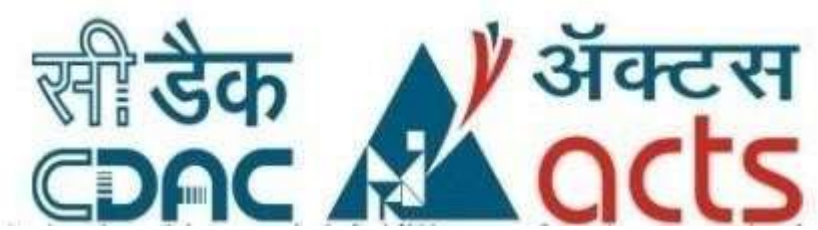
Project Report

On

CREDIT WORTHNESS FORECASTER



Submitted

In partial fulfilment

For the award of the Degree of

PG-Diploma in Big Data Analytics

(C-DAC, ACTS (Pune))

Guided By:

Mr. Milind Kapase

Submitted By:

Neeraj Singh (240340125030)

Nikhil Khaladkar (240340125031)

Sanjay Nannaware (240340125038)

Shreeya Gondhalekar (240340125043)

Vaibhav Phatangare (240340125054)

Centre for Development of Advanced Computing

(C-DAC), ACTS (Pune- 411008)

Acknowledgement

This is to acknowledge our indebtedness to our Project Guide, **Mr. Milind Kapase ,** C-DAC ACTS, Pune for his constant guidance and helpful suggestion for preparing this project **Credit Worthiness Forecaster Model .** We express our deep gratitude towards him for his inspiration, personal involvement, constructive criticism that he provided us along with technical guidance during the course of this project.

We take this opportunity to thank Head of the department **Mr. Gaur Sunder** for providing us such a great infrastructure and environment for our overall development.

We express sincere thanks to **Mrs. Namrata Ailawar (Process Owner)** for their kind cooperation and extendible support towards the completion of our project.

It is our great pleasure in expressing sincere and deep gratitude towards **Mrs. Risha P R (Program Head)** and **Ms. Pratiksha Gacche (Course Coordinator, PG-DBDA)** for their valuable guidance and constant support throughout this work and help to pursue additional studies.

Also, our warm thanks to **C-DAC ACTS Pune**, which provided us this opportunity to carry out, this prestigious Project and enhance our learning in various technical fields.

Submitted By:

Neeraj Singh (240340125030)

Nikhil Khaladkar (240340125031)

Sanjay Nannaware (240340125038)

Shreeya Gondhalekar (240340125043)

Vaibhav Phatangare(240340125054)

ABSTRACT

Currently, lenders face challenges in effectively predicting which individuals are at higher risk of defaulting on their loans. Inaccurate predictions can lead to significant financial losses, either by extending credit to high-risk individuals or by denying credit to individuals who are actually creditworthy. So our project consists of machine-learning model that can accurately predict the likelihood of an individual defaulting on a loan based on their historical loan repayment behaviour and transactional activities.The highest accuracy achieved is 89% with the XGBoost model. The project is deployed as an interactive web application using Streamlit, allowing users to input loan details and receive real-time predictions. Additionally, data visualization is performed using Tableau and Power BI, creating reports and dashboards that provide valuable insights.

Table of Contents

Introduction

In the financial industry, managing credit risk is critical to the sustainability and profitability of lending institutions. One of the most significant challenges faced by these institutions is the risk of loan defaults, where borrowers fail to meet their repayment obligations. Loan defaults can lead to substantial financial losses, not only for lenders but also for the broader economy, affecting credit availability and interest rates.

To address this challenge, predicting loan defaults before they occur has become a priority for financial institutions. By leveraging data-driven approaches and machine learning techniques, it is possible to assess the likelihood of default at the time of loan application, enabling lenders to make more informed decisions. This project aims to develop a predictive model that accurately identifies potential loan defaulters based on historical data and a range of borrower characteristics.

Aim :

The goal of this project is to build a robust loan default prediction model that can help financial institutions minimize risks and optimize their loan portfolios. By analyzing factors such as credit history, income levels, loan amounts, and other relevant variables, the model seeks to provide early warnings and allow lenders to take appropriate actions to mitigate potential losses.

Problem Statement :

Currently, lenders face challenges in effectively predicting which individuals are at higher risk of defaulting on their loans. Inaccurate predictions can lead to significant financial losses, either by extending credit to high-risk individuals or by denying credit to individuals who are actually creditworthy. The goal is to create a robust machine-learning model that can accurately predict the likelihood of an individual defaulting on a loan based on their historical loan repayment behavior and transactional activities.

Objective :

Use EDA to comprehend how loan and consumer characteristics affect the tendency of default.

Constraints :

• When an individual requests for a loan, the organisation may make one of two decisions:
• Loan accepted: Three situations are outlined below in the event that the company authorises the loan:
• Fully paid: The loan (principal and interest rate) has been paid in full by the applicant.
• Current: The loan has not yet reached the end of its tenure; the applicant is currently paying the installments. There is no designation of "defaulted" for these candidates.

• Charged-off: The applicant has defaulted on the loan because they have not made the required monthly payments over an extended period of time.
• Loan denied: The loan had been rejected by the company (because to the candidate does not meet their requirements, among other reasons). Because the loan was denied, the applicants' transaction history with the company is non existent, making this data unavailable to the company (and hence in this dataset)

Data Collection

Once the understanding of the objective is over, the next step is to collect the data. Data collection involves the understanding of initial observations of the data to identify the useful subsets from hypotheses of the hidden information. We use the data from Kaggle.

This dataset contains information on loans made through the **Lending Club platform**, a peer-to-peer lending company that connects borrowers with investors. The dataset includes data on loans issued between *2012* and *2016*.

The goal of this dataset is to predict whether a borrower will **fully pay** off their loan or **default**.

Data Summary

• lending_club_loan_two.csv file contains 396031 rows and 27 columns.

• There are two types of attributes Loan Attribute and Customer attributes

Columns:

1. **loan_amnt**: The amount of money requested by the borrower.
2. **term**: The loan's term length, typically in months (e.g., "36 months").
3. **int_rate**: The interest rate on the loan.
4. **installment**: The fixed monthly payment that the borrower has to make.
5. **grade**: The loan's grade as assigned by LendingClub (e.g., A, B, C).
6. **sub_grade**: The finer classification within each grade (e.g., B1, B2).
7. **emp_title**: The job title of the borrower.

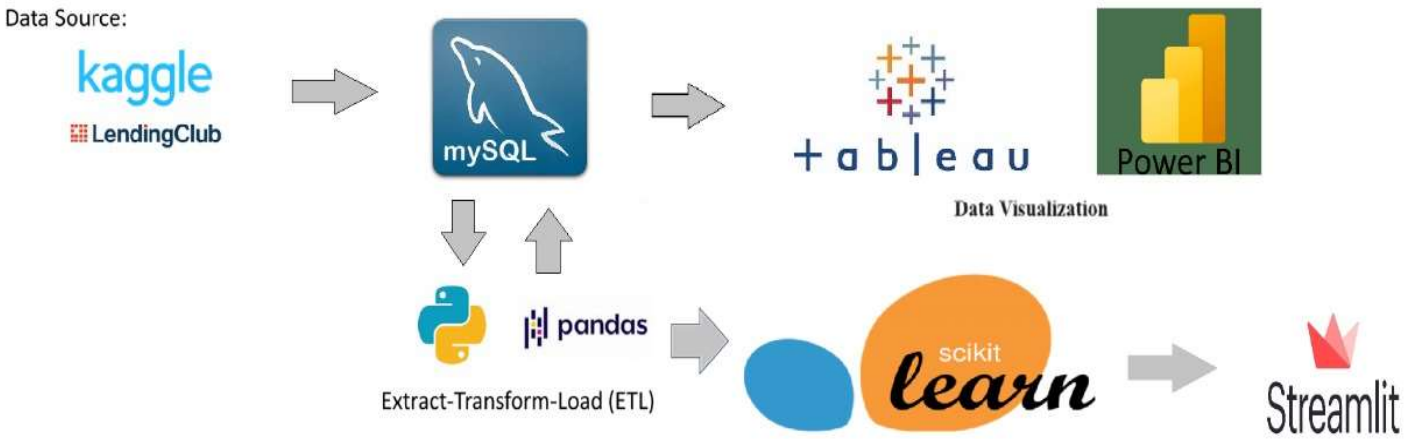
8. **emp_length**: The length of employment in years.
9. **home_ownership**: The status of the borrower's home ownership (e.g., RENT, MORTGAGE).
10. **annual_inc**: The annual income of the borrower.
11. **verification_status**: Indicates whether the borrower's income was verified.
12. **issue_d**: The date when the loan was issued.
13. **loan_status**: The status of the loan (e.g., Fully Paid, Charged Off).
14. **purpose**: The purpose of the loan (e.g., debt consolidation, home improvement).
15. **title**: The title or description provided by the borrower for the loan.
16. **dti**: Debt-to-income ratio of the borrower.
17. **earliest_cr_line**: The date when the borrower's earliest reported credit line was opened.
18. **open_acc**: The number of open credit lines in the borrower's credit file.
19. **pub_rec**: Number of derogatory public records.
20. **revol_bal**: The total credit revolving balance.
21. **revol_util**: The revolving line utilization rate as a percentage.
22. **total_acc**: The total number of credit lines currently in the borrower's credit file.
23. **initial_list_status**: The initial listing status of the loan (e.g., "f", "w").
24. **application_type**: Indicates whether the loan is an individual or joint application.
25. **mort_acc**: The number of mortgage accounts.
26. **pub_rec_bankruptcies**: The number of public record bankruptcies.
27. **address**: The borrower's address.

Target Variable:

The loan_status column is the dataset's target variable. This column shows if the borrower has defaulted or has paid off their loan in full.

A value of "Default" denotes that the borrower has failed to make loan payments, a value of "Fully Paid" indicates that the borrower has entirely paid off their debt.

Workflow


Data Source:
kaggle
LendingClub
mySQL
+ a b | e a u
Power BI
Data Visualization
pandas
Extract-Transform-Load (ETL)
scikit
learn
Streamlit

Methodology and Techniques

Project Overview:

EDA : Analysing and understanding the data is the task of the exploratory data analysis step. It involves comprehending the relationship between the features, spotting outliers, and visualising the data.
Data cleaning : it is the process of dealing with missing numbers, removing outliers, and eliminating features that are not essential to our research.
Categorical to Numerical: Using various encoding strategies, we transform categorical data into numerical ones.
Feature engineering: we can enhance the performance of our model by generating new features from preexisting ones.
Model Training: Training models with different techniques, including Random Forest Classifier (RFC), XGBoost, Logistic Regression, and Artificial Neural Networks (ANN).
Front-end Development: We use Streamlit, a Python package for developing interactive web applications, to create a user-friendly front-end.

STEPS:

To conduct an effective Exploratory Data Analysis (EDA) on the given dataset, the following operations should be performed:

1. Data Understanding and Initial Exploration

- **Preview the Data:** Use head(), info(), and describe() to get an initial understanding of the data types, non-null counts, and summary statistics.
- **Check for Missing Values:** Use isnull().sum() to identify columns with missing values and their proportions.

2. **Data Cleaning**

For the provided dataset, the following data cleaning operations should be performed to ensure that it is ready for analysis and model building:

A. Handling Missing Values

- Identify Missing Values:
    - Use data.isnull().sum() to identify the columns with missing values.
- Imputation for Numerical Columns:

- mort_acc: Impute missing values with the median or mean, as this field is numerical and might be skewed.
- pub_rec_bankruptcies: Since the number of missing values is small, you can either drop these rows or impute them with the mode or median.
- revol_util: Impute missing values with the median or mean.

B. **Correcting Data Types**

- Convert term to Integer:
  - The term column currently contains string values like "36 months" or "60 months". Convert these to integer values (36 or 60) using string manipulation.
- Convert Date Columns to DateTime Format:
  - Convert issue_d and earliest_cr_line from string format to a datetime format. This will allow you to perform time-based calculations and analysis.
- Correct Numerical Columns with String Values:
  - Columns like int_rate might have percentage signs (e.g., "13.56%"). Convert these to float by removing the percentage sign and dividing by 100.

C. **Removing or Transforming Unnecessary Columns**

- Drop Irrelevant Columns:
  - Drop columns that do not add value to the analysis or model, such as url or desc if present, which may contain unstructured text data not useful for the current analysis.
- Transform Derived Features:
  - Create new features if necessary (e.g., loan_income_ratio, credit_age) based on existing columns.

D. Data Conversions vs. Derived Columns, EDA

- **Data Conversions:**
  - Convert term to numeric (36 or 60).
  - Parse date columns such as issue_d and earliest_cr_line to datetime format.
- **Derived Columns:**
  - Create a new column loan_income_ratio as loan_amnt divided by annual_inc.
  - Create a loan_duration derived from issue_d and earliest_cr_line.
- **Exploratory Data Analysis (EDA):**

- Perform EDA to understand the distribution and relationship between variables, especially focusing on loan_status (whether the loan was paid off, defaulted, etc.).

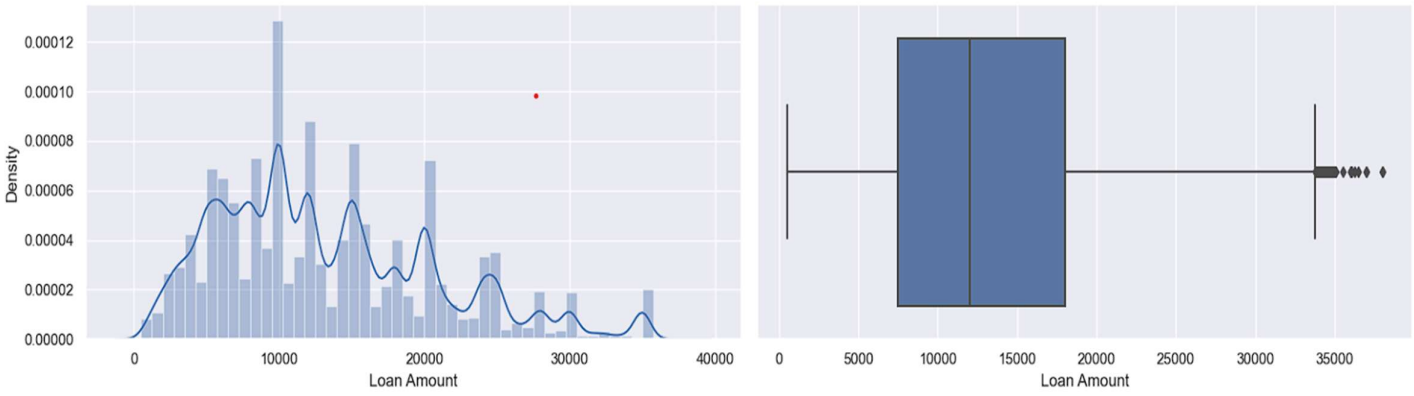
E. Dropping/Imputing Rows

- **Dropping Rows:**
  - Rows with missing values in crucial columns like loan_amnt, int_rate, or loan_status should be dropped if they are very few.
- **Imputing Rows:**
  - For columns like emp_title, emp_length, and mort_acc, imputation can be done using mean, median, or a more complex method.

3. Univariate Analysis

- **Continuous Variables:**
  - Analyze distributions of variables like loan_amnt, int_rate, installment, and annual_inc.
  - Visualize using histograms or KDE plots to understand skewness, central tendency, and spread.

- **Loan Amount**



Observations:

mean 13185.748989

min 500.000000

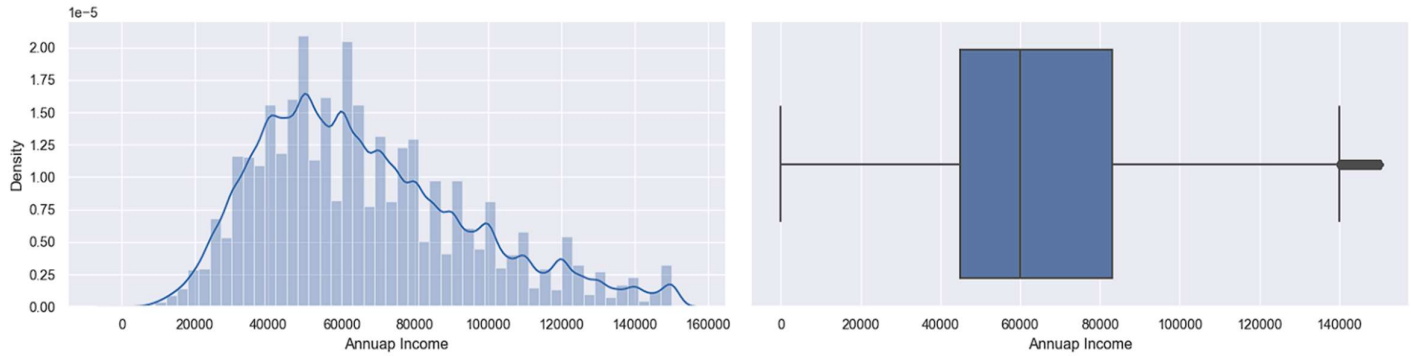
25% 7475.000000

50% 12000.000000

max 38000.000000

- Most of the loan amount applied was in the range of 5k-14k.
- Max Loan amount applied was 38k.

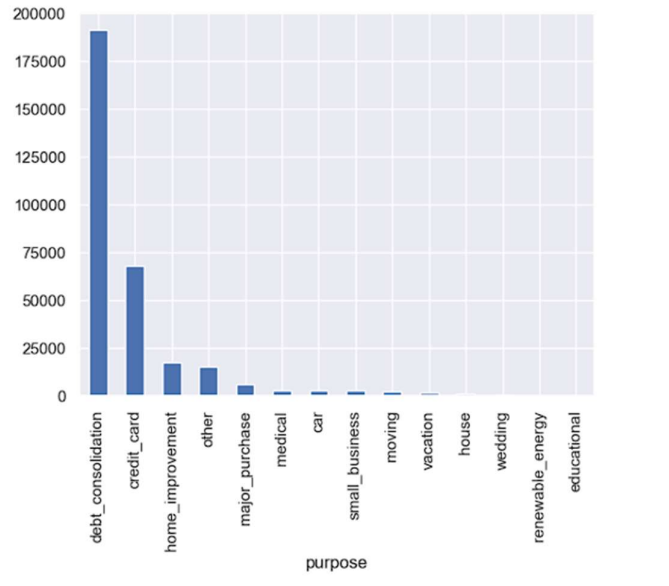
Annual Income



- The Annual income of most if applicants lies between 40k-75k.
- Average Annual Income is : 59883.0
- Average Annual Income is : 67790.0

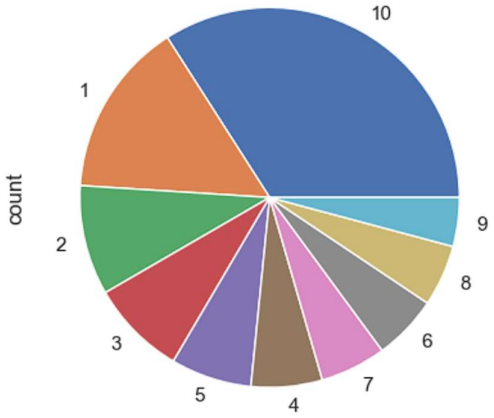
Categorical Variables:

- Analyze categorical variables like grade, home_ownership, and loan_status using bar charts to understand the frequency distribution.



Observations:

Most of the loans are taken for the purpose of Debt_consolidation.

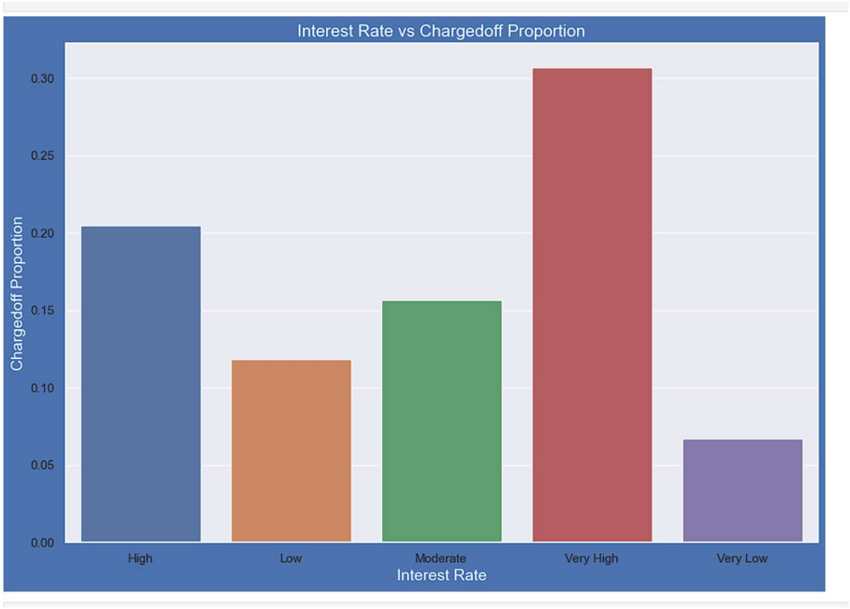


Observations :

- Most of the loan applicants are for debt_consolidations.
- Most of the applications are having 10+ yrs of Exp

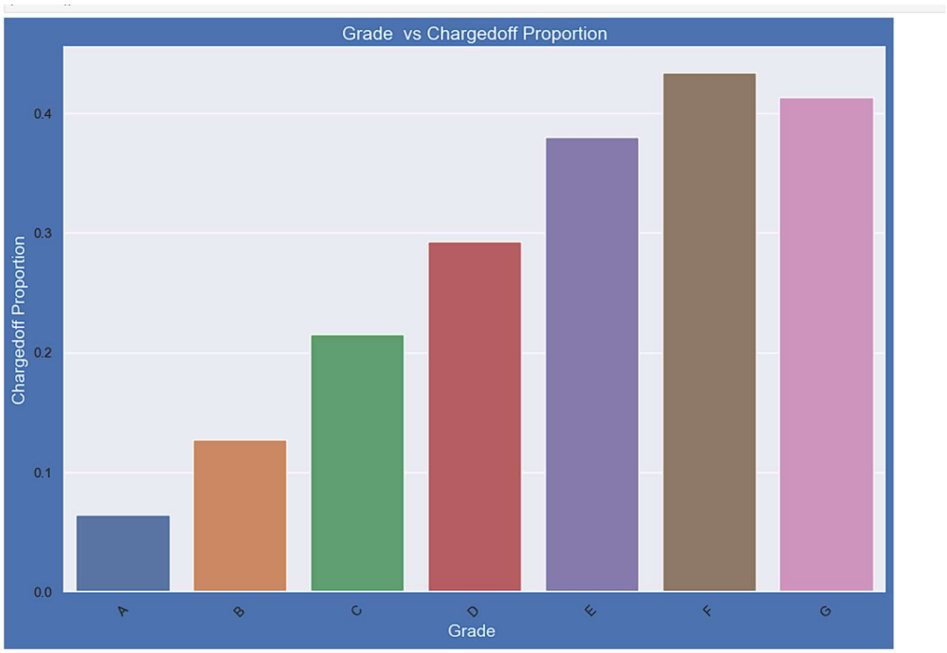
4. Bivariate Analysis

- **Relationships:**
    - Explore relationships between pairs of variables, such as int_rate vs. loan_amnt, loan_status vs. grade, and annual_inc vs. dti.
- **Visualization:**
    - Use scatter plots, box plots, and violin plots to visualize these relationships.



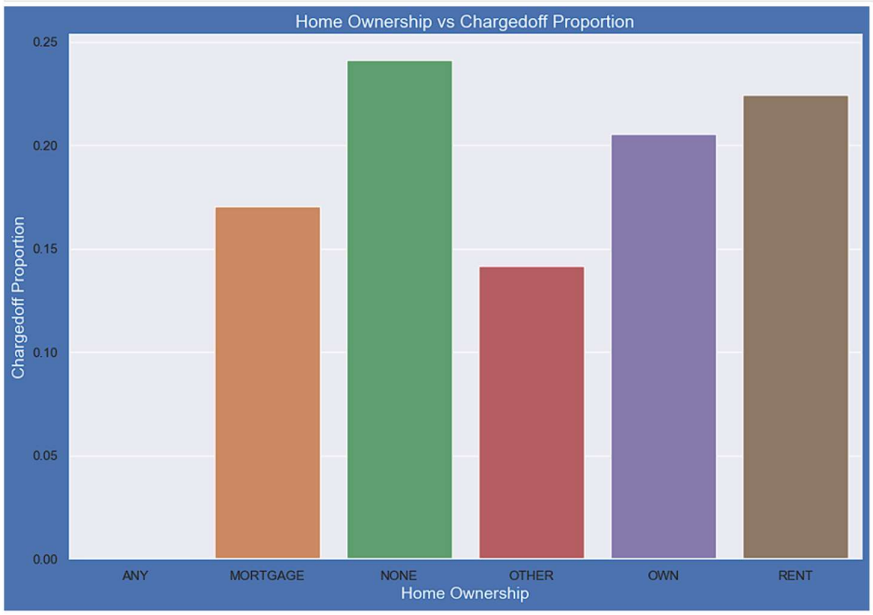
Observations:

- Interest rate less than 10% or very low has very less chances of charged off. Intrest rates are starting from minimum 5 %.
- Interest rate more than 16% or very high has good chnaces of charged off as compared to other category intrest rates.
- Charged off proportion is increasing with higher intrest rates.



Observations:

- 1.The Loan applicants with loan Grade G is having highest Loan Defaults.
- 2. The Loan applicants with Grade A is having lowest Loan Defaults.



Observations:

- Those who are not owning the home is having high chances of loan defaulter.
- From the graph even shows high chances of charged off. Proportions, but data available is very limited compared to other points

5**. Handling Outliers**

- Detect Outliers:
    - Identify outliers in key numerical columns like loan_amnt, annual_inc, int_rate, and dti.
- Treatment of Outliers:
    - Use techniques like capping, transformation (e.g., log transformation), or removing extreme outliers to prevent them from skewing the analysis.

6. Correlations

- **Correlation Matrix:**
  - Calculate and visualize the correlation matrix to identify strong linear relationships between numerical variables.
  - Focus on how loan_amnt, int_rate, and annual_inc are related to other variables.
- **Important Findings:**
  - Investigate highly correlated pairs to reduce multicollinearity in modeling.

