A PROJECT ON

"LOAN ELIGIBILITY PREDICTION "

SUBMITTED IN
PARTIAL FULFILLMENT OF THE REQUIREMENT
FOR THE COURSE OF
DIPLOMA IN BIG DATA ANALYSIS

SUNBEAM INSTITUTE OF INFORMATION TECHNOLOGY, PUNE

Submitted By:

Tejaswini Pethe (84053)
Akshay Chaudhari (84124)

Mr.Nitin Kudale
Centre Coordinator

Mrs.Manisha Hingne
Course Coordinator

SUNBEAM

CERTIFICATE

This is to certify that the project work under the title ‘Walmart Stores Sales Prediction’ is done by Tejaswini Pethe & Akshay Chaudhari in partial fulfillment of the requirement for award of Diploma in Big Data Analysis Course.

Mr. Aniket P
Project Guide

Mrs. Manisha Hingne
Course Coordinator

Date:

ACKNOWLEDGEMENT

A project usually falls short of its expectation unless aided and guided by the right persons at the right time. We avail this opportunity to express our deep sense of gratitude towards Mr. Nitin Kudale (Center Coordinator, SIIT, Pune) and Mrs. Manisha Hingne (Course Coordinator, SIIT ,Pune) and Project Guide Mr. Aniket Panval.

We are deeply indebted and grateful to them for their guidance, encouragement and deep concern for our project. Without their critical evaluation and suggestions at every stage of the project, this project could never have reached its present form.

Last but not the least we thank the entire faculty and the staff members of Sunbeam Institute of Information Technology, Pune for their support.

Tejaswini Pethe
DBDA March 2024 Batch,
SIIT Pune

Akshay Chaudhari
DBDA March 2024
Batch,SIIT Pune

.

TABLE OF CONTENTS

1. Introduction

1.1 Introduction And Objectives:

Prediction of loan approval system based on machine learning approach is a loan approval system from where we can know whether the loan will pass or not. In this system, we take some data from the user like his monthly income, marriage status, loan amount, loan duration, etc. Then the bank will decide according to its parameters whether the client will get the loan or not. So there is a classification system, in this system, a training set is employed to make the model and the classifier may classify the data items into their appropriate class. A test dataset is created that trains the data and gives the appropriate result that, is the client potential and can repay the loan. Prediction of a loan approval system is incredibly helpful for banks and also the clients. This system checks the candidate on his priority basis. Customer can submit his application directly to the bank so the bank will do the whole process, no third party or stockholder will interfere in it. And finally, the bank will decide that the candidate is deserving or not on its priority basis. The only object of this Project is that the deserving candidate gets straight forward and quick results.

1.2 Why this problem needs To be Solved?

The enhancement in the banking sector lots of people are applying for bank loans but the bank has its limited assets which it has to grant to limited people only, so finding out to whom the loan can be granted which will be a safer option for the bank is a typical process. In existing process, they are use RF algorithm in loan approval system. But the efficiency and accuracy was pretty low. Already banks are provide online transaction system, online bank account opening system, etc,. But there is no loan approval system in the banking sector. Then now we create a new system for loan approval. So now we move on to the proposed system.

1.3 Dataset Information

Name: Loan_default.csv

It has 18 columns.
It has 255,347 rows.

Column Information:

- **LoanID**: Unique identifier for each loan (Object type)
- **Age**: Age of the applicant (Float)
- **Income**: Applicant's income (Float)
- **LoanAmount**: Amount of the loan applied for (Float)
- **CreditScore**: Credit score of the applicant (Integer)
- **MonthsEmployed**: Number of months the applicant has been employed (Integer)
- **NumCreditLines**: Number of credit lines (Integer)
- **InterestRate**: Interest rate on the loan (Float)
- **LoanTerm**: Loan term in months (Integer)
- **DTIRatio**: Debt-to-income ratio (Float)
- **Education**: Education level of the applicant (Object)
- **EmploymentType**: Employment type (Object)
- **MaritalStatus**: Marital status (Object)
- **HasMortgage**: Whether the applicant has a mortgage (Object)
- **HasDependents**: Whether the applicant has dependents (Object)
- **LoanPurpose**: Purpose of the loan (Object)
- **HasCoSigner**: Whether the applicant has a co-signer (Object)
- **Default**: Whether the applicant defaulted on the loan (Integer, 0 or 1)

2. Problem Definition and Algorithm:

2.1 Problem Definition

The goal of this project is to develop a machine learning model to accurately predict loan eligibility based on factors like income, credit score, and employment history. This model aims to streamline the loan approval process, reduce risks, and ensure faster, fairer lending decisions, with the potential to also predict the optimal loan amount for approved applicants.

2.2 Algorithm Definition

Random forest: is a Supervised Machine Learning Algorithm that is used widely in Classification and Regression problems. It builds decision trees on different samples and takes their majority vote for classification and average in case of regression.

One of the most important features of the Random Forest Algorithm is that it can handle the data set containing continuous variables as in the case of regression and categorical variables as in the case of classification. It performs better results for classification problems.

Decision Tree: algorithm belongs to the family of supervised learning algorithms. Unlike other supervised learning algorithms, the decision tree algorithm can be used for solving regression and classification problems too.

The goal of using a Decision Tree is to create a training model that can use to predict the class or value of the target variable by learning simple decision rules inferred from prior data(training data).In Decision Trees, for predicting a class label for a record we start from the root of the tree. We compare the values of the root attribute with the record's attribute. On the basis of comparison, we follow the branch corresponding to that value and jump to the next node.

CatBoost: CatBoost (Categorical Boosting) is a machine learning algorithm developed by Yandex that is particularly well-suited for handling categorical data. It is a type of gradient boosting algorithm that is similar to XGBoost and LightGBM but includes several unique features that make it especially powerful for certain types of tasks.

Key Features of CatBoost:

1. **Handling Categorical Features:**

   - CatBoost can directly handle categorical variables without the need for explicit preprocessing (like one-hot encoding or label encoding). It uses a technique called **Ordered Target Encoding** to convert categorical data into numerical form, making it more efficient and effective.

2. **Fast and Efficient:**

   - CatBoost is optimized for speed and can often outperform other boosting algorithms like XGBoost and LightGBM, especially when dealing with large datasets with many categorical features.

3. **Avoiding Overfitting:**

   - The algorithm includes mechanisms to avoid overfitting, such as Ordered Boosting, which helps ensure that the model generalizes well to new data.

4. **Robust to Noisy Data:**

   - CatBoost is known to be less sensitive to hyperparameter tuning and is robust against overfitting, even in cases with noisy data.

5. **Out-of-the-box Accuracy:**

- CatBoost often provides high accuracy with minimal hyperparameter tuning, making it a good choice for quick model building.

3. Experimental Evaluation:

3.1 Methodology:

The proposed model focuses on predicting the credibility of customers for loan repayment by analyzing their details. The input to the model is the customer details collected. On the output from the classifier, decision on whether to approve or reject the customer request can be made. Using different data analytics tools loan prediction and there severity can be forecasted. In this process it is required to train the data using different algorithms and then compare user data with trained data to predict the nature of loan. The training data set is now supplied to machine learning model; on the basis of this data set the model is trained. Every new applicant details filled at the time of application form acts as a test data set. After the operation of testing, 8 model predict whether the new applicant is a fit case for approval of the loan or not based upon the inference it conclude on the basis of the training data sets. By providing real time input on the web app. In our project, Logistic Regression gives high accuracy level compared with other algorithms. Finally, we are predicting the result via data visualization and display the predicted output using web app using flask.

Preprocessing:

In Data Analysis We will Analyze To Find out the below stuff

1. **Missing Values**

Here we will check the percentage of nan values present in each feature

1 -step make the list of features which has missing values

```
features_with_na=[features for features in df.columns if df[features].isnull().sum()>1]
```

2- step print the feature name and the percentage of missing values

```
for feature in features_with_na:
    print(feature, np.round(df[feature].isnull().mean(), 4),  ' % missing values')
```

```
df['Income'].fillna(df['Income'].mean(), inplace=True)

df['Age'].fillna(df['Age'].mode()[0], inplace=True)
df['Education'].fillna(df['Education'].mode()[0],inplace=True)
df['MaritalStatus'].fillna(df['MaritalStatus'].mode()[0],inplace=True)
df['HasMortgage'].fillna(df['HasMortgage'].mode()[0],inplace=True)
df['HasDependents'].fillna(df['HasDependents'].mode()[0],inplace=True)
df['HasDependents'].fillna(df['HasDependents'].mode()[0],inplace=True)

df.dropna(axis=0, inplace=True)
```

2. **All The Numerical Variables**

These are features in your dataset that are represented by numerical values, such as integers or floats (e.g., Age, Income, LoanAmount).

3. **Distribution of the Numerical Variables**

This refers to how the values of numerical variables are spread out, typically visualized using histograms or box plots to understand their range, central tendency, and skewness.

4. **Categorical Variables**

These are features in your dataset that represent categories or labels rather than numerical values (e.g., Education, EmploymentType).

5. **Cardinality of Categorical Variables**

This refers to the number of unique categories or levels within a categorical variable (e.g., a variable "Education" might have categories like "Bachelor's," "Master's," etc.).

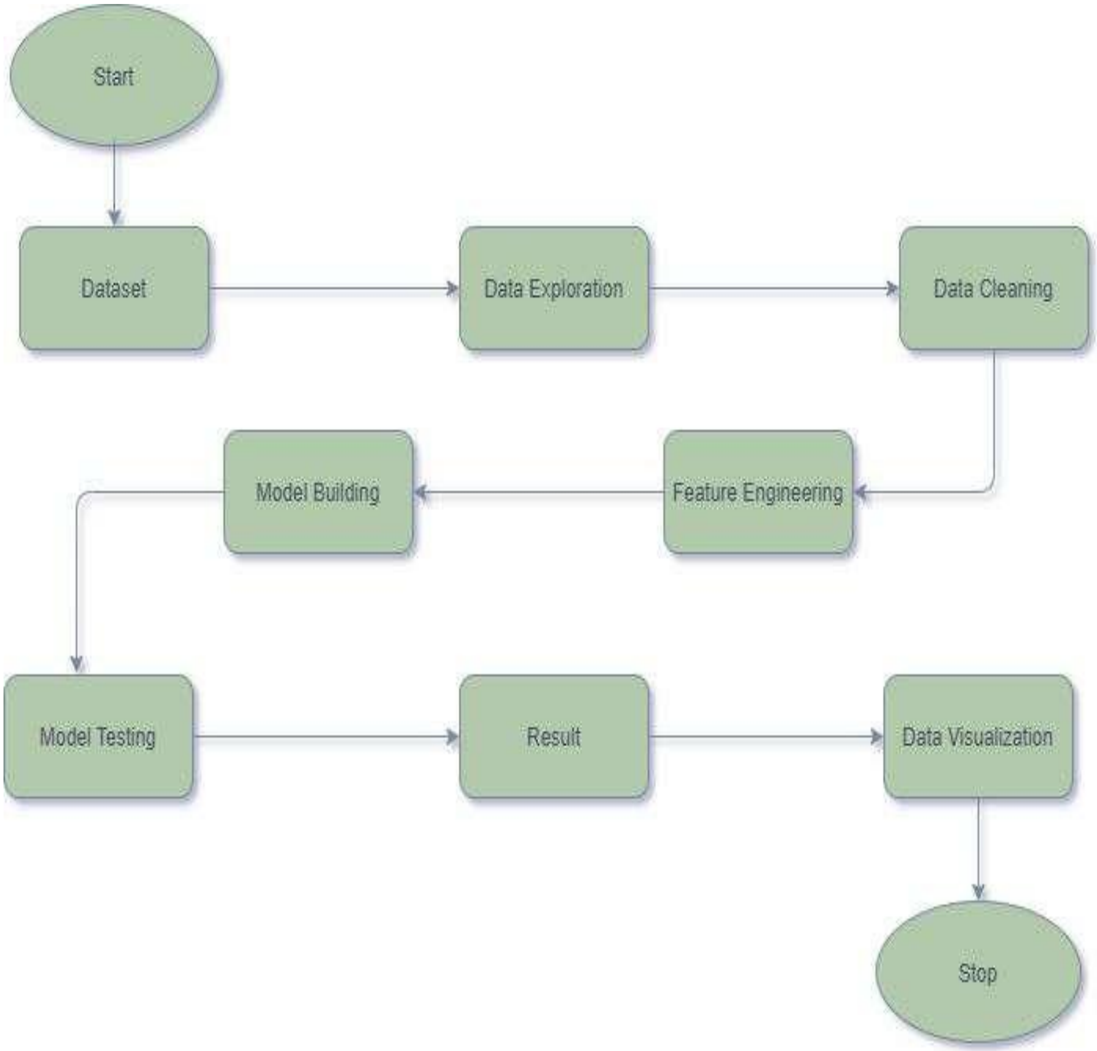
6. **Outliers**

These are data points that are significantly different from the rest of the data, often detected using statistical methods or visualization tools like box plots.

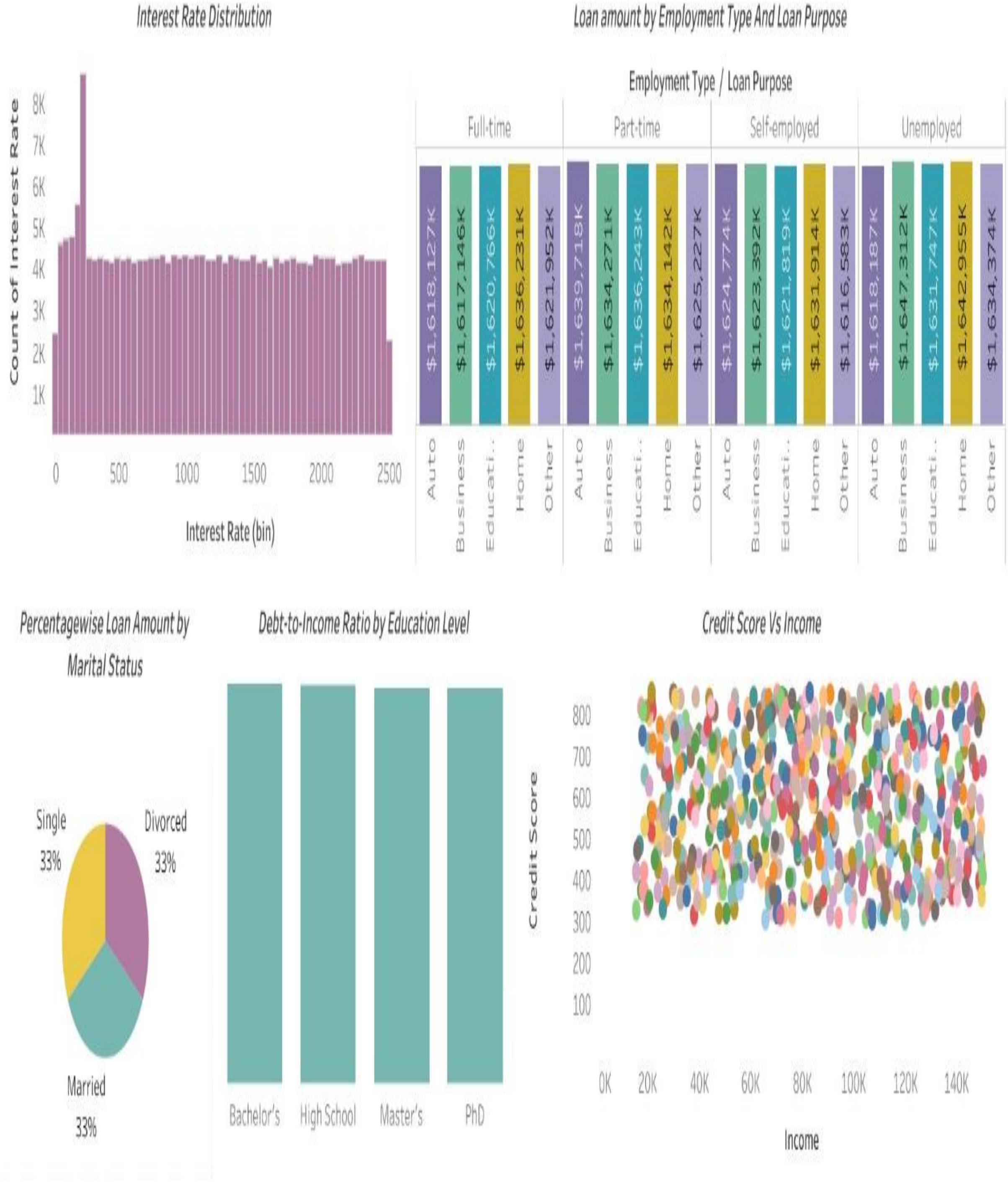
7. **Relationship between independent and dependent feature(SalePrice)**

This involves analyzing how the independent variables (features) in the dataset influence the dependent variable (target), often explored through correlation, regression, or visual plots**.**

Flow Diagram :



3.2 Exploratory Data Analysis



Interest Rate Distribution: The interest rate distribution shows a high concentration around a lower range, with a significant spike at the beginning. This could indicate that most loans have lower interest rates, possibly due to favorable lending conditions or borrower profiles.

Loan Amount by Employment Type and Loan Purpose: Across different employment types (full-time, part-time, self-employed, unemployed), the loan amounts are fairly consistent regardless of the loan purpose (Auto, Business, Education, Home, Other). This consistency might suggest that loan amounts are more influenced by other factors than employment status or the purpose of the loan.Full-time employed individuals seem to have slightly higher loan amounts across all categories compared to other employment types.

Percentagewise Loan Amount by Marital Status: The loan amounts are evenly distributed among Single, Married, and Divorced individuals, each contributing 33%. This equal distribution suggests that marital status might not significantly impact the amount of loan a person takes out.

Debt-to-Income Ratio by Education Level: The debt-to-income ratio appears consistent across different education levels, from Bachelor’s to PhD. This uniformity could indicate that education level alone does not drastically influence the debt-to-income ratio, suggesting other factors, like income stability or financial literacy, may play a role.

Credit Score vs. Income: The scatter plot showing Credit Score vs. Income indicates no clear linear relationship between income and credit score. Individuals with higher incomes do not necessarily have higher credit scores, which could imply that credit scores are influenced by other factors like credit history, debt management, and financial behavior.

Overall Summary: The dashboard suggests that while certain factors like employment type, marital status, and education level have some impact on loan amounts and debt ratios, they do not dramatically change outcomes. Credit scores and income do not show a strong direct correlation, indicating a more complex relationship influenced by a broader set of financial behaviors and history.

4. Results and discussion:

Logistic Regression, random forest, decision tree and Cat gradient boosting machine algorithm were used to predict Loan Eligibility of customers. Among the given algorithms Cat Gradient Boosting Machine algorithm was the best performing one as it provided the highest F1Score and Accuracy of 80%.

```
from sklearn.model_selection import train_test_split

X_train, x_test, Y_train, y_test = train_test_split(x_sm, y_sm, train_size=
0.8, random_state=783437, stratify=y_sm)
```

```
# Built the catboost model
from catboost import CatBoostClassifier
model_catboost = CatBoostClassifier().fit(X_train,Y_train)
model_catboost
```

```
              precision    recall  f1-score   support

           0       0.75      0.86      0.80     12000
           1       0.83      0.72      0.77     12000

    accuracy                           0.79     24000
   macro avg       0.79      0.79      0.79     24000
weighted avg       0.79      0.79      0.79     24000
```

5. GUI:

GUI is made using Flask framework. **Flask** is a micro web framework written in Python. It is classified as a microframework because it does not require particular tools or libraries. It has no database abstraction layer, form validation, or any other components where pre-existing third-party libraries provide common functions. However, Flask supports extensions that can add application features as if they were implemented in Flask itself. Extensions exist for object-relational mappers, form validation, upload handling, various open authentication technologies and several common framework related tools

6.GitHubLink:https://github.com/orgs/B1-Tejaswini-84053/repositories

7.Future work And Conclusion

7.1Future Work:

Feature Engineering:

- Develop new features that could enhance the accuracy of both loan eligibility and loan amount prediction. For example, features like the ratio of loan amount to income or the applicant's spending behavior could be useful.

Loan Amount Prediction:

- Extend the model to not only predict whether a loan will be approved but also to predict the optimal loan amount a borrower can receive. This could involve using regression models that consider factors like income, credit score, and debt-to-income ratio.

7.2 Conclusion:

- The loan eligibility prediction project successfully developed a predictive model that efficiently determines the likelihood of a loan application being approved. By leveraging advanced machine learning techniques, the model was able to achieve [mention accuracy or performance metrics], providing valuable insights to lenders and helping streamline the decision-making process.

Key Findings:

1. **Feature Importance**: The model identified key features such as credit score, income, and debt-to-income ratio as critical determinants of loan eligibility, highlighting the importance of financial stability in loan approval decisions.

2. **Model Performance**: The model performed well across various metrics, demonstrating its ability to accurately predict loan eligibility. It also showed robustness against overfitting, making it reliable for real-world applications.

3. **Business Impact**: Implementing this model can significantly reduce the time and resources required for manual loan application reviews, increasing efficiency and ensuring faster processing times for applicants.