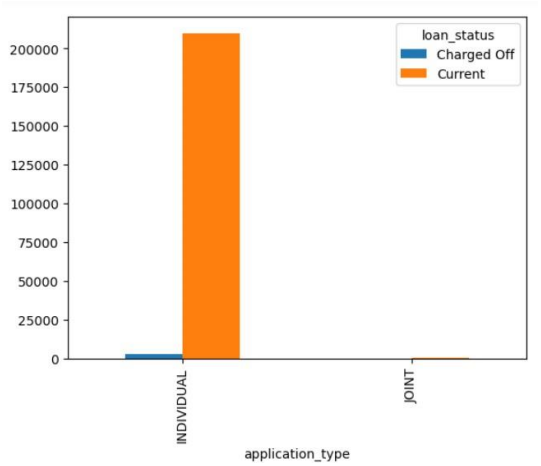
- Debt-to-income ratio and number of delinquencies in the past 2 years are higher in current category than charged off.

Bivariate analysis of categoric Vs categoric variable:









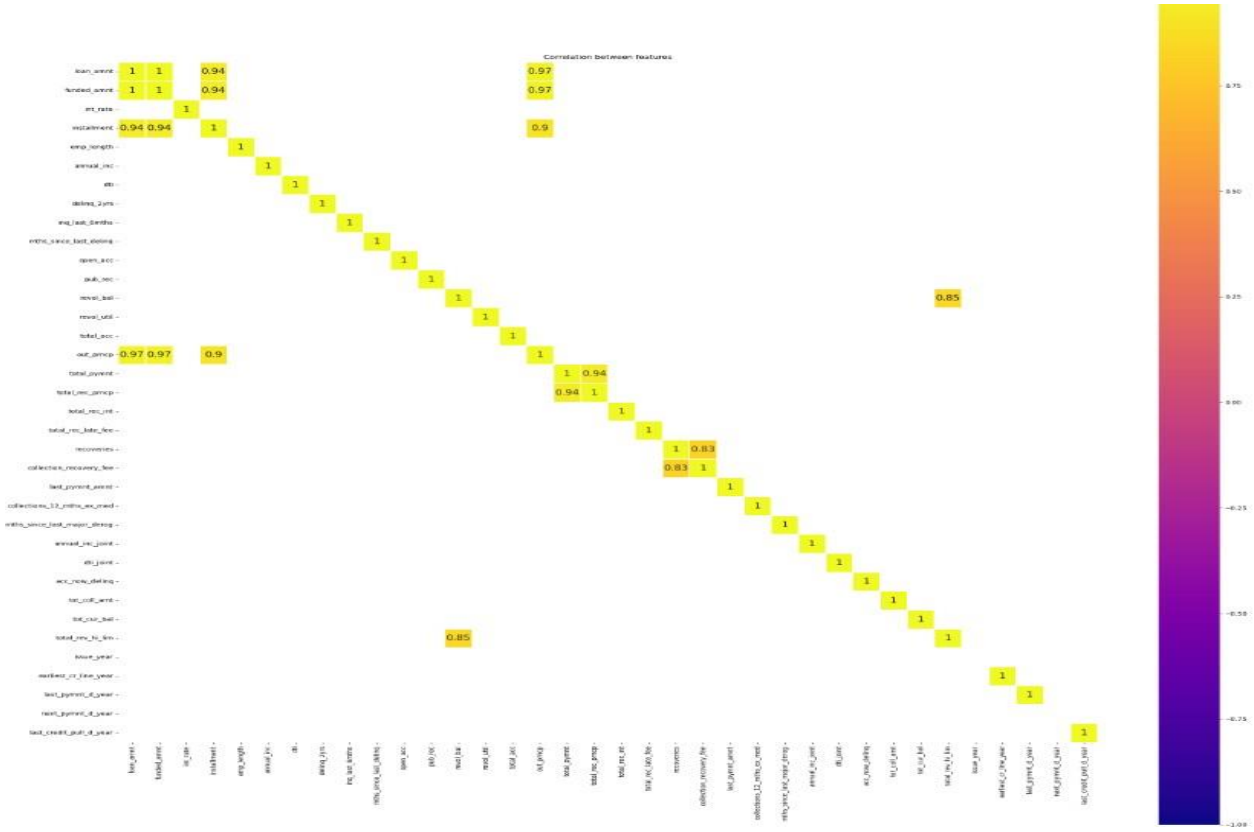
Inference:

- In all categories, current is greater than charged off.
- Most of the people who tend to be charged off have Mortgage or Rent home ownership.

3.3. Checking for Multicollinearity

- Multicollinearity occurs when two or more independent variables in a regression model are strongly correlated. It can lead to difficulties in distinguishing the individual effects of each predictor on the dependent variable. In the presence of multicollinearity, the estimated regression coefficients may become unstable and have inflated standard errors. This makes it challenging to assess the statistical significance of each predictor.

- Multicollinearity makes it difficult to interpret the contribution of each predictor variable independently. The coefficients may change significantly when predictors are added or removed from the model.
- The Variance Inflation Factor (VIF) is a common metric used to detect multicollinearity. It measures how much the variance of an estimated regression coefficient increases when predictors are correlated. High VIF values (usually above 5) indicate potential multicollinearity.
- Strategies to address multicollinearity include removing one of the correlated predictors, combining correlated predictors into a single variable, or collecting more data to reduce correlation.



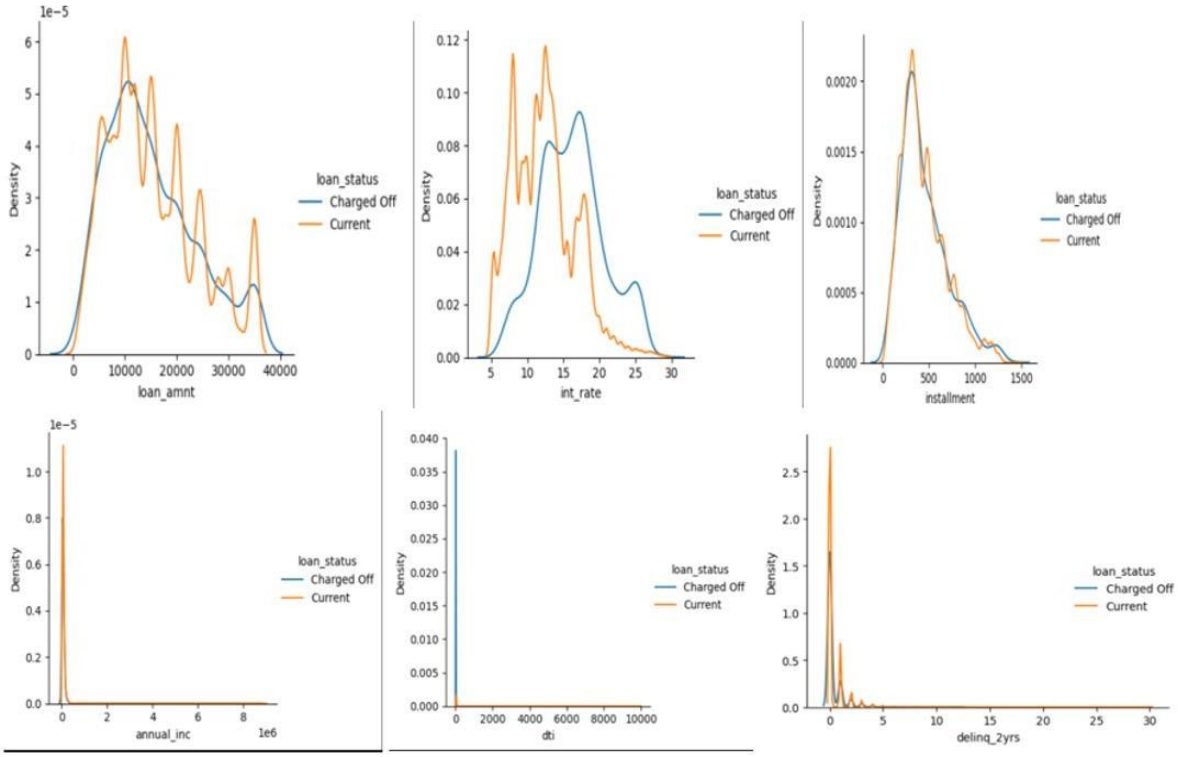
Inference:

- There is high correlation between loan amount with funded amount, instalment, and outstanding principal. As loan amount increases installment and principal amount increases.
- The total revolving credit limit is typically related to the current revolving balance. This correlation is expected.
- As the principal is repaid, the installment amount increases.
- The total payment should ideally be the sum of the received principal, and a high correlation is expected.
- Recovery and collection recovery fee are related as they both pertain to the recovery of funds after a loan default.

- As very high multicollinearity exists between these columns, we need to drop some pairs.
- Loan amount and installment are very important feature so as of now we will not drop them. Columns to be dropped are mentioned below:

FUNDED_AMNT
OUT_PRNCP
RECOVERIES
TOTAL_REV_HI_LIM
OPEN_ACC
TOTAL_REC_PRNCP

3.4. Checking for distribution of variables



Inference:

Fig1:

- Many individuals in both the 'Charged Off' and 'Current' categories are concentrated within the loan amount range of $10,000 to $20,000. This indicates that this loan amount range is popular among borrowers, irrespective of their loan status.
- Notably, the number of individuals in the 'Charged Off' category declines consistently after the $10,000 loan amount threshold. This trend implies that borrowers who eventually default ('Charged Off') tend to opt for lower loan amounts compared to those who remain in good standing ('Current').

Fig2:

- Most individuals in the 'Current' category exhibit a preference for interest rates between 6% and 15%. Beyond a 15% interest rate, the number of individuals in the 'Current' category decreases significantly, indicating a diminishing proportion of borrowers willing to accept higher interest rates.
- Conversely, individuals who eventually default ('Charged Off') tend to accept higher interest rates, with a concentration observed between 13% and 18%. This suggests that borrowers in the 'Charged Off' category may have been more willing to accept higher interest rates initially, potentially due to factors such as financial urgency or less favourable credit profiles.

Fig3:

- In the 'installment' column, while both 'Current' and 'Charged Off' categories exhibit similar distributions overall, there are instances where the 'Current' category shows higher peaks, suggesting that individuals who are currently in good standing with their loans may have opted for higher installment amounts, possibly indicating a higher financial stability or willingness to take on larger loan obligations.
- On the other hand, the 'Charged Off' category tends to have lower peaks, implying that individuals in this category may have opted for lower installment amounts, potentially indicating a lower financial capacity or a higher risk of default.

Fig4:

- In the 'annual_inc', there is a noticeable peak for the 'Current' category, indicating that a significant proportion of individuals in this category have relatively higher annual incomes. Conversely, the absence of a peak for the 'Charged Off' category suggests that individuals who have defaulted on their loans ('Charged Off') generally have lower annual incomes.

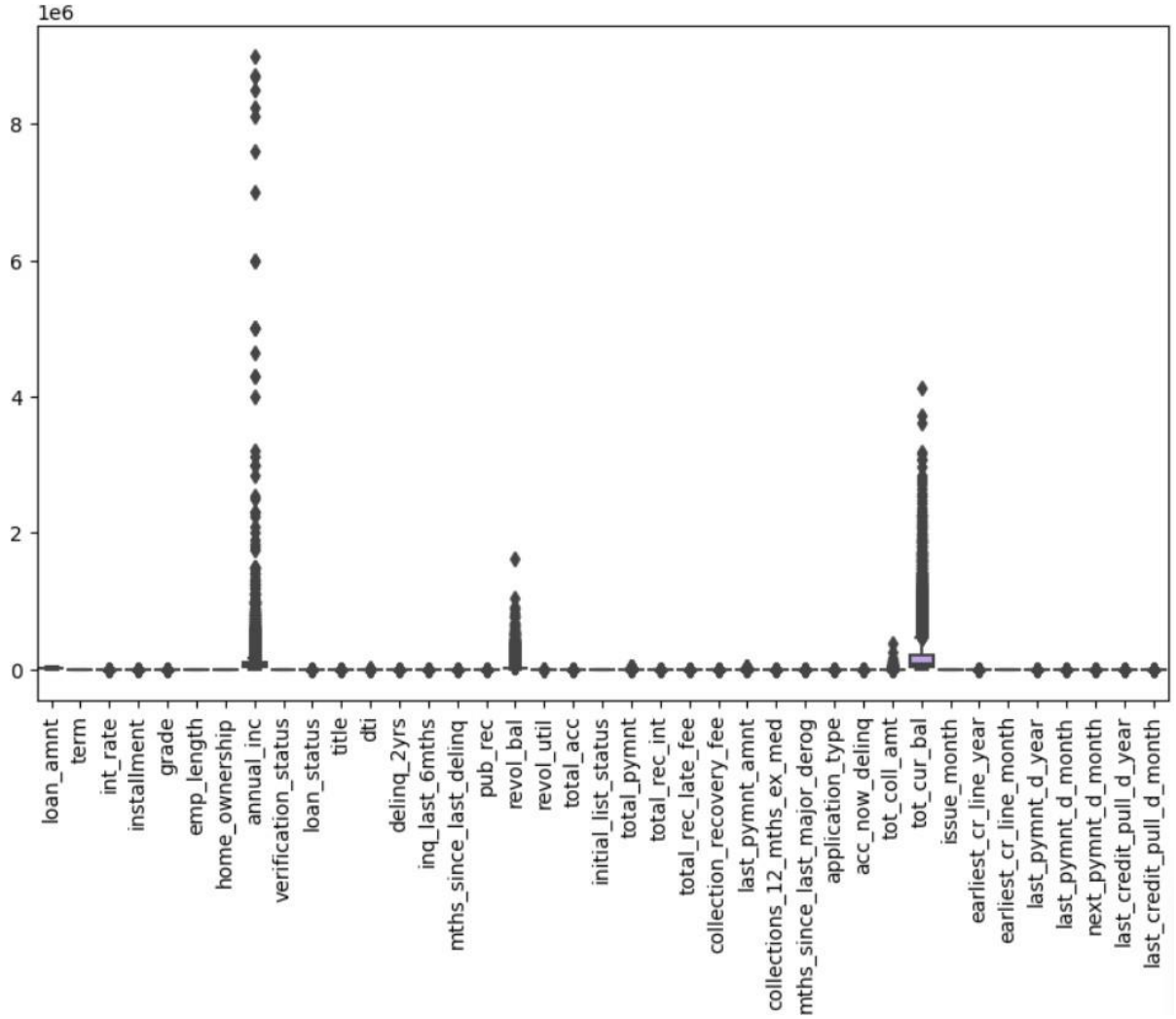
Fig5:

- In the 'dti', the higher peak observed for 'Charged Off' compared to the relatively lower peak for 'Current' suggests that individuals who have defaulted on their loans ('Charged Off') tend to have higher debt-to-income ratios. This indicates that a significant proportion of individuals in the 'Charged Off' category may have taken on loans that strain their financial capacity, potentially leading to their inability to repay the loans.
- Conversely, the lower peak for 'Current' suggests that individuals in good standing with their loans tend to have lower debt-to-income ratios, indicating a healthier financial position and a lower risk of default."

Fig6:

- In the 'delinq_2yrs' column, the higher distribution and peak for 'Current' compared to 'Charged Off' suggest that individuals who are currently in good standing with their loans ('Current') generally have a lower incidence of delinquencies in the past 2 years. This implies that borrowers who have managed to keep up with their payments ('Current') tend to have a better track record of timely payments, potentially indicating a higher level of financial responsibility and a lower risk of default.
- Conversely, the lower distribution and peak for 'Charged Off' indicate that individuals who have defaulted on their loans ('Charged Off') are more likely to have a history of delinquencies, which could have contributed to their inability to meet their loan obligations

3.5. Checking for presence of outliers and its treatment



- We can see many outliers in our dataset. Since this is a classification problem and many values will be helpful for our analysis.
- We will not be performing any outlier treatment.

3.6. Checking for statistical significance of variables

3.5.1. Pearson correlation coefficient

- The objective of the Pearson correlation coefficient test is to determine the strength and direction of the linear relationship between two continuous variables. It helps to quantify the extent to which changes in one variable are associated with changes in another variable in a linear fashion.
- We perform the Pearson correlation coefficient test to assess whether there is a statistically significant linear relationship between two continuous variables. This analysis is commonly used in fields such as statistics, psychology, economics, and

social sciences to explore associations between variables and to make predictions based on those associations.

- **Null Hypothesis (H0):** The null hypothesis for the Pearson correlation coefficient test states that there is no linear relationship between the two continuous variables. It suggests that the correlation coefficient (Pearson's r) calculated from the sample data is not significantly different from zero, indicating no linear association between the variables. Mathematically, it can be written as:
  H0: There is no significant linear relationship between Variable X and Variable Y.

- **Alternative Hypothesis (H1):** The alternative hypothesis for the Pearson correlation coefficient test contradicts the null hypothesis by suggesting that there is a significant linear relationship between the two continuous variables. It implies that changes in one variable are associated with changes in the other variable in a linear fashion. Mathematically, it can be written as:
- **H1:** There is a significant linear relationship between Variable X and Variable Y.

- From the above test we get following significant features:

INT_RATE
INSTALLMENT
EMP_LENGTH
ANNUAL_INC
DTI
INQ_LAST_6MTHS
REVOL_BAL
REVOL_UTIL
TOTAL_PYMNT
TOTAL_REC_INT
TOTAL_REC_LATE_FEE
COLLECTION_RECOVERY_FEE
LAST_PYMNT_AMNT
TOT_CUR_BAL
LAST_PYMNT_D_YEAR
LAST_CREDIT_PULL_D_YEAR

3.5.2. Chi-square test

- The objective of the chi-square test for independence is to determine whether there is a significant association between two categorical variables. It helps to understand if changes in one variable are related to changes in another variable, or if they occur independently of each other.
- We perform the chi-square test for independence to investigate whether there is evidence of a relationship between two categorical variables in a population. This analysis is essential in various fields such as social sciences, market research, biology, and healthcare, where understanding the association between different categorical

factors is crucial for making informed decisions.

- **Null Hypothesis (H0):** The null hypothesis for the chi-square test for independence states that there is no association between the two categorical variables. It suggests that changes in one variable are independent of changes in the other variable. Mathematically, it can be written as:
  H0: There is no significant association between Variable A and Variable B.

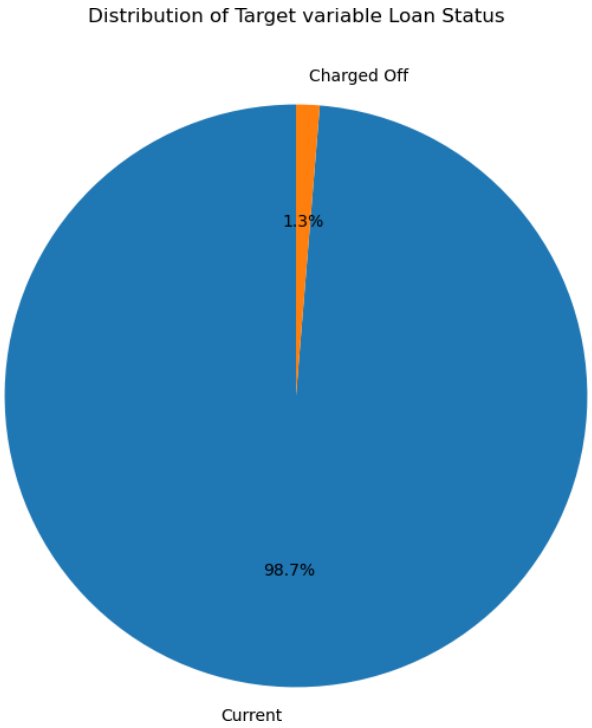
- **Alternative Hypothesis (H1):** The alternative hypothesis for the chi-square test for independence contradicts the null hypothesis by suggesting that there is a significant association between the two categorical variables. It implies that changes in one variable are associated with changes in the other variable. Mathematically, it can be written as:
- **H1:** There is a significant association between Variable A and Variable B.

- From chi-square test we get below features as important:

GRADE
HOME_OWNERSHIP
LOAN_STATUS
TITLE
INITIAL_LIST_STATUS
APPLICATION_TYPE
ISSUE_MONTH
EARLIEST_CR_LINE_MONTH
LAST_PYMNT_D_MONTH
NEXT_PYMNT_D_MONTH
LAST_CREDIT_PULL_D_MONTH

3.7. Checking for class imbalance and its treatment

- Class imbalance, where one class has significantly fewer instances, can impact the performance of machine learning models. Treatment strategies include resampling techniques such as oversampling and undersampling, algorithmic approaches like cost-sensitive learning and ensemble methods, and data augmentation which involves SMOTE analysis.
- **SMOTE (Synthetic Minority Over-sampling Technique):** SMOTE is a resampling technique used to address class imbalance by generating synthetic samples for the minority class
- Evaluation metrics, modifying decision thresholds, and leveraging transfer learning are additional considerations to address class imbalance. A comprehensive approach involves experimenting with various strategies to find the most effective balance for improved model performance.



1. Our target variable, is having two classes 0 (Current) and 1 (Charged off) with count and % as follows:
   - Current – 2,10,226 (98.7%)
   - Charged Off – 2,773 (1.3%)
2. Our dataset is not balanced.

4) Feature Engineering

4.1. Whether any transformations required

Yes, we have transformed variables having datetime as data type, to extract month and year from it mentioned below:

1. ISSUE_D
2. EARLIEST_CR_LINE
3. LAST_PYMNT_D
4. NEXT_PYMNT_D
5. LAST_CREDIT_PULL_D

4.2. Scaling the data

- Scaling is a preprocessing step used to standardize or normalize the range of features or variables in a dataset. It ensures that all variables contribute equally to the analysis and prevents features with larger scales from dominating those with smaller scales.

- The primary goal of scaling data is to bring all features to a similar scale, making comparisons and interpretations more meaningful.
- **Normalization:** Also known as min-max scaling, it rescales features to a range between 0 and 1. It is suitable when the distribution of data is unknown or when the standard deviation is small.
- **Standardization:** It centers the data around the mean and scales it to have a standard deviation of 1. It is preferred when the data has outliers or follows a Gaussian distribution.
- Standardization is particularly beneficial for algorithms that are sensitive to the scale of input features, such as support vector machines and k-nearest neighbors.
- It helps in improving the convergence of optimization algorithms, especially in gradient-based optimization used in many machine learning models.
- Standardization is widely applied in various machine learning models, including linear models, support vector machines, and neural networks, to ensure consistent feature scales and enhance model performance.
- **In our dataset we are using Standard Scaler as our scaling technique.**

	loan_amnt	int_rate	installment	emp_length	annual_inc	dti	delinq_2yrs	inq_last_6mths	mths_since_last_delinq	pub_rec
0	0.550651	-0.026276	0.920112	-1.468642	-0.156334	0.042573	-0.377181	0.505032	-0.102693	-0.353802
1	-0.494440	0.067856	-0.292213	0.256301	-0.461657	0.163165	-0.377181	-0.655429	0.212065	1.155774
2	-0.958925	-0.567535	-0.874500	-0.031189	-0.559361	0.408735	-0.377181	0.505032	-0.102693	-0.353802
3	-0.610561	-0.332205	-0.461660	1.118773	-0.363954	0.363943	-0.377181	1.665492	-0.102693	-0.353802
4	-0.662816	1.785763	-0.349685	-0.031189	-0.554769	0.412807	-0.377181	2.825953	2.289468	-0.353802

4.3. Encoding the categorical variables

- Encoding is the process of converting categorical data or text-based information into numerical representations that can be utilized by machine learning algorithms. It is a crucial step in data preprocessing, as many machine learning models require numerical input.
- Here, we have used label encoding and mapping for the variables mentioned below:

GRADE
HOME_OWNERSHIP
VERIFICATION_STATUS
TITLE
ISSUE_MONTH
EARLIEST_CR_LINE_MONTH,
LAST_PYMNT_D_MONTH
NEXT_PYMNT_D_MONTH
LAST_CREDIT_PULL_D_MONTH
TERM
LOAN_STATUS
APPLICATION_TYPE
INITIAL_LIST_STATUS

4.4. Feature selection

- We have selected following features from statistical analysis:
    - Numerical Features

INT_RATE
INSTALLMENT
EMP_LENGTH
ANNUAL_INC
DTI
INQ_LAST_6MTHS
REVOL_BAL
REVOL_UTIL
TOTAL_PYMNT
TOTAL_REC_INT
TOTAL_REC_LATE_FEE
COLLECTION_RECOVERY_FEE
LAST_PYMNT_AMNT
TOT_CUR_BAL
LAST_PYMNT_D_YEAR
LAST_CREDIT_PULL_D_YEAR

    - Categorical Features

GRADE
HOME_OWNERSHIP
LOAN_STATUS
TITLE
INITIAL_LIST_STATUS
APPLICATION_TYPE
ISSUE_MONTH
EARLIEST_CR_LINE_MONTH
LAST_PYMNT_D_MONTH
NEXT_PYMNT_D_MONTH
LAST_CREDIT_PULL_D_MONTH

4.5. Dimensionality Reduction

- Dimensionality Reduction is a technique used in data analysis and machine learning to reduce the number of input variables or features in a dataset while preserving as much relevant information as possible. This is particularly valuable when dealing with high-dimensional data, as it can lead to improved model performance, reduced computational complexity, and enhanced interpretability.

- Principal Component Analysis (PCA), t-Distributed Stochastic Neighbor Embedding (t-SNE) and Linear Discriminant Analysis (LDA) are some dominant techniques for dimension reduction. **We have applied feature selection to reduce the number of features as both dimension reduction and feature selection aims to reduce the columns of the dataset.**

5) Assumptions for base model (Logistic Regression)

1. **Binary Outcome:** The dependent variable (outcome) must be binary, meaning it has only two possible outcomes or categories. Logistic regression is specifically designed for modeling binary outcomes, such as 0 or 1, yes or no, success or failure, etc.
2. **Linearity of Log Odds:** The relationship between the predictor variables and the log odds of the outcome should be linear. This means that the log odds of the outcome should change linearly with changes in the predictor variables. This assumption can be assessed using methods like plotting the logit function.
3. **No Multicollinearity:** The predictor variables should not be highly correlated with each other (multicollinearity). High multicollinearity can lead to unstable estimates of the coefficients and inflated standard errors, making interpretation difficult. Multicollinearity can be assessed using measures like Variance Inflation Factor (VIF) or correlation matrices.
4. **Sufficient Sample Size:** Logistic regression requires a sufficiently large sample size to produce reliable estimates and valid statistical inference. While there is no strict rule for the minimum sample size, a commonly cited guideline is to have at least 10-20 cases with the least frequent outcome for each predictor variable to avoid issues with model convergence and estimation instability.

Checking the assumptions:

1. In our case dependent variable ‘Loan status’ is Binary with two unique categories as ‘Current’ and ‘Charged Off.’ We met assumption of binary outcome.
2. **Multicollinearity** occurs when independent variables in a regression model are highly correlated, causing instability in coefficients, inflated standard errors, and reduced precision. It is detected using methods like VIF and correlation matrices. Handling methods include variable selection, combining variables, and regularization. Prevention involves collecting more data or creating new features. Addressing multicollinearity is crucial for reliable regression analysis results.

    - By using VIF, we have dropped below mentioned columns as we were having multicollinearity:

LAST_PYMNT_D_MONTH
EARLIEST_CR_LINE_MONTH
HOME_OWNERSHIP
GRADE
LAST_CREDIT_PULL_D_MONTH
APPLICATION_TYPE

	VIF Factor	features
14	3.738857	total_rec_int
13	3.399861	total_pymnt
1	2.809001	title
3	2.710642	issue_month
0	2.534753	verification_status
6	2.122778	installment
5	1.887024	int_rate
17	1.723577	last_pymnt_amnt
2	1.437408	initial_list_status
18	1.424455	tot_cur_bal
11	1.422293	revol_bal
8	1.209051	annual_inc
4	1.187674	next_pymnt_d_month
12	1.176308	revol_util
16	1.126932	collection_recovery_fee
20	1.103418	last_credit_pull_d_year
10	1.095138	inq_last_6mths
19	1.074753	last_pymnt_d_year
7	1.015529	emp_length
9	1.009519	dti
15	1.006086	total_rec_late_fee

- There is no multi-collinearity now as the VIF of each of the variables is less than the threshold value of 5.

6) Performance Metrics for our base model

Since the dataset available with us is an Imbalanced Dataset, we cannot simply use Accuracy as a metric for evaluating the performance of the model. There are some metrics that work well with imbalanced datasets, as mentioned below:

- **ROC-AUC Score:** This metric is insensitive to class imbalance. It works by ranking the probabilities of prediction of the positive class label and calculating the Area under the ROC Curve which is plotted between True Positive Rates and False Positive Rates for each threshold value.
- **Recall Score:** It is the ratio of the True Positives predicted by the model and the total number of Actual Positives. It is also known as True Positive Rate.
- **Precision Score:** It is the ratio of True Positives and the Total Positives predicted by the model.

$$recall = \frac{true\ positives}{true\ positives\ +\ false\ negatives} \qquad precision = \frac{true\ positives}{true\ positives + false\ positives}$$

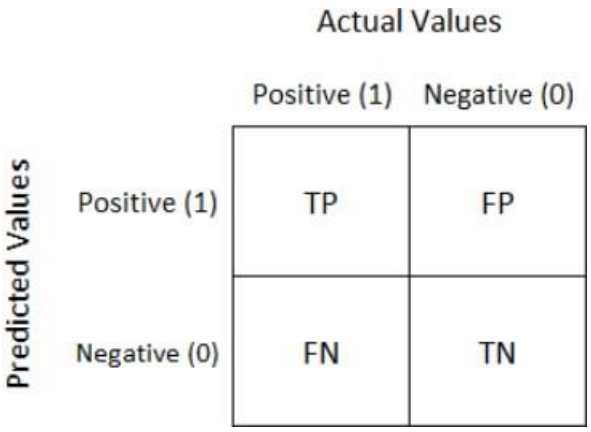
Recall and Precision Score Formulae

- **F1 score:** It is a metric commonly used in binary classification to assess the model's performance by considering both precision and recall. It is particularly useful when there is an uneven class distribution. The F1 score is the harmonic mean of precision and recall, and it provides a balance between the two.

$$F1 = \frac{2 \times (\text{Precision} \times \text{Recall})}{\text{Precision} + \text{Recall}}$$

F1 score formula

- **Confusion Matrix:** The confusion matrix helps us to visualize the mistakes made by the model on each of the classes, be it positive or negative. Hence, it tells us about misclassifications for both classes.



Confusion Matrix Formulae

In the context of predicting whether a user will default on housing loan or not, both precision and recall are useful. However, which one to prioritize depends on the business objective.

We shall use f1 score as the measure of performance as both precision and recall are important.

7) Steps in building the Logistic Regression Model

7.2.1. Split the Dataset for training and testing

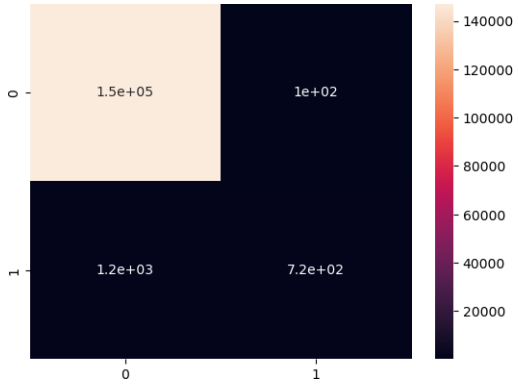
- We have split the dataset containing 212999 rows and 21 columns using the stratify parameter while splitting data in the ratio, 70%:30%. This parameter is only taking the output label as an argument. Hence, we must pass the target variable (y) as the parameter.

7.2.2. Measure of Model Performance using original data

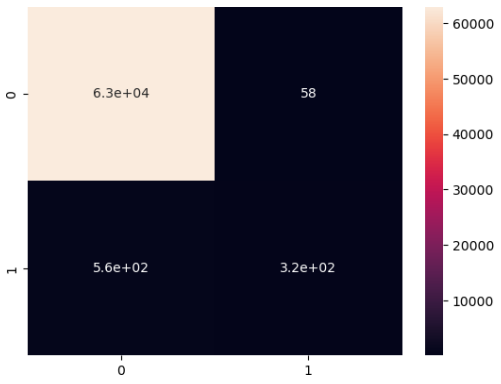
1. Using Default model parameters

Confusion Matrix

Train Data



Test Data



Performance on Training data

```
Training performance
              precision    recall  f1-score   support

           0       0.99      1.00      1.00    147204
           1       0.87      0.38      0.53      1895

    accuracy                           0.99    149099
   macro avg       0.93      0.69      0.76    149099
weighted avg       0.99      0.99      0.99    149099

AUC score: 0.9740495859532324
```

Performance on Testing data

```
Testing performance
              precision    recall  f1-score   support

           0       0.99      1.00      1.00     63022
           1       0.84      0.36      0.50       878

    accuracy                           0.99     63900
   macro avg       0.92      0.68      0.75     63900
weighted avg       0.99      0.99      0.99     63900

AUC score: 0.9698071772889951
```

The model has high accuracy and AUC scores on both training and testing data. It suggests that it generalizes well to unseen data but the model is over fit.

The relatively low recall for default cases indicates that the model may benefit from resampling techniques such undersampling and oversampling which improve its ability to correctly identify defaults

8) Methodology of Model Building

In our loan risk analysis dataset, where the target variable represents loan status (default and non-default) class imbalance is present where 'Charged Off' class is significantly less frequent than the 'Current' i.e. non- default class.

This class imbalance can lead to several challenges such as:

- Biased Model Learning : Models trained on imbalanced datasets tend to be biased towards the majority class (Current), as they have more instances to learn from. Consequently, they may struggle to accurately predict instances belonging to the minority class (Charged Off)
- Misleading Performance Metrics: Traditional performance metrics such as accuracy will not accurately reflect the model's effectiveness, especially in imbalanced datasets. A model may achieve high accuracy by simply predicting the majority class, while failing to identify instances of the minority class, which are more critical in loan risk assessment.
- Risk Assessment: Correctly identifying instances of the Charged Off class (defaults) is crucial in loan risk analysis as it directly impacts the lender's financial risk. Misclassifying defaults as non-defaults (false negatives) can lead to significant financial losses for the lender.

To address these challenges and improve model performance in loan risk analysis, we have used resampling techniques such as oversampling the minority class (Charged Off) or undersampling the majority class (Current).

8.1 Resampling Techniques:

1. Oversampling:

- Oversampling involves increasing the number of instances in the minority class to balance it with the majority class. This is typically achieved by generating synthetic examples that resemble the existing minority class instances.
- Synthetic Minority Over-sampling Technique (SMOTE) is one of the most popular oversampling techniques. It works by creating synthetic instances along the line segments connecting minority class instances in the feature space. This helps to fill in the gaps between existing minority class instances, effectively expanding the minority class.
- We employed oversampling techniques to balance the dataset by adjusting the ratio between the minority class (e.g., "Charged Off") and the majority class (e.g., "Current").
- Four different oversampling proportions were tested: 50-50, 60-40, 70-30, and 85-15.
- We evaluated the performance of several machine learning models, including Decision Tree, Random forest, Xgboost and Gaussian Naïve Bayes model etc.

2. Undersampling:

- Undersampling involves reducing the number of instances in the majority class to balance it with the minority class. This is achieved by randomly removing instances from the majority class until the desired balance is achieved.

- Random Undersampling is the simplest form of undersampling, where instances from the majority class are randomly selected and removed until the desired class balance is achieved.
- Undersampling techniques were applied to balance the dataset by adjusting the ratio between the minority class (e.g., "Charged Off") and the majority class (e.g., "Current").
- Four different undersampling proportions were tested: 50-50, 60-40, 70-30, and 85-15.
- The performance of several machine learning models, including Decision Tree, Random forest, Xgboost and Gaussian Naïve Bayes model was evaluated.

```
In [313]: df_oversampling
```

Out[313]:

	Split	Model Name	Training F1score	Testing F1score	Training Recall	Testing Recall	Training AUC Score	Testing AUC Score
0	50-50	Decision Tree	0.9719	0.9720	0.9600	0.9500	0.9774	0.9771
1	50-50	Random Forest	1.0000	1.0000	1.0000	1.0000	0.9999	0.9999
2	50-50	Xgboost	1.0000	1.0000	1.0000	1.0000	0.9999	0.9999
3	50-50	GaussianNB	0.8897	0.8880	0.8500	0.8500	0.9721	0.9714
4	70-30	Decision Tree	0.9505	0.9512	0.9900	0.9800	0.9689	0.9694
5	70-30	Random Forest	1.0000	1.0000	1.0000	1.0000	0.9999	0.9999
6	70-30	Xgboost	1.0000	1.0000	1.0000	1.0000	0.9999	0.9999
7	70-30	GaussianNB	0.9109	0.9130	0.8800	0.8900	0.9721	0.9726
8	60-40	Decision Tree	0.9728	0.9733	0.9600	0.9600	0.9775	0.9776
9	60-40	Random Forest	0.9993	0.9999	1.0000	1.0000	0.9999	0.9993
10	60-40	Xgboost	0.9999	0.9999	1.0000	1.0000	0.9999	0.9999
11	60-40	GaussianNB	0.9103	0.9117	0.8900	0.9000	0.9729	0.9729
12	85-15	Decision Tree	0.9581	0.9577	0.9900	0.9900	0.9691	0.9690
13	85-15	Random Forest	0.9999	0.9999	1.0000	1.0000	0.9999	0.9999
14	85-15	Xgboost	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999
15	85-15	GaussianNB	0.9157	0.9140	0.8900	0.8800	0.9718	0.9709

```
In [328]: df_undersampling
```

Out[328]:

	Split	Model Name	Training F1score	Testing F1score	Training Recall	Testing Recall	Training AUC Score	Testing AUC Score
0	50-50	Decision Tree	0.9767	0.9695	0.97	0.95	0.9832	0.9742
1	50-50	Random Forest	1.0000	0.9812	1.00	0.97	0.9999	0.9992
2	50-50	Xgboost	0.9997	0.9958	1.00	0.99	0.9999	0.9996
3	50-50	GaussianNB	0.8791	0.8701	0.84	0.82	0.9698	0.9627
4	70-30	Decision Tree	0.9744	0.9675	0.95	0.94	0.9750	0.9685
5	70-30	Random Forest	0.9999	0.9761	1.00	0.97	0.9999	0.9999
6	70-30	Xgboost	0.9999	0.9981	1.00	1.00	1.0000	0.9999
7	70-30	GaussianNB	0.8059	0.8072	0.72	0.73	0.9708	0.9733
8	60-40	Decision Tree	0.9709	0.9818	0.96	0.97	0.9778	0.9867
9	60-40	Random Forest	0.9992	0.9898	1.00	0.99	0.9999	0.9985
10	60-40	Xgboost	0.9999	0.9994	1.00	1.00	1.0000	0.9999
11	60-40	GaussianNB	0.8246	0.8273	0.74	0.76	0.9671	0.9707
12	85-15	Decision Tree	0.9730	0.9708	0.95	0.94	0.9737	0.9717
13	85-15	Random Forest	0.9999	0.9999	1.00	1.00	0.9999	0.9999
14	85-15	Xgboost	0.9999	0.9999	1.00	1.00	0.9999	0.9999
15	85-15	GaussianNB	0.7470	0.7451	0.65	0.66	0.9720	0.9752

We will interpret the performance of each individual model across different splits and resampling techniques:

Decision Tree:
- F1 Score: The decision tree model consistently achieves high F1 scores across all splits and resampling techniques, indicating a good balance between precision and recall.
- Recall: The decision tree model shows high recall values for both training and testing datasets, suggesting that it effectively identifies instances of the positive class (Charged Off).
- AUC Score: The AUC scores for the decision tree model are consistently high, indicating its ability to distinguish between positive and negative classes.
- We can conclude that Decision Tree model with over sampling technique having split 85-15 % performs the best across all other Decision Tree models.

Random Forest:
- We can conclude that for all sampling techniques and different splits, Random Forest is giving us overfit results.

XGBoost:
-We can conclude that for all sampling techniques and different splits, Random Forest is giving us overfit results.

GaussianNB:
- F1 Score: GaussianNB achieves relatively high F1 scores across all splits and resampling techniques, indicating good balance between precision and recall.
- Recall: GaussianNB shows high recall values for both training and testing datasets, suggesting its ability to identify instances of the positive class.
- AUC Score: The AUC scores for GaussianNB are consistently high, indicating its ability to discriminate between positive and negative classes.
-We can conclude that GaussianNB model with oversampling technique and 60-40 % split performs the best across all other GaussianNB model.

9) Model Evaluation

In the case of a loan risk analysis dataset with a severe class imbalance, where the majority of loans are non-default (negative class) and only a small fraction is default (positive class), it's essential to choose evaluation metrics that are robust to class imbalance. In such scenarios, metrics like accuracy can be misleading because a naive model that predicts everything as the majority class would achieve high accuracy due to the class imbalance.
Recall measures the ability of the classifier to find all positive instances (defaults) in the dataset. In loan risk analysis, recall is crucial because missing even a single default can be costly for the lender. High recall indicates that the model can effectively identify defaults, minimizing the risk of missing potentially problematic loans.

True Positive (TP): A loan correctly identified as high-risk.
False Negative (FN): A high-risk loan incorrectly classified as low-risk.
A high recall indicates that the model is effective in capturing a significant portion of the high-risk loans, minimizing the chances of approving loans that may default.

F1-score is also an important performance metric. It provides a balance between precision and recall, which is useful when there's an imbalance between false positives and false negatives.
F1-score is particularly valuable in loan risk analysis because it considers both the ability to identify defaults (recall) and the ability to avoid misclassifying non-defaults (precision).

Here, we have considered Recall score of 0.99 as best score. Hence,we will compare all models having Recall value of 0.99 in under sampling and oversampling.

```
1 df_oversampling[df_oversampling['Testing Recall']==0.99]
```

	Split	Model Name	Training F1score	Testing F1score	Training Recall	Testing Recall	Training AUC Score	Testing AUC Score
12	85-15	Decision Tree	0.9581	0.9577	0.99	0.99	0.9691	0.969

```
1 df_undersampling[df_undersampling['Testing Recall']==0.99]
```

	Split	Model Name	Training F1score	Testing F1score	Training Recall	Testing Recall	Training AUC Score	Testing AUC Score
2	50-50	Xgboost	0.9997	0.9958	1.0	0.99	0.9999	0.9996
9	60-40	Random Forest	0.9992	0.9898	1.0	0.99	0.9999	0.9985

Based on above data, it seems that the Decision Tree model with an 85-15 split in oversampling achieves a testing recall score of 0.99. In contrast, in undersampling, both XGBoost and Random Forest models have recall values of 0.99.

While XGBoost and Random Forest models exhibit high recall values, it's noted that they demonstrate overfitting according to the F1-score and AUC_score metrics. Therefore, the Decision Tree model is chosen as the best model due to its robust performance across multiple metrics and its ability to generalize well to unseen data.

It's crucial to prioritize models that exhibit good generalization performance and avoid overfitting, as overfit models may not perform well on unseen data. By selecting the Decision Tree model, we are opting for a model that strikes a balance between performance and generalization ability, making it a suitable choice for deployment in real-world scenarios.

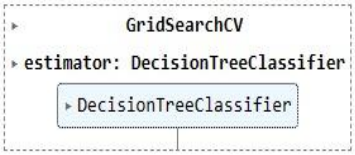
9.1 Hyper-parameter Tuning:

Hyperparameter tuning is a crucial step in model development, regardless of the initial performance of the model. It ensures that the model is optimized for the specific task at hand, leading to better performance, interpretability, and generalization ability.

We will apply Grid search on our best model to further optimize it's performance.

```
from sklearn.model_selection import GridSearchCV
X_train, X_test, y_train, y_test = train_test_split(X,y,test_size=0.3,random_state=42)
dt_classifier = DecisionTreeClassifier()
param_grid = {
    'criterion': ['gini', 'entropy'],
    'max_depth': [5, 10, 15],
    'min_samples_split': [2, 5, 10],
    'min_samples_leaf': [1, 2, 4]
}

grid_search = GridSearchCV(estimator=dt_classifier, param_grid=param_grid, cv=5,
                           scoring='accuracy', n_jobs=-1)
grid_search.fit(X_train, y_train)
```




```
performance(y_train,X_train,dtmodel)
```

	precision	recall	f1-score	support
0	1.00	1.00	1.00	147016
1	1.00	1.00	1.00	125226
accuracy			1.00	272242
macro avg	1.00	1.00	1.00	272242
weighted avg	1.00	1.00	1.00	272242

AUC score: 0.9999989453687095
f1_score: 0.9999600709140566

```
performance(y_test,X_test,best_model)
```

	precision	recall	f1-score	support
0	1.00	1.00	1.00	63210
1	1.00	1.00	1.00	53466
accuracy			1.00	116676
macro avg	1.00	1.00	1.00	116676
weighted avg	1.00	1.00	1.00	116676

AUC score: 0.9996533146224016
f1_score: 0.999625923501356

Upon conducting hyperparameter tuning using grid search for our best model, Decision Tree, we encountered an overfitting issue. Despite our efforts to optimize the model's parameters, the resulting model exhibited signs of overfitting, indicating that it may have learned noise or irrelevant patterns from the training data that do not generalize well to unseen data.

This overfitting phenomenon suggests that the model's complexity may have been unnecessarily

increased, leading to a lack of generalization ability.

Hence, we will select our original model Decision Tree with 85-15 split as best model as it does not need further optimization in it's performance.

Decision Tree Model Evaluation:

Type of split: 85-15 %
Resampling technique: Oversampling

Classification Report:

```
1 performance(y_train,X_train,DT_model)
```

	precision	recall	f1-score	support
0	0.99	0.94	0.96	147131
1	0.93	0.99	0.96	125111
accuracy			0.96	272242
macro avg	0.96	0.96	0.96	272242
weighted avg	0.96	0.96	0.96	272242

AUC score: 0.9691896500665753
f1_score: 0.9582405799466206

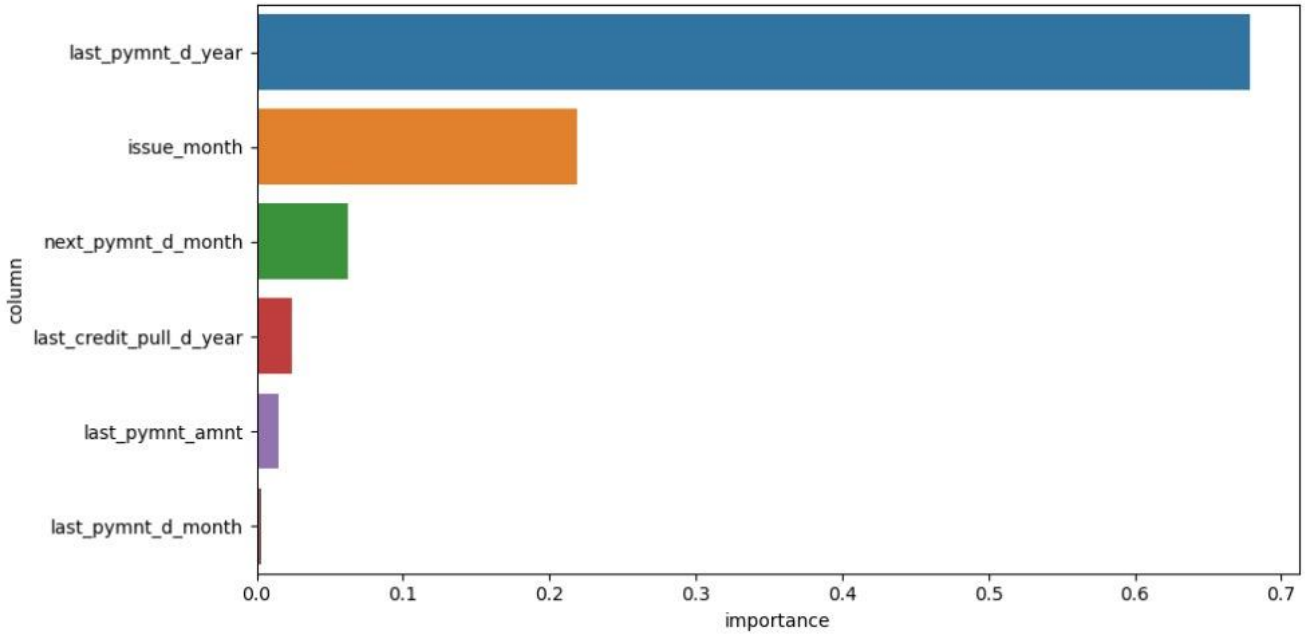
```
1 performance(y_test,X_test, DT_model)
```

	precision	recall	f1-score	support
0	0.99	0.94	0.96	63095
1	0.93	0.99	0.96	53581
accuracy			0.96	116676
macro avg	0.96	0.96	0.96	116676
weighted avg	0.96	0.96	0.96	116676

AUC score: 0.9690427421350195
f1_score: 0.9579319727891157

10) Visualization

Here's an interpretation of the importance of each feature from our best model (Decision Tree) in the context of loan risk analysis:



last_pymnt_d_year (0.678385):
This feature appears to be highly important according to the model. It represents the year of the last payment date. A higher importance suggests that the year of the last payment has a substantial impact on predicting loan risk.

issue_month (0.218913):
The month when the loan was issued is also deemed important. The model suggests that there might be a discernible pattern or relationship between the issuance month and loan risk.

next_pymnt_d_month (0.061897):
The month of the next payment date is considered moderately important. It implies that the timing of the upcoming payment might provide insights into the likelihood of loan risk.

last_credit_pull_d_year (0.023733):
The year of the last credit pull is less important than some other features but still contributes to the model's predictions. It could indicate that recent credit inquiries have some influence on loan risk.

last_pymnt_amnt (0.014467):
The amount of the last payment is a feature with moderate importance. It suggests that the size of the last payment made by the borrower may be indicative of loan risk.

last_pymnt_d_month (0.002605):
The month of the last payment is considered less important. While it has some impact on the model, other features may have a more substantial influence on predicting loan risk.

11) Business Impact and Recommendations

In the context of the loan risk analysis problem and the identified important features, the solution and

recommendations based on the feature importance scores can have significant implications for the business. Here are some observations and recommendations:

I. Impact on Decision-Making:

- The identified important features, such as last_pymnt_d_year, issue_month, and next_pymnt_d_month, suggest that temporal factors related to loan repayment, including the timing of the last payment and the month of loan issuance, significantly impact the assessment of loan risk.
- Consider placing more emphasis on recent payment behavior, as indicated by the last_pymnt_d_year and next_pymnt_d_month features. Loans with more recent payments may be perceived as less risky, and this information could be crucial in decision-making.

II. Seasonality Considerations:

- The issue_month feature's importance implies that seasonality or other temporal patterns related to the month of loan issuance should be taken into account. For example, borrowers taking loans during certain months may have different risk profiles.

III. Creditworthiness Insights:

- The last_credit_pull_d_year feature indicates that recent credit inquiries or credit checks play a role in assessing loan risk. Lending institutions may benefit from considering borrowers' recent credit activities to gain insights into their creditworthiness.

IV. Payment Behavior Impact:

- The last_pymnt_amnt feature highlights the importance of the amount of the last payment in evaluating loan risk. This suggests that borrowers making higher last payments may be considered less risky, and this factor should be factored into risk assessments.

V. Recommendations:

- Recommendations are made with confidence based on the importance scores, but it's crucial to continuously monitor and validate the model's performance. Regular updates to the model, incorporating new data and re-evaluating feature importance, ensure that recommendations remain relevant over time.

VI. Adaptation and Continuous Improvement:

- Lending institutions should be open to adapting their models based on changing business conditions, regulatory requirements, and shifts in customer behavior. Regular model evaluations and updates are essential for continuous improvement.

12) Limitations and Future Enhancements

Considering the information provided in the Business Impact and Recommendations section, we can identify several potential limitations and areas for future enhancement:

1. Seasonal Variations:

- Limitation: While the model considers the month of loan issuance (issue_month) as an important feature, it may not fully capture all seasonal variations and trends that affect loan repayment behavior.
- Enhancement: Further analysis could explore additional temporal factors, such as economic indicators, holidays, or seasonal employment patterns, to improve the model's ability to capture seasonality in loan risk assessment.

2. Data Quality and Completeness:

- Limitation: The model's performance and recommendations heavily rely on the quality and completeness of the dataset, particularly in terms of historical payment and credit information.
- Enhancement: Conducting thorough data quality assessments and implementing data cleansing and enrichment techniques can enhance the reliability and accuracy of the model predictions. Additionally, integrating alternative data sources or external datasets may provide additional insights into borrower behavior and creditworthiness.

3. Model Interpretability:

- Limitation: While Decision Trees offer interpretability, complex interactions between features and nonlinear relationships may not be fully captured or understood.
- Enhancement: Exploring alternative modeling techniques, such as ensemble methods or explainable AI approaches, can improve model interpretability while maintaining predictive performance. Additionally, providing stakeholders with clear explanations of model decisions and recommendations can enhance trust and usability.

4. Generalizability:

- Limitation: The model's performance may vary across different demographic groups, loan types, or market conditions, limiting its generalizability.
- Enhancement: Conducting robust model validation and testing across diverse datasets and scenarios can assess the model's generalizability and identify potential biases or limitations. Additionally, incorporating demographic or segment-specific features into the model can enhance its ability to address diverse borrower profiles and market conditions.

Overall, addressing these limitations and pursuing future enhancements can lead to a more robust and effective loan risk assessment model, enabling lending institutions to make informed decisions and mitigate risks more effectively.

13) Closing Reflections and Future Directions

Throughout this process, several key insights and lessons have been gained, shaping our approach to future projects. Here are the closing reflections on what we have learned and what we would do differently next time:

1. **Model Evaluation Importance:** We have reinforced the significance of robust model evaluation techniques in assessing performance accurately. Understanding the limitations of evaluation metrics

and the potential for overfitting is crucial for developing reliable models.

2. **Hyperparameter Tuning Considerations:** We have learned the importance of balancing model complexity and generalization when performing hyperparameter tuning. Next time, we would explore more sophisticated techniques to prevent overfitting during parameter optimization.

3. **Feature Importance Interpretation:** The process of interpreting feature importance has highlighted the importance of domain knowledge and context in understanding model decisions. Next time, we would collaborate more closely with domain experts to ensure a deeper understanding of feature relevance.

4. **Data Quality and Preprocessing:** We have realized the critical role of data quality and preprocessing in model performance. Next time, we would invest more time in data exploration, cleaning, and feature engineering to enhance the quality and relevance of the input data.

5**. Continuous Learning and Improvement:** The iterative nature of model development underscores the importance of continuous learning and improvement. Next time, we would prioritize regular model updates, incorporating new data and insights to ensure the model remains relevant and effective over time.

6**. Collaboration and Communication:** Effective collaboration and communication among team members, stakeholders, and domain experts have been instrumental in driving project success. Next time, we would foster even stronger collaboration and communication channels to facilitate knowledge sharing and decision-making.

Overall, this process has provided valuable insights into the complexities of machine learning model development and deployment. By reflecting on these lessons and implementing improvements in future projects, we aim to enhance our capabilities and deliver even more impactful solutions.

14) Appendix

Data Dictionary

1. id: To uniquely identify every loan in the dataset.
2. member_id: To identify the borrower to who has applied for the loan.
3. loan_amnt: The listed amount of the loan applied for by the borrower.
4. funded_amnt : The amount that was sanctioned by the LC.
5. term : The number of payments on the loan. Values are in months and can be either 36 or 60.
6. int_rate: Interest Rate on the loan
7. installment : The monthly payment owed by the borrower if the loan originates.
8. grade: LC assigned loan grade which depends on the borrower’s credit score.
9. sub_grade: LC assigned loan subgrade
10. emp_title: The job title supplied by the Borrower when applying for the loan.*
11. emp_length : Employment length in years. Possible values are between 0 and 10 where 0 means less than one year and 10 means ten or more years.
12. home_ownership: The home ownership status provided by the borrower during registration or obtained from the credit report. Our values are: RENT, OWN, MORTGAGE, OTHER.

13. annual_inc : The self-reported annual income provided by the borrower during registration.
14. verification_status : Indicates if income was verified by LC, not verified, or if the income source was verified
15. issue_d : The month which the loan was funded
16. loan_status : Current status of the loan
17. purpose : A category provided in the form of a code to indicate the purpose for the loan.
18. title : Explaining the 'purpose' of the loan.
19. dti : The debt to income ratio is the ratio of how much the borrower owes every month to the borrower's income every month.
20. delinq_2yrs : The number of delinquencies(late installment payment) by the borrower in the past 2 years.
21. earliest_cr_line : The month-year the borrower's earliest reported credit line was opened
22. inq_last_6mths : Inquiries for loans made by the borrower over the past 6 months.
23. mths_since_last_delinq : Months that have passed since the borrower last missed the timely payment of installment.
24. open_acc : The number of open credit lines in the borrower's credit file.
25. pub_rec Number of derogatory public records
26. revol_bal : Total credit revolving balance
27. revol_util : Revolving line utilization rate, or the amount of credit the borrower is using relative to all available revolving credit.
28. total_acc : The total number of credit lines currently in the borrower's credit file
29. initial_list_status : The initial listing status of the loan. Possible values are – W(whole), F(fractional)
30. out_prncp : Remaining outstanding principal for total amount funded
31. total_pymnt : Payments received to date for the total amount funded.
32. total_rec_prncp : Principal received till date.
33. total_rec_int Interest received till date.
34. total_rec_late_fee : Late fees received to date.
35. recoveries : Total recovery procedures initiated against the borrower.
36. collection_recovery_fee : The fees collected during the recovery procedures.
37. last_pymnt_d The last month when payment was received.
38. last_pymnt_amnt : The last payment amount received.
39. next_pymnt_d : Next scheduled payment date.
40. last_credit_pull_d : The most recent month LC pulled credit for this loan
41. collections_12_mths_ex_med : Number of collections in 12 months excluding medical collections
42. mths_since_last_major_derog : Months since most recent 90-day delinquency or worse rating
43. application_type Indicates whether the loan is an individual application or a joint application with two co-borrowers
44. annual_inc_joint : The combined self-reported annual income provided by the co-borrowers during registration
45. dti_joint : A ratio calculated using the co-borrowers' total monthly payments on the total debt obligations, excluding mortgages and the requested LC loan, divided by the co-borrowers' combined self-reported monthly income
46. acc_now_delinq : The number of accounts on which the borrower is now delinquent
47. tot_coll_amt : Total collection amounts ever owed by the borrower
48. tot_cur_bal : Total current balance of all accounts owned by the borrower
49. total_rev_hi_lim : Total high credit/credit limit