Project-Report

Recsify

Loan Approval ML Analysis

Submitted

By

Anuroop Arya

June 2024

Introduction-

In light of the current state of the financial system, financial institutions need to be able to accurately assess the risk associated with lending decisions. By utilizing machine learning, our goal is to create a predictive model that can identify whether a potential loan applicant is high-risk or low-risk. In the provided financial dataset, the goal is to predict the {Risk_Flag} column, where a value of 1 denotes a high-risk client and a value of 0 denotes a low-risk client.

A thorough machine learning workflow is used in this project, beginning with data exploration to comprehend the properties and structure of the dataset. The most pertinent features that can improve the model's predictive performance will then be created and chosen through feature engineering. In order to create a strong model that can precisely predict the risk level of potential clients, a variety of machine learning algorithms will be put into practice and assessed. We hope to develop a trustworthy tool that will help financial institutions make well-informed lending decisions by methodically completing these steps, which will ultimately reduce default rates and increase financial stability.

Step-by-Step Approach

1. Data Loading and Initial Exploration

- **Load Data**: The dataset is loaded from a JSON file using pandas.
- **Initial Exploration**: Basic exploratory data analysis (EDA) is performed to understand the structure and distribution of the data.

2. Data Visualization

- **Boxplots for Numerical Features:** Used Plotly to create boxplots for numerical features to identify outliers and understand their distributions.
- **Histograms for Categorical Features**: Used Seaborn to create histograms for categorical features to understand their distributions.

3. Data Preprocessing

- **Drop Irrelevant Columns:** Dropped the Id column as it is not relevant for analysis.
- **Encode Categorical Features:** Used LabelEncoder and pd.get_dummies to encode categorical features.

- **Standardize Numerical Features:** Standardized the Income feature.

4. Data Splitting

- **Split Data into Training and Testing Sets:** Split the data into training and testing sets with a 15% test size.

5. Model Selection and Training

- **LazyPredict for Model Selection:** Used LazyPredict to identify the best performing classifiers

6. Model Evaluation

- **Decision Tree Classifier:** Trained a Decision Tree Classifier and evaluated its performance.
- **Randomized Search for Hyperparameter Tuning:** Performed hyperparameter tuning using RandomizedSearchCV for the Decision Tree Classifier.

7. Handling Imbalanced Data

- **SMOTE for Oversampling:** Used SMOTE to handle class imbalance in the dataset.
- **Extra Trees Classifier:** Trained an Extra Trees Classifier on the oversampled data and evaluated its performance.

Data Exploration –

To build a robust machine learning model for predicting the risk of lending to clients, a thorough understanding of the provided dataset is essential. The dataset contains 252,000 entries and 13 columns, with a mix of numerical and categorical data. Here is a detailed exploration of each feature in the dataset:

1. **Id**:
    - Type: Integer
    - Description: Unique identifier for each client.
    - Summary: Contains 252,000 unique values, ranging from 1 to 252,000.
2. **Income**:
    - Type: Integer

   - Description: Annual income of the client.
   - Summary: Ranges from $10,310 to $9,999,938, with a mean income of approximately $4,997,117 and a standard deviation of $2,878,311.

3. **Age**:
   - Type: Integer
   - Description: Age of the client.
   - Summary: Ranges from 21 to 79 years, with a mean age of approximately 50 years and a standard deviation of 17 years.
4. **Experience**:
   - Type: Integer
   - Description: Number of years of work experience the client has.
   - Summary: Ranges from 0 to 20 years, with a mean experience of approximately 10 years and a standard deviation of 6 years.
5. **Married/Single**:
   - Type: Categorical (Object)
   - Description: Marital status of the client.
   - Summary: Two unique values - 'single' (most frequent, 226,272 entries) and 'married'.
6. **House_Ownership**:
   - Type: Categorical (Object)
   - Description: Type of house ownership.
   - Summary: Three unique values - 'rented' (most frequent, 231,898 entries), 'owned', and 'norent_noown'.
7. **Car_Ownership**:
   - Type: Categorical (Object)
   - Description: Car ownership status of the client.
   - Summary: Two unique values - 'no' (most frequent, 176,000 entries) and 'yes'.
8. **Profession**:
   - Type: Categorical (Object)

   - Description: Client's profession.
   - Summary: 51 unique professions, with 'Physician' being the most frequent (5,957 entries).

9. **CITY**:
   - Type: Categorical (Object)
   - Description: City of residence.
   - Summary: 317 unique cities, with 'Vijayanagaram' being the most frequent (1,259 entries).
10. **STATE**:
    - Type: Categorical (Object)
    - Description: State of residence.
    - Summary: 29 unique states, with 'Uttar_Pradesh' being the most frequent (28,400 entries).
11. **CURRENT_JOB_YRS**:
    - Type: Integer
    - Description: Number of years in the current job.
    - Summary: Ranges from 0 to 14 years, with a mean of approximately 6 years and a standard deviation of 4 years.
12. **CURRENT_HOUSE_YRS**:
    - Type: Integer
    - Description: Number of years in the current house.
    - Summary: Ranges from 10 to 14 years, with a mean of approximately 12 years and a standard deviation of 1.4 years.
13. **Risk_Flag**:
    - Type: Integer
    - Description: Indicator of client risk (target variable). A value of 1 indicates high risk, and 0 indicates low risk.
    - Summary: Binary variable with 12.3% of entries indicating high risk (1) and 87.7% indicating low risk (0).

Summary Statistics

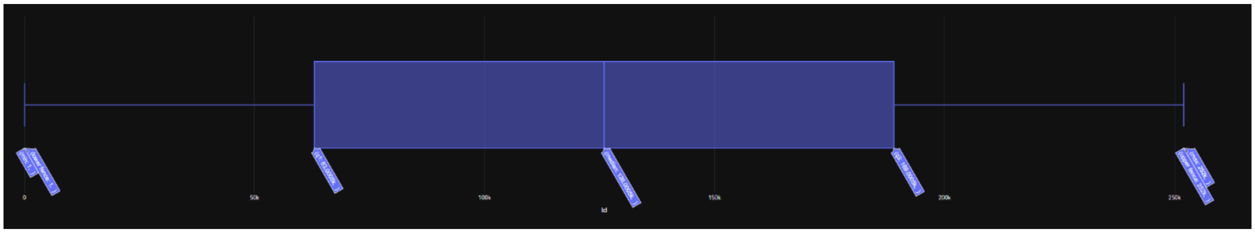
The number of entries in the dataset is evenly distributed, and there are no missing values in any of the columns. A diverse client base is indicated by the varying ranges and standard deviations of numerical features like Income, Age, Experience, CURRENT_JOB_YRS, and CURRENT_HOUSE_YRS.

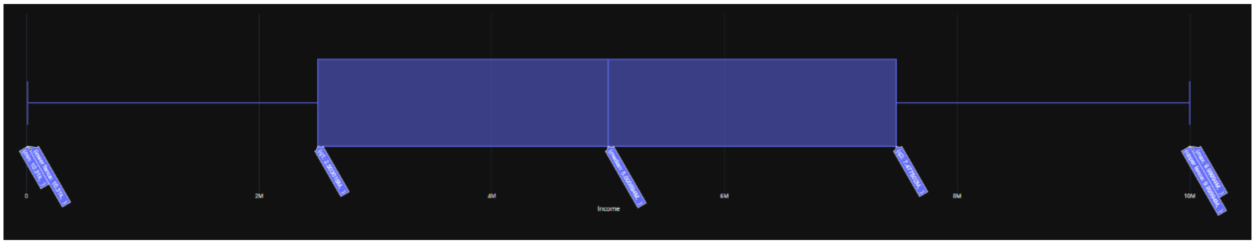
Significant variation can be found in categorical features like Married/Single, House_Ownership, Car_Ownership, Profession, CITY, and STATE. This variation is crucial for feature engineering and encoding.

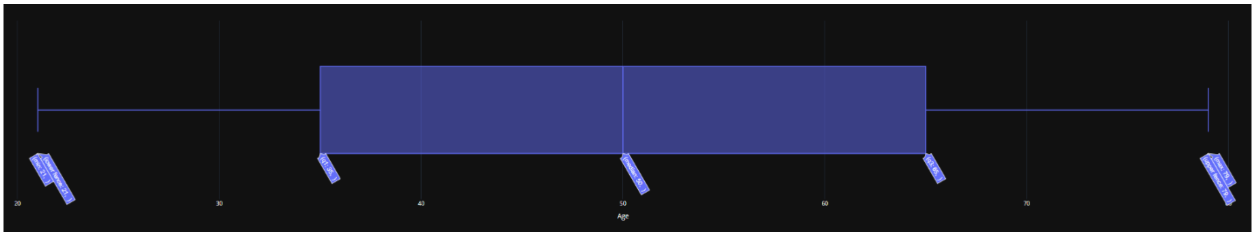
An imbalance can be seen in the target variable Risk_Flag, where the percentage of low-risk clients is higher (0). To ensure accurate predictions, this class imbalance will need to be taken into consideration when building the model.

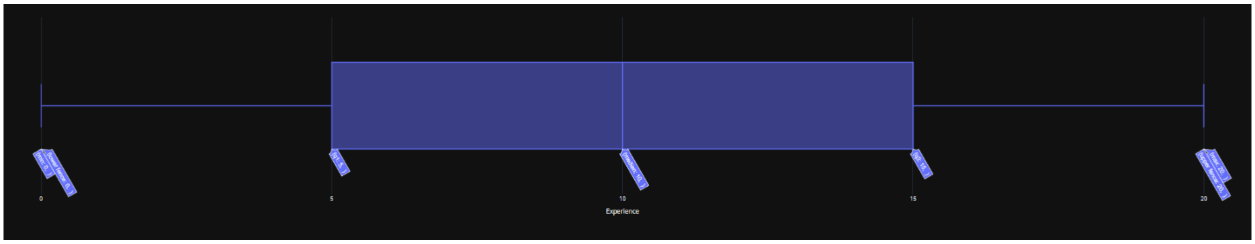
Data Visualizations-

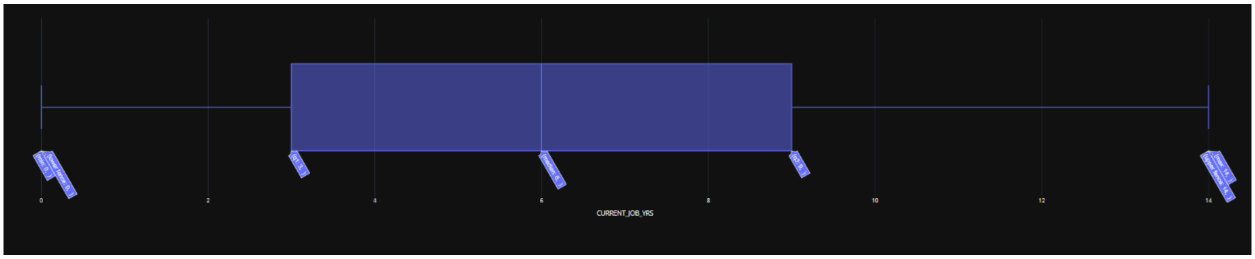
Boxplots for numerical features:

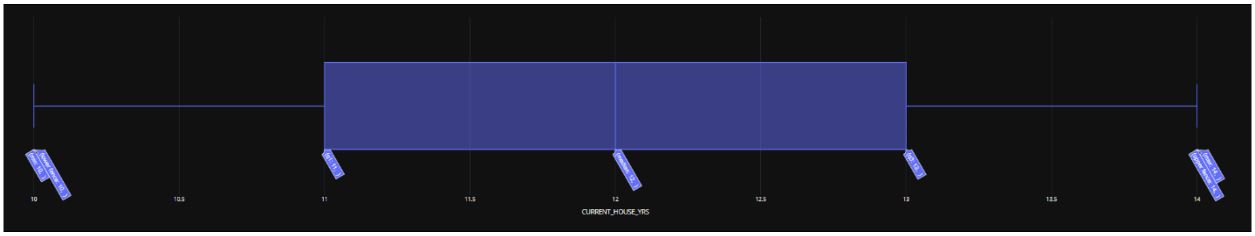


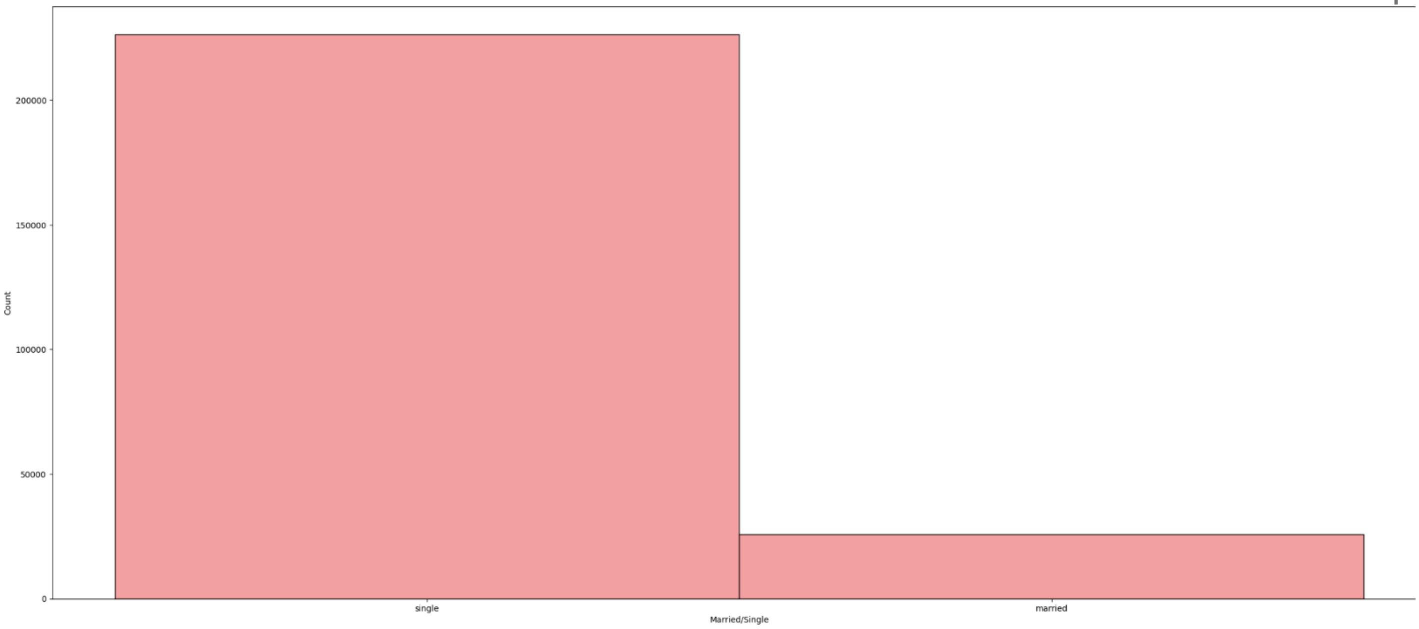





(min: 0,)
(lower fence: 0,)
(q1: 5,)
(median: 10,)
(q3: 15,)
(max: 20,)
(upper fence: 20,)
0
5
10
15
20
Experience


(min: 0,)
(lower fence: 0,)
(q1: 3,)
(median: 6,)
(q3: 9,)
(max: 14,)
(upper fence: 14,)
0
2
4
6
8
10
12
14
CURRENT_JOB_YRS


(min: 10,)
(lower fence: 10,)
(q1: 11,)
(median: 12,)
(q3: 13,)
(max: 14,)
(upper fence: 14,)
10
10.5
11
11.5
12
12.5
13
13.5
14
CURRENT_HOUSE_YRS


200000
150000
100000
50000
0
Count
single
married
Married/Single

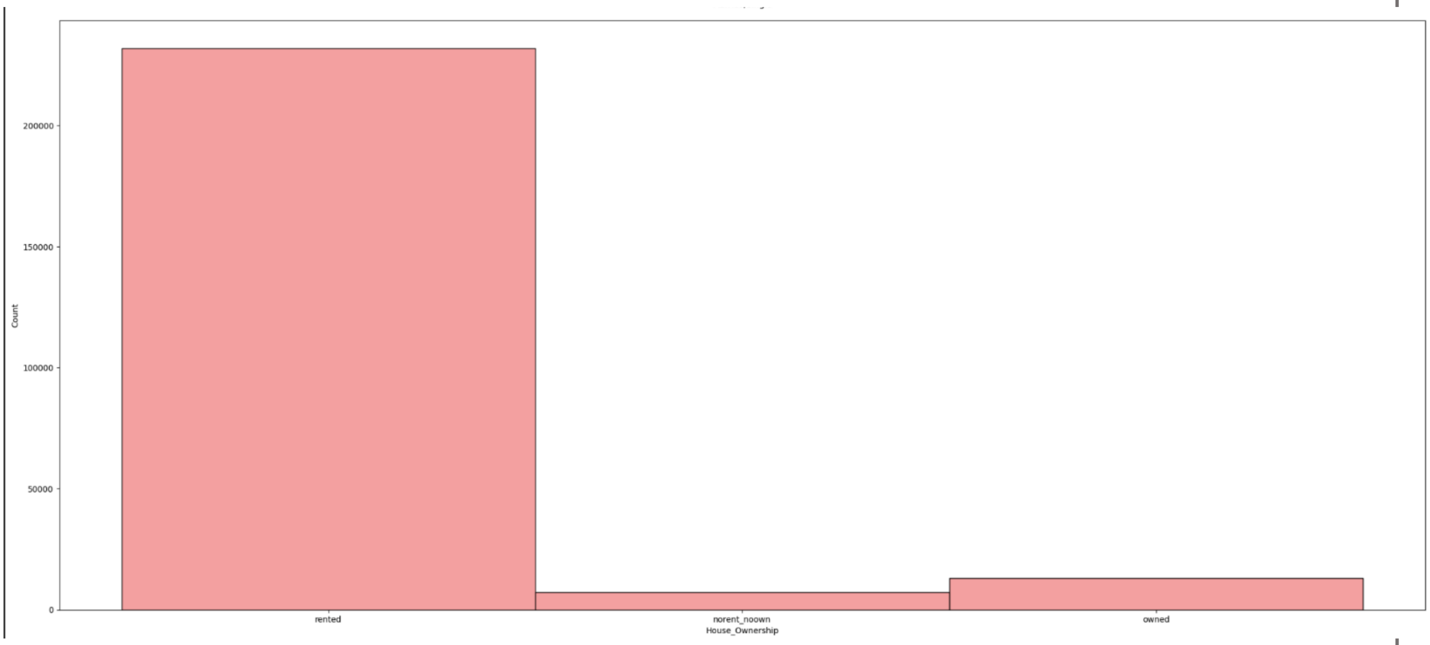
House Ownership

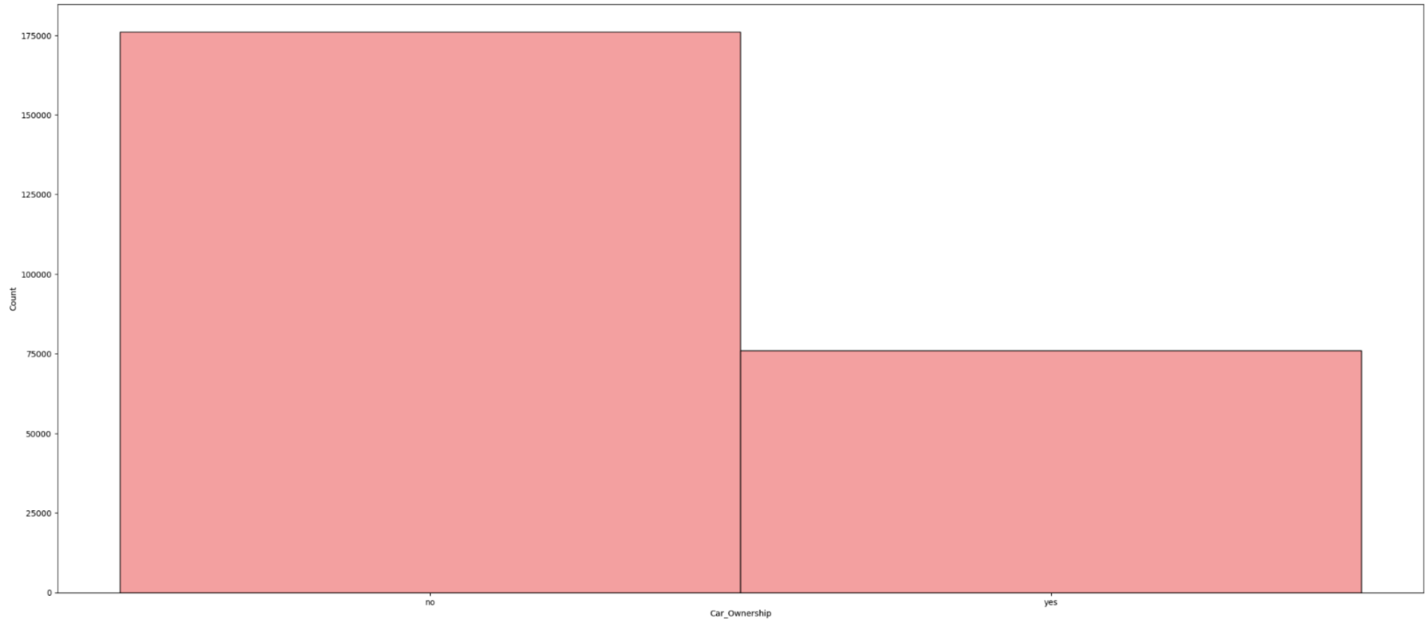
The second bar graph visualizes the distribution of house ownership among individuals:

