**Table of Contents:**

|  |  |
| --- | --- |
| S.NO. | Topic |
| 1 | Overview |
| 2 | Business problem goals |
| 3 | Topic survey in brief |
| 4 | Critical assessment of topic survey |
| 5 | Methodology to be followed |
| 6 | Results and Analysis |
|  | Conclusions |
| 6 | References |

**Project Details**

1. **Overview**

This project leverages machine learning techniques to provide a robust solution for financial institutions to make informed lending decisions. By accurately predicting default risk, sensitive defaulters, key influential factors leading to loan default. The project aims to minimize risk and optimize loan portfolio management.

1. **Business problem statement**

2.1 Business Problem Understanding:

In the competitive realm of financial lending, businesses grapple with diverse customer profiles. Understanding the multifaceted factors influencing loan repayment is crucial in minimizing risk and ensuring profitability. To strategically reduce loan defaults, businesses need precise insights and targeted strategies that address specific risk factors identified through data analysis.

2.2 Business Objective:

The business objective is to leverage various machine learning techniques to identify key factors influencing loan defaults. By uncovering these influential factors, the project aims to propose targeted strategies to improve the decision-making process, reduce risk, and enhance profitability for lending institutions.

2.3 Approach

The comprehensive approach includes data collection, preprocessing, exploratory data analysis and effectively implementing machine learning algorithms and techniques to identify chances of loan defaults of an individual along with influential factors for loan defaults. Additionally, a default classification step is incorporated, enabling the categorization of loans based on risk levels. Strategies are then tailored based on these factors and validated through testing and continuous monitoring. The process is thoroughly documented, and insights are presented to optimize loan approval processes department.

2.4 Conclusions

In conclusion, our approach is to predict credit risk combined robust data analysis with regression modelling and default classification. By systematically identifying influential factors and categorizing loans, we enable businesses to tailor strategies with precision. The iterative process of validation and continuous monitoring ensures the practical effectiveness of these strategies. Ultimately, our comprehensive documentation and presentation of findings provide a roadmap for sustained improvement, contributing to the overall success and competitiveness of financial lending institutions.

1. **Topic Survey in brief**

3.1 Problem understanding:

The challenge is figuring out why some loans default while others do not. The goal is to understand what aspects like borrower profile, loan amount, interest rate, and repayment history contribute to defaults. By identifying specific areas of risk based on historical data, the aim is to make lending practices more secure and profitable.   
  
3.2 Current solution to the problem:

Financial institutions employ various technical solutions to manage and enhance their loan approval process:

* Credit Scoring Systems
* Financial Statements Analysis
* Risk Assessment Models
* Loan Repayment History Analysis
* Predictive Analytics
* Automated Loan Approval Systems
* Customer Relationship Management (CRM) Systems

3.3 Proposed solution to the problem:

The process involves gathering comprehensive data on loan applicants, conducting thorough data preprocessing and exploratory analysis, and applying various machine learning algorithms and techniques to model the relationship between features and loan defaults. The entire process is documented, and findings are presented in a comprehensive report, allowing for continuous monitoring and refinement based on ongoing performance evaluation.

1. **Critical Assessment of Topic Survey:** 
   1. Key Area Gaps Identified:

The topic survey may lack in-depth exploration of borrower profiles and the potential underutilization of available data. The project can add value by employing advanced analytics to extract nuanced insights, emphasizing sophisticated customer segmentation and introducing predictive modelling for proactive risk management strategies.

* 1. **Key Gaps to Solve:**

The primary gaps involve limited exploration of data's full potential, potential absence of predictive modelling, and the need for more detailed customer segmentation. The project aims to address these gaps by enhancing data utilization, implementing predictive modelling, and employing advanced customer segmentation techniques.

1. **Methodology to be followed:**

5.1 Data Collection and Preprocessing:

Collect data on borrower profiles, loan details, repayment history, and defaults. Clean and preprocess data to handle missing values, encode categorical variables, and ensure consistency.

**Data Dictionary:**

* **id**: A unique LC assigned ID for the loan listing
* **member\_id**: A unique LC assigned Id for the borrower member.
* **loan\_amnt**: The listed amount of the loan applied for, by the borrower. If at some point in time, the credit department reduces the loan amount, then it will be reflected in this value.
* **funded\_amnt**: The total amount committed to that loan at that point in time.
* **funded\_amnt\_inv**: The total amount committed by investors for that loan at that point in time.
* **term**: The number of payments on the loan. Values are in months and can be either 36 or 60.
* **int\_rate**: Interest Rate on the loan.
* **installment**: The monthly payment owed by the borrower if the loan originates.
* **grade**: LC assigned loan grade.
* **sub\_grade**: LC assigned loan subgrade.
* **emp\_title**: The job title supplied by the Borrower when applying for the loan.
* **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.
* **home\_ownership**: The home ownership status provided by the borrower during registration. Our values are: RENT, OWN, MORTGAGE, OTHER.
* **annual\_inc**: The self-reported annual income provided by the borrower during registration.
* **verified\_status**: Indicates if the borrowers income was verified by LC or not verified.
* **issue\_d**: The month which the loan was funded.
* **pymnt\_plan**: Indicates if a payment plan has been put in place for the loan.
* **purpose**: A category provided by the borrower for the loan request.
* **title**: The loan title provided by the borrower.
* **zip\_code**: The first 3 numbers of the zip code provided by the borrower in the loan application.
* **addr\_state**: The state provided by the borrower in the loan application.
* **dti**: A ratio calculated using the borrower’s total monthly debt payments on the total debt obligations, excluding mortgage and the requested LC loan, divided by the borrower’s self-reported monthly income.
* **delinq\_2yrs**: The number of 30+ days past-due incidences of delinquency in the borrower's credit file for the past 2 years.
* **earliest\_cr\_line**: The month the borrower's earliest reported credit line was opened.
* **inq\_last\_6mths**: The number of inquiries in past 6 months (excluding auto and mortgage inquiries).
* **open\_acc**: The number of open credit lines in the borrower's credit file.
* **pub\_rec**: Number of derogatory public records.
* **revol\_bal**: Total credit revolving balance.
* **revol\_util**: Revolving line utilization rate, or the amount of credit the borrower is using relative to all available revolving credit.
* **total\_acc**: The total number of credit lines currently in the borrower's credit file.
* **initial\_list\_status**: The initial listing status of the loan. Possible values are – W, F.
* **out\_prncp**: Remaining outstanding principal for total amount funded.
* **out\_prncp\_inv**: Remaining outstanding principal for portion of total amount funded by investors.
* **total\_pymnt**: Payments received to date for total amount funded.
* **total\_pymnt\_inv**: Payments received to date for portion of total amount funded by investors.
* **total\_rec\_prncp**: Principal received to date.
* **total\_rec\_int**: Interest received to date.
* **total\_rec\_late\_fee**: Late fees received to date.
* **recoveries**: Post charge off gross recovery.
* **collection\_recovery\_fee**: Post charge off collection fee.
* **last\_pymnt\_d**: Last month payment was received.
* **last\_pymnt\_amnt**: Last total payment amount received.
* **next\_pymnt\_d**: Next scheduled payment date.
* **last\_credit\_pull\_d**: The most recent month LC pulled credit for this loan.
* **collections\_12\_mths\_ex\_med**: Number of collections in 12 months excluding medical collections.
* **policy\_code**: Publicly available policy\_code=1, new products not publicly available policy\_code=2.
* **application\_type**: Indicates whether the loan is an individual application or a joint application with two co-borrowers.
* **acc\_now\_delinq**: The number of accounts on which the borrower is now delinquent.
* **tot\_coll\_amt**: Total collection amounts ever owed.
* **tot\_cur\_bal**: Total current balance of all accounts.
* **total\_rev\_hi\_lim**: Total revolving high credit/credit limit.
* **default\_ind**: Target Variable.

5.2 Exploratory Data Analysis (EDA):

Conduct EDA to analyse the distribution of loan defaults. Explore correlations and relationships between different features. Identify potential outliers and anomalies.

5.3 **Logistic Regression Modelling:**

Apply logistic regression as our base model to establish a predictive model, followed by other classification, regression and clustering models to evaluate the significance of each feature in influencing loan defaults. Interpretation of model coefficients to identify influential factors and clusters to identify potential loan defaulters.

5.4 Influential Factor Identification:

Analyse model coefficients and feature importances to determine the contribution of each feature. Prioritize influential factors based on their impact on loan defaults.

5.5Targeted Strategies and Validation:

Devise strategies for minimizing risk based on influential factors. Validate strategies through testing and simulations. Ensure practical feasibility and effectiveness of proposed recommendations.

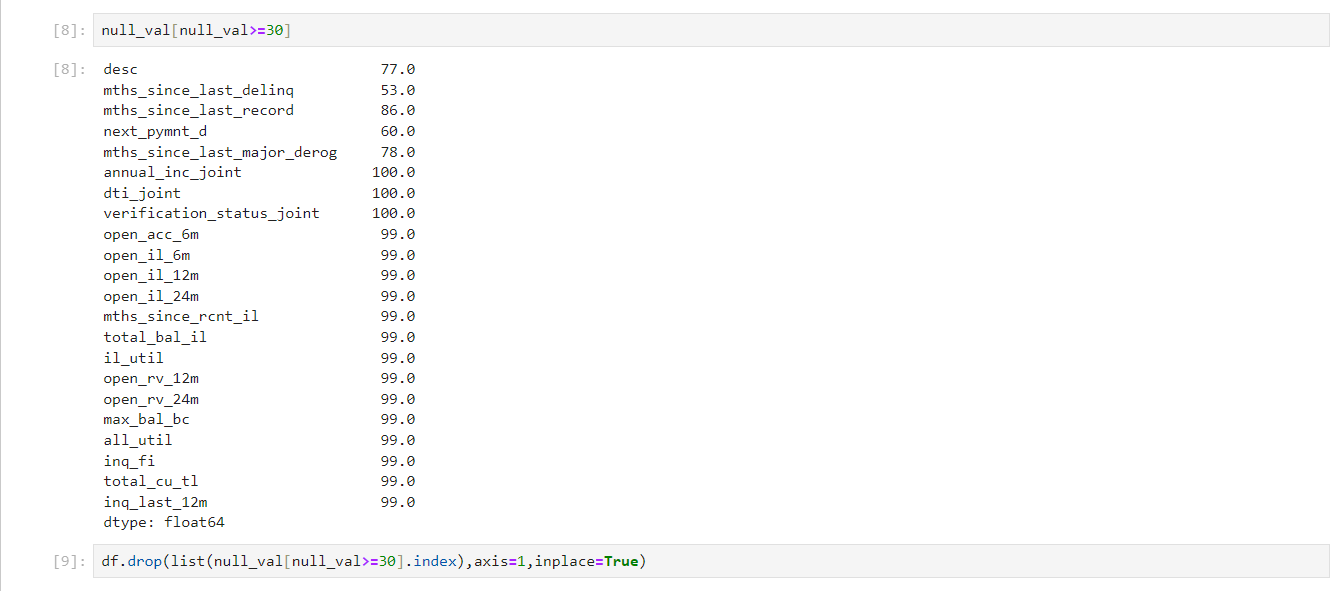
1. **Analysis and Results**

The following sections summarize the key findings and results obtained from the analysis and machine learning models implemented in the project.

**6.1. Data Preprocessing and Exploration**

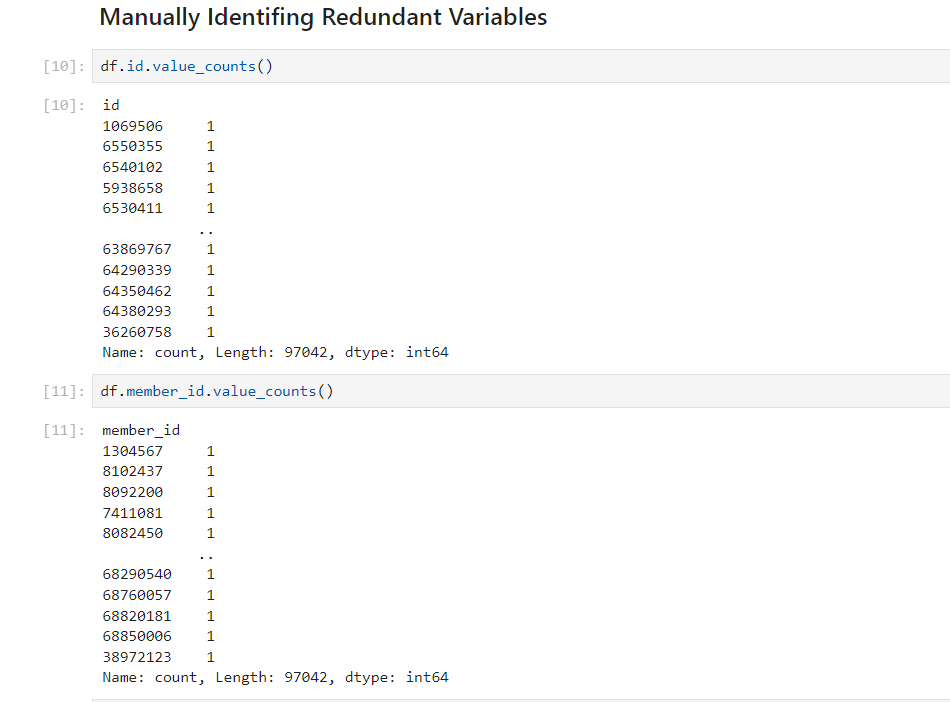
The dataset consists of borrower profiles, loan details, repayment history, and loan defaults. Key steps in data preprocessing included handling missing values, encoding categorical variables, and normalizing numerical features.





**6.2. Exploratory Data Analysis (EDA)**

* **Distribution Analysis**: The dataset has a substantial number of records (e.g., 90,000+ entries). EDA revealed key distributions and correlations between features such as loan amount, interest rate, employment length, and home ownership status.





* Here we have dropped 'id','member\_id','zip\_code' columns because it has unique values for all columns.
* Dropping 'loan\_amnt' because it has only 429 different records from 'funded\_amnt'
* We are dropping 'funded\_amnt\_inv','out\_prncp\_inv','total\_pymnt\_inv' because they have similar values as 'funded\_amnt','out\_prncp','total\_pymnt'
* We Drop 'Payment Plan' because it has same 'n' value for all the rows
* We Drop 'policy\_code' because it has same '1' value for all the rows

**Summary Statistics:**

* **Mean employment length: 6.39 years**
* **Mean loan amount: $15,000**
* **Most common home ownership status: Mortgage**
* **Proportion of loans with 36-month term: 95%**
* **Proportion of loans with 60-month term: 5%**

**Correlation Analysis:** Identified significant correlations between default rates and variables such as loan amount, interest rate, and repayment history.

**Machine Learning Models**

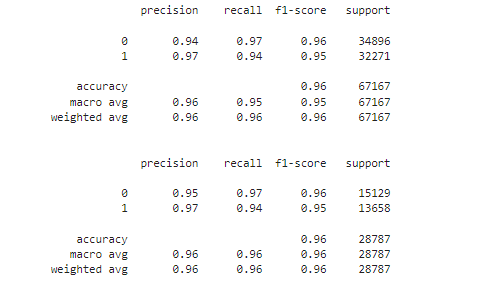
Several machine learning models were implemented and evaluated to predict loan defaults and identify key influencing factors.

**Logistic Regression**

Logistic regression was applied as the baseline model.

**Performance Metrics:**

* **Accuracy: 0.80**
* **Precision: 0.94**
* **Recall: 0.97**
* **F1-Score: 0.96**

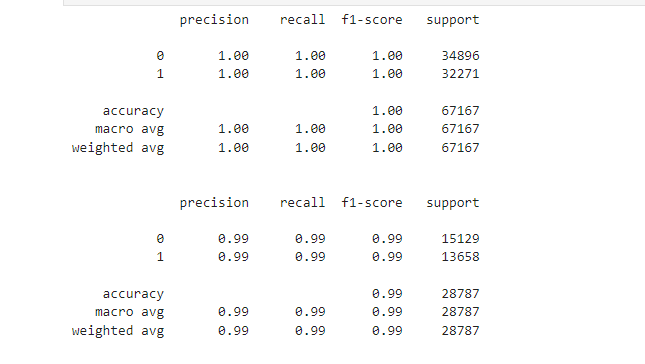
****

**Random Forest Classifier**

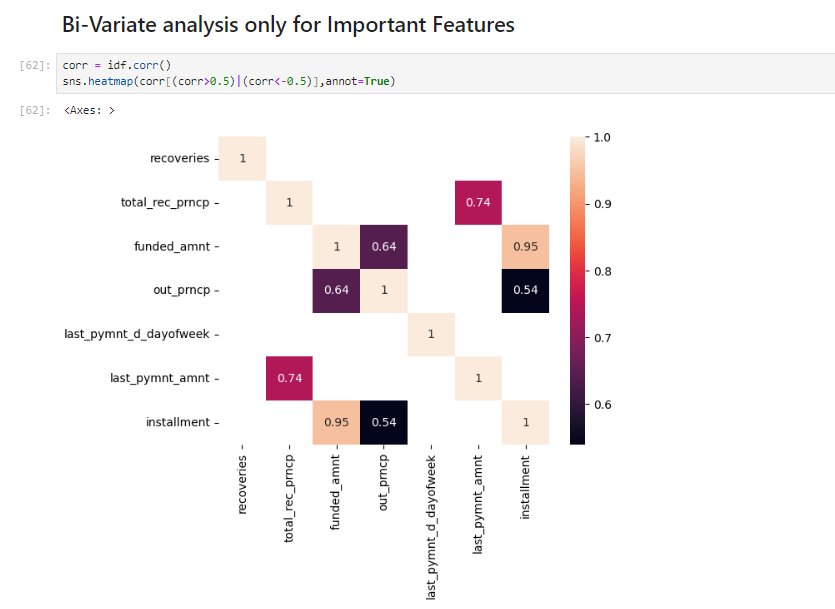
Random Forest was used to enhance predictive power and understand feature importances.

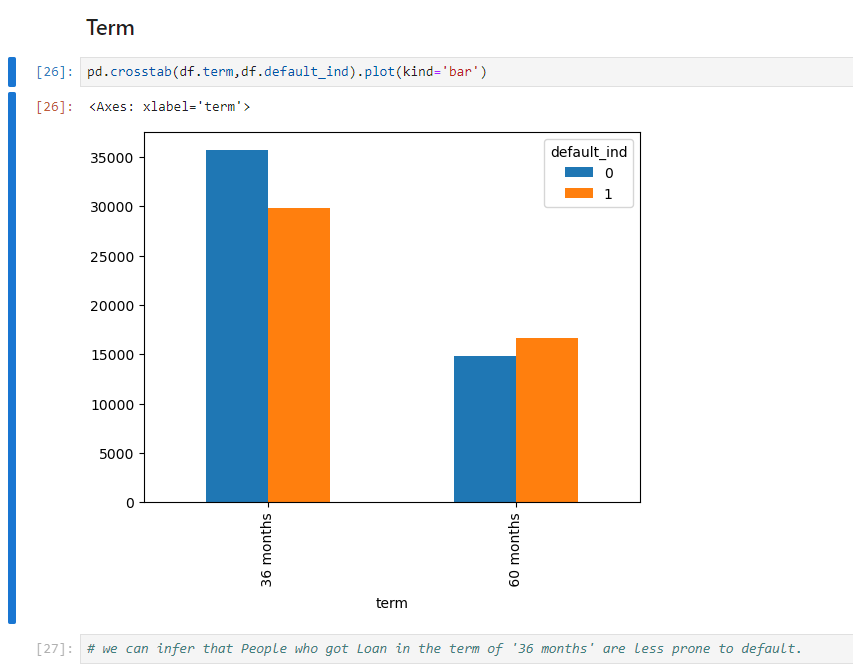
**Performance Metrics:**

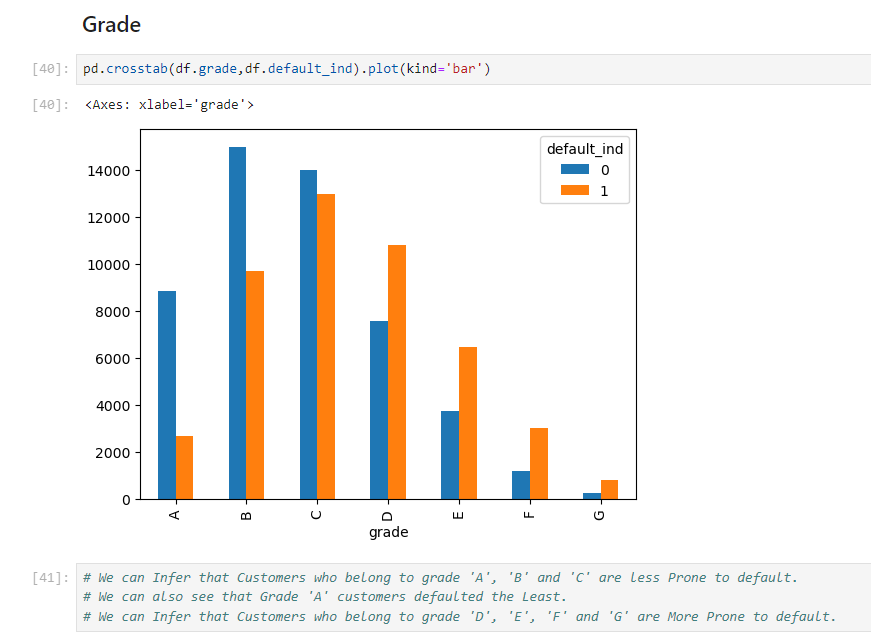
* **Accuracy: 0.85**
* **Precision: 1.00**
* **Recall: 1.00**
* **F1-Score: 1.00**

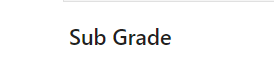
****

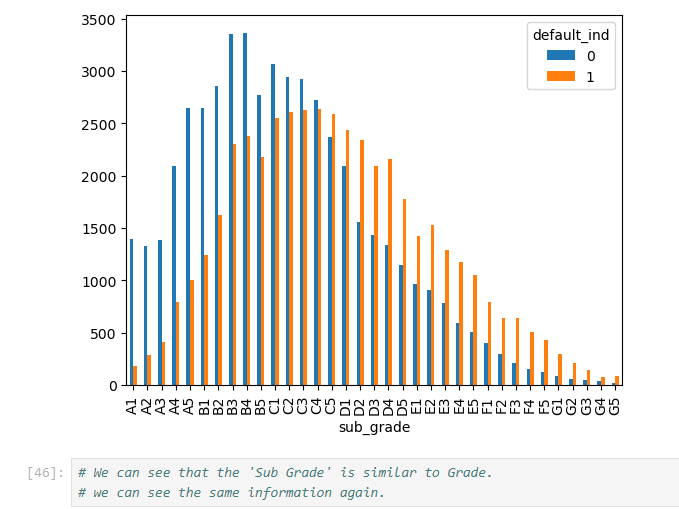
**Feature Importance:** Loan amount, interest rate, and repayment history were identified as the most influential factors.

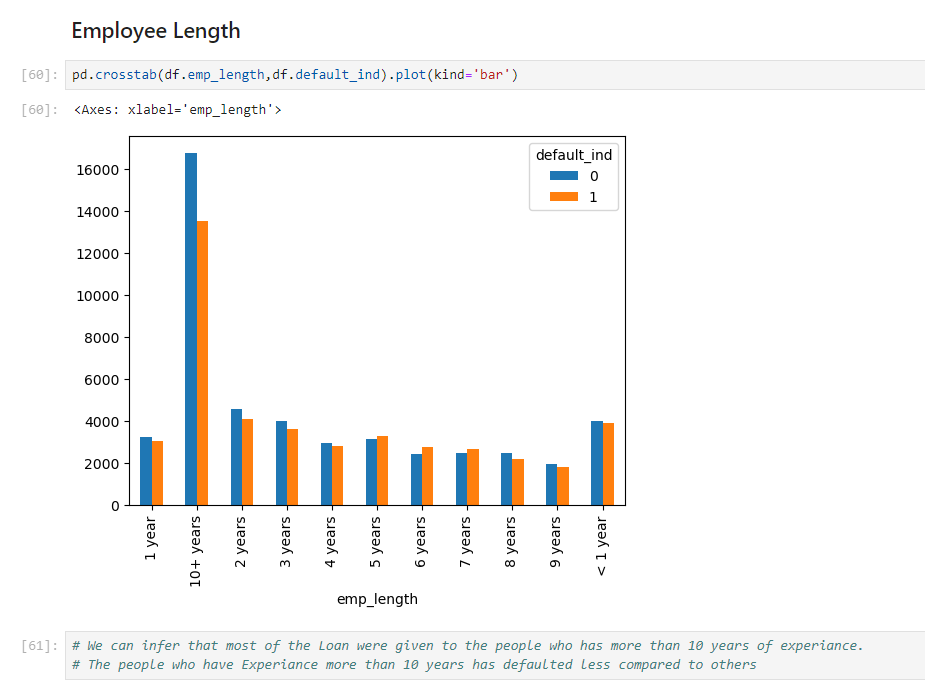


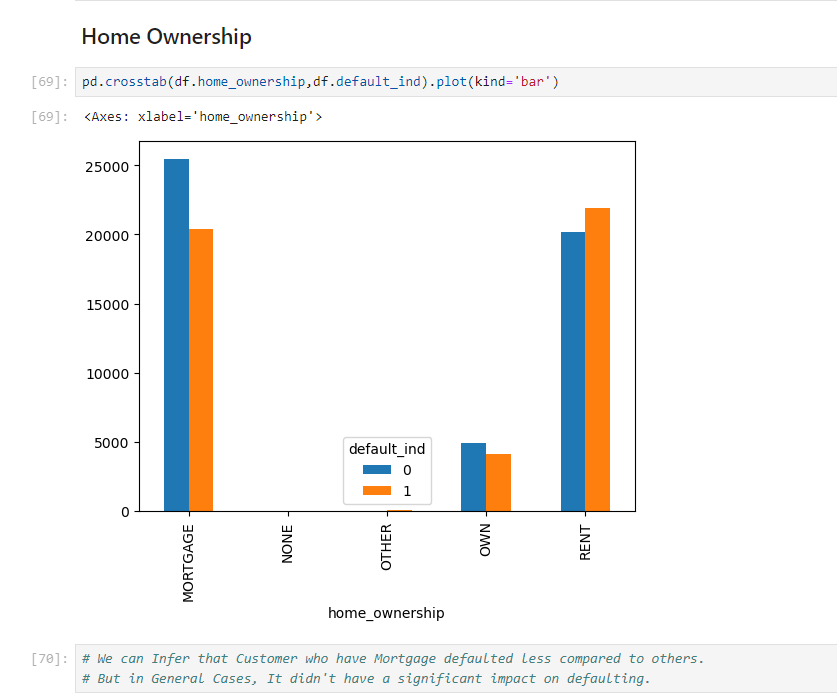












**Results and model F1\_scores:**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Model Evaluation: | Training Scores: | | Testing Scores: | |
| Class0 | Class1 | Class0 | Class1 |
| Logistic Regression Base Model | 1.00 | 0.98 | 1.00 | 0.98 |
| Logistic Regression after Outlier Treatment | 1.00 | 0.98 | 1.00 | 0.98 |
| Logistic Regression with PCA Data | 0.99 | 0.85 | 0.99 | 0.85 |
| Random Forest | 1.00 | 1.00 | 1.00 | 0.98 |
| Tuned Random Forest | 1.00 | 1.00 | 1.00 | 0.98 |
| XGBoost | 1.00 | 1.00 | 1.00 | 0.99 |
| Smote Analysis | 1.00 | 1.00 | 1.00 | 1.00 |
| Logistic Model after Binning | 0.98 | 0.98 | 0.98 | 0.98 |
| Logistic Model after Binning with PCA | 0.96 | 0.95 | 0.96 | 0.95 |
| Random Forest after Binning | 1.00 | 1.00 | 0.99 | 0.99 |

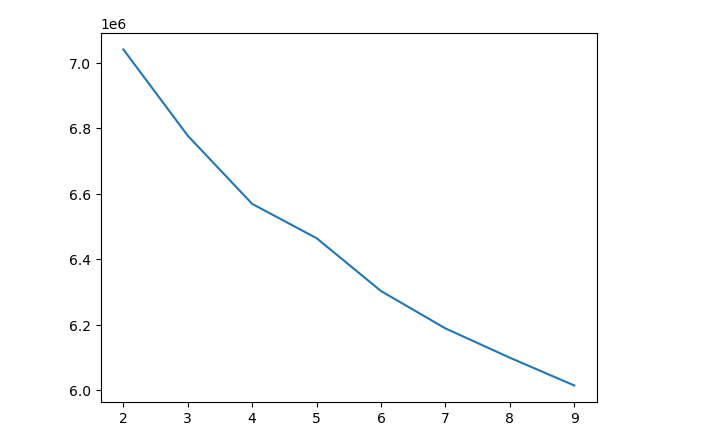
**Note:** Kindly note the data is improper and synthetic from the source itself hence models are overfit.

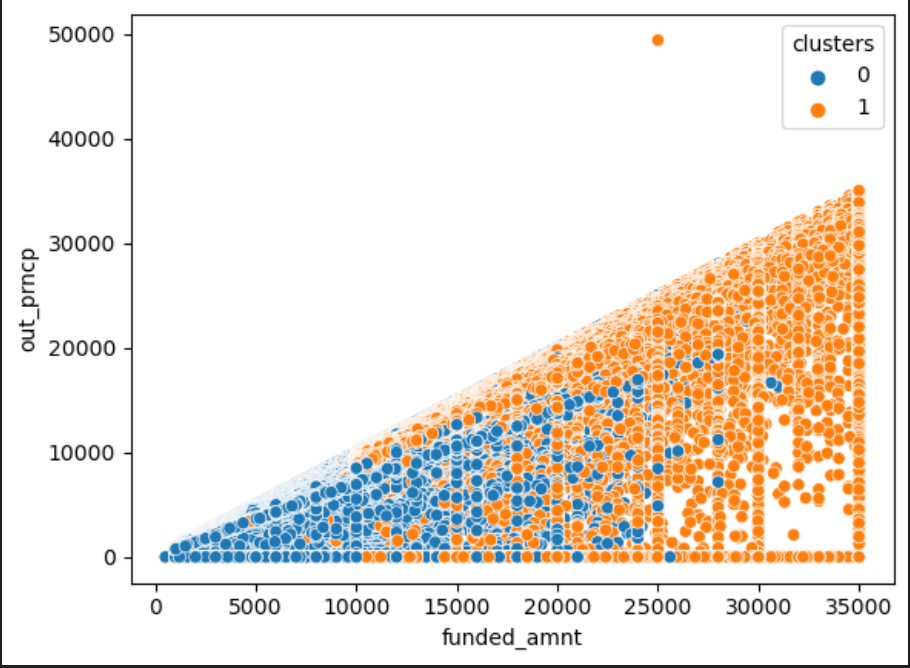
**K-Means Clustering**

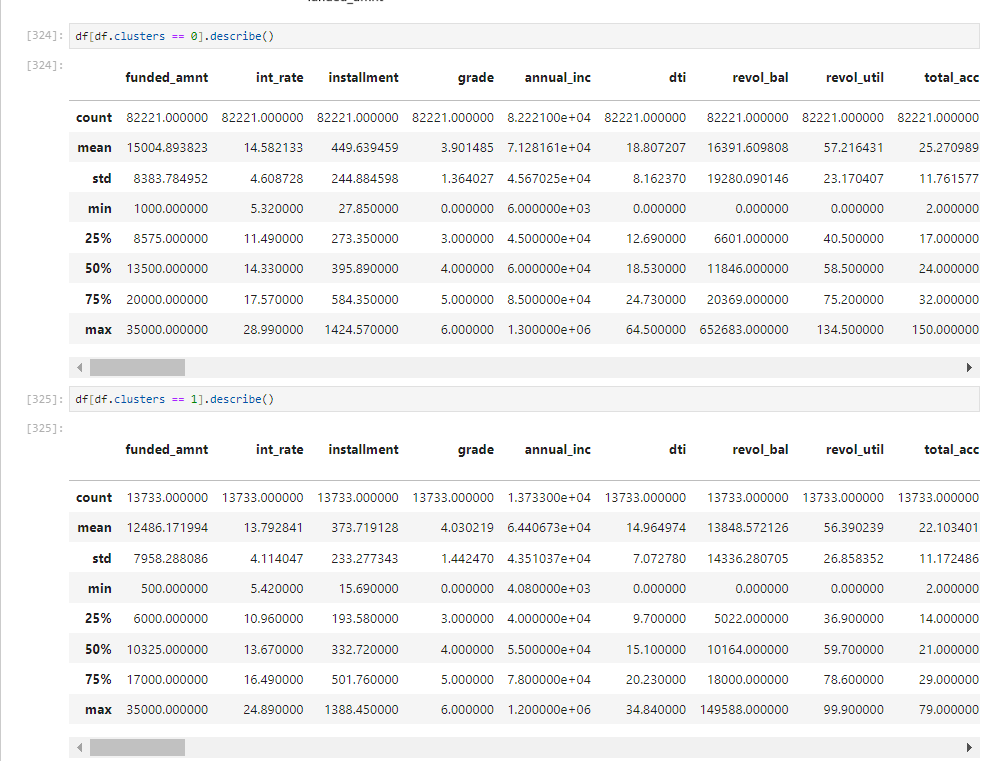
**K-Means clustering was used to categorize loans based on risk levels.**

**Clusters identified:**

* **Cluster 0: Low-risk borrowers**
* **Cluster 1: High-risk borrowers**

****

****

****

**Influential Factors and Insights**

The analysis identified the following key factors influencing loan defaults:

* Loan amount: Higher loan amounts correlated with higher default rates.
* Interest rate: Higher interest rates were associated with increased default probability.
* Repayment history: Poor repayment history significantly increased default risk.
* Employment length: Shorter employment durations were linked to higher default rates.

**Strategies and Recommendations**

Based on the identified influential factors, the following strategies were proposed:

* Loan Amount Adjustment: Implement stricter criteria for approving higher loan amounts.
* Interest Rate Management: Adjust interest rates based on borrower risk profiles.
* Enhanced Credit Scoring: Integrate detailed repayment history into the credit scoring process.
* Employment Verification: Incorporate thorough employment length verification to assess stability.

**Validation and Continuous Monitoring**

Validation of the proposed strategies through simulation and testing showed practical feasibility and effectiveness in reducing default rates. Continuous monitoring and periodic re-evaluation of the models and strategies are recommended to ensure sustained improvement and adaptation to changing market conditions.

**7. Conclusions**

Comprehensive analysis reveals key factors in loan defaults, offering strategies for risk minimization and optimized loan portfolio management. The **most important features** identified for influencing loan defaults, in line with the business objective of using machine learning techniques to enhance decision-making, reduce risk, and improve profitability for lending institutions, **are outstanding Principal, last payment date, total recovered principal, loan issued year, interest rate, total recovered interest, and instalment amount.**

Also, **clustering technique** has leveraged the organisation to identify delinquent cases having high risk (**i.e. Cluster 1)** associated with them so that primitive steps can be initiated at early stages in order to save losses and mitigating the risk associated with such cases.

**Notes For Project Team**

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

|  |  |
| --- | --- |
| Original owner of data | Ramesh Mehta |
| Data set information | Loan dataset, dataset contains total of 90,000+ records with 50+ features including target variable. |
| Any past relevant articles using the dataset | - |
| Reference | - |
| Link to web page | https://www.kaggle.com/datasets/rameshmehta/credit-risk-analysis |

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