7. Feature Engineering

- **Derived Columns:**
  - **Loan-to-Income Ratio:** Create a new feature as loan_amnt divided by annual_inc.
  - **Credit Age:** Calculate the number of months between earliest_cr_line and the issue_d.
  - **Debt-to-Income Ratio (Enhanced):** Create an enhanced dti by including other factors like revol_bal.
- **Categorical Encoding:**
  - **Label Encoding/One-Hot Encoding:** Encode categorical variables like grade, sub_grade, and home_ownership for use in machine learning models.

8. Target Variable Analysis

- **Analyze loan_status:** Since loan_status is the target variable, analyze its distribution and relationship with other features. Create binary flags if necessary for predicting defaults.

9. Dimensionality Reduction (Optional)

- **PCA:** Perform Principal Component Analysis (PCA) to reduce dimensionality if there are too many correlated features.
- **Feature Selection:** Use techniques like SelectKBest or feature importance from tree-based models to select the most relevant features.

10. Conclusion from EDA

- Summarize the key findings from EDA, including important correlations, relationships, outlier treatment, and any potential feature engineering steps that would improve model performance.

These EDA operations will provide deep insights into the data and will prepare the dataset for building a robust machine learning model for creditworthiness prediction.

11. Visualization of Key Insights

- Loan Status Analysis**:** Visualized the distribution of loan_status and explored how other variables affect the likelihood of default.
- Correlation Heatmap**:** Visualized the correlations among all numerical features to detect potential multicollinearity.
- Data visualization is performed using Tableau and Power BI, creating reports and dashboards that provide valuable insights.

Model Building

To achieve good accuracy with a machine learning model on the dataset you provided, a series of well-structured machine learning operations should be performed. Here's a step-by-step outline:

1. Data Preprocessing

- **Feature Selection:**
    - Identify and select the most relevant features using techniques like correlation analysis, feature importance from tree-based models, or L1/L2 regularization methods. This helps in reducing dimensionality and removing irrelevant or redundant features.
- **Data Splitting:**
    - Split the data into training and testing sets, typically using an 80-20 or 70-30 ratio. Additionally, consider creating a validation set if needed.
- **Feature Scaling:**
    - Scale numerical features using StandardScaler or MinMaxScaler, especially if the model is sensitive to feature scaling (e.g., Logistic Regression, SVM, KNN).
- **Handling Imbalanced Classes:**
    - If the target variable (loan_status or similar) is imbalanced (e.g., more "non-defaults" than "defaults"), use techniques like:
        - **Resampling:** Over-sample the minority class (SMOTE) or under-sample the majority class.
        - **Class Weights:** Use model-specific options to assign higher weights to the minority class.

2. Model Selection

- **Baseline Models:**
    - Start with simple models like Logistic Regression and Decision Trees to establish a baseline performance. This will provide a reference point for more complex models.

- **Advanced Models:**
    - **Random Forest:** A powerful ensemble method that can handle both categorical and numerical data, providing good accuracy and robustness against overfitting.
    - **Gradient Boosting Machines (GBM, XGBoost, Catboost):** Gradient Boosting models like XGBoost or Catboost are highly effective for tabular data and often outperform other models when tuned correctly.
    - **Support Vector Machines (SVM):** Effective in high-dimensional spaces, though more computationally intensive.
    - **Neural Networks:** If the dataset is large and complex, a deep learning model like a feed-forward neural network might be considered.
    - **Stacking Models:** Combine multiple models to create a stronger meta-model (stacking), which often boosts performance.

3. Model Evaluation

- **Cross-Validation:**
    - Use k-fold cross-validation (e.g., 5 or 10 folds) to evaluate model performance and ensure the model generalizes well to unseen data.
- **Metrics to Track:**
    - **Accuracy:** Measures the overall correctness of the model.
    - **Precision, Recall, and F1-Score:** Especially important if the classes are imbalanced. F1-score balances precision and recall, providing a single metric that accounts for both false positives and false negatives.
    - **ROC-AUC:** Measures the trade-off between true positive rate and false positive rate. A higher AUC indicates a better model.
    - **Confusion Matrix:** Provides insights into the types of errors the model is making.

4. Hyperparameter Tuning

- **Grid Search or Random Search:**
    - Used Grid Search or Random Search to find the optimal hyperparameters for our models. This involves testing different combinations of parameters (e.g., n_estimators, max_depth for Random Forest) and selecting the ones that yield the best performance.
- **Bayesian Optimization:** More advanced than Grid or Random Search, this method can efficiently explore the hyperparameter space to find the optimal settings.

5. Feature Engineering

- **Polynomial Features:**
    - Create interaction terms or polynomial features to capture non-linear relationships between variables.
- **Domain-Specific Features:**
    - Incorporate domain knowledge to create new features, such as loan_income_ratio or credit_age, which can significantly improve model accuracy.
- **Encoding Categorical Variables:**
    - Use One-Hot Encoding for nominal categories or Target Encoding for high-cardinality categories.

6. Model Ensembling

- **Bagging:** Combine the predictions of several base models (e.g., multiple decision trees) to reduce variance (e.g., Random Forest).
- **Boosting:** Sequentially build models that correct the errors of the previous ones, such as in Gradient Boosting Machines (GBM, XGBoost).
- **Stacking:** Use the predictions of multiple models as input features for a final meta-model to improve predictive performance.

7. Final Model Testing and Validation

- **Final Evaluation on Test Set:**
    - After tuning and selecting the best model, we evaluate it on the holdout test set to ensure that it performs well on unseen data.
- **Calibration:**
    - If necessary, calibrate the final model's predicted probabilities to better align with actual outcomes, using techniques like Platt Scaling or Isotonic Regression.

8. Deployment and Monitoring

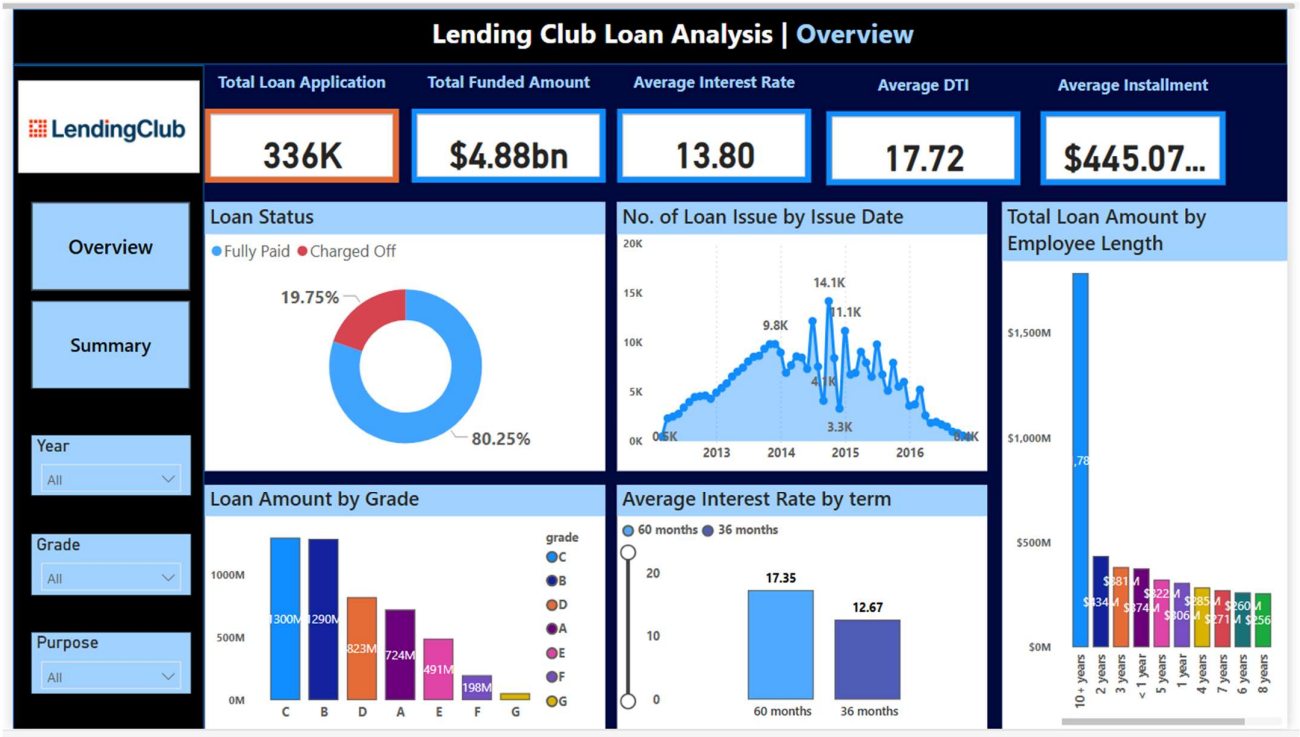
The project is deployed as an interactive web application using Streamlit, allowing users to input loan details and receive real-time predictions.

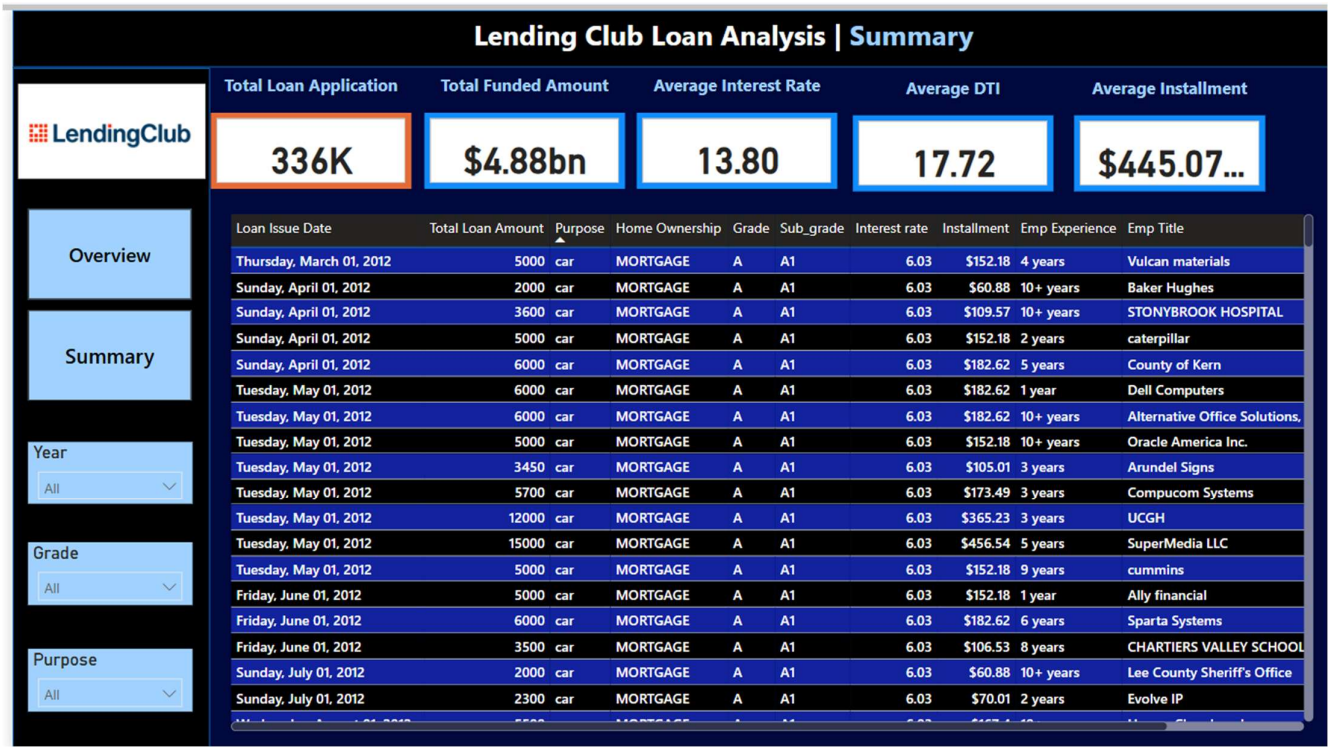
These steps, when carefully followed, will help in building a robust machine learning model with good accuracy, capable of predicting credit worthiness effectively.

Streamlit Interface


Public Records
0
Revolving Balance
1000.00
Revolving Utilization Rate (%)
30.00
Total Accounts
10
Mortgage Accounts
1
Postal Code
0.00
Predict
Congratulations the loan is likely to be approved.

PowerBi Dashboard





Loan Issue Date	Total Loan Amount	Purpose	Home Ownership	Grade	Sub_grade	Interest rate	Installment	Emp Experience	Emp Title
Thursday, March 01, 2012	5000	car	MORTGAGE	A	A1	6.03	$152.18	4 years	Vulcan materials
Sunday, April 01, 2012	2000	car	MORTGAGE	A	A1	6.03	$60.88	10+ years	Baker Hughes
Sunday, April 01, 2012	3600	car	MORTGAGE	A	A1	6.03	$109.57	10+ years	STONYBROOK HOSPITAL
Sunday, April 01, 2012	5000	car	MORTGAGE	A	A1	6.03	$152.18	2 years	caterpillar
Sunday, April 01, 2012	6000	car	MORTGAGE	A	A1	6.03	$182.62	5 years	County of Kern
Tuesday, May 01, 2012	6000	car	MORTGAGE	A	A1	6.03	$182.62	1 year	Dell Computers
Tuesday, May 01, 2012	6000	car	MORTGAGE	A	A1	6.03	$182.62	10+ years	Alternative Office Solutions,
Tuesday, May 01, 2012	5000	car	MORTGAGE	A	A1	6.03	$152.18	10+ years	Oracle America Inc.
Tuesday, May 01, 2012	3450	car	MORTGAGE	A	A1	6.03	$105.01	3 years	Arundel Signs
Tuesday, May 01, 2012	5700	car	MORTGAGE	A	A1	6.03	$173.49	3 years	Compucom Systems
Tuesday, May 01, 2012	12000	car	MORTGAGE	A	A1	6.03	$365.23	3 years	UCGH
Tuesday, May 01, 2012	15000	car	MORTGAGE	A	A1	6.03	$456.54	5 years	SuperMedia LLC
Tuesday, May 01, 2012	5000	car	MORTGAGE	A	A1	6.03	$152.18	9 years	cummins
Friday, June 01, 2012	5000	car	MORTGAGE	A	A1	6.03	$152.18	1 year	Ally financial
Friday, June 01, 2012	6000	car	MORTGAGE	A	A1	6.03	$182.62	6 years	Sparta Systems
Friday, June 01, 2012	3500	car	MORTGAGE	A	A1	6.03	$106.53	8 years	CHARTIERS VALLEY SCHOOL
Sunday, July 01, 2012	2000	car	MORTGAGE	A	A1	6.03	$60.88	10+ years	Lee County Sheriff's Office
Sunday, July 01, 2012	2300	car	MORTGAGE	A	A1	6.03	$70.01	2 years	Evolve IP

Conclusion and Future Scope

The loan defaulter prediction model developed in this project demonstrates the potential to significantly improve the decision-making process for lenders. By accurately identifying individuals or entities at higher risk of defaulting on their loans, financial institutions can take proactive measures to mitigate risk, such as adjusting interest rates, requiring additional collateral, or declining high-risk applications.

The model's performance, as measured by metrics like accuracy, precision, recall, and AUC-ROC, indicates its robustness and reliability. However, it is essential to recognize that the model's effectiveness relies on the quality and quantity of input data. Continuous monitoring and retraining of the model with new data are crucial to maintain its predictive accuracy over time.

Looking forward, the model could be further refined by incorporating additional features, such as macroeconomic indicators, or by exploring more advanced techniques like ensemble learning and deep learning. Additionally, ethical considerations, such as fairness and transparency in the prediction process, should be carefully addressed to ensure that the model's deployment does not inadvertently discriminate against specific groups.

Overall, this project lays a solid foundation for enhancing the loan approval process, reducing financial losses, and contributing to a more resilient financial system.

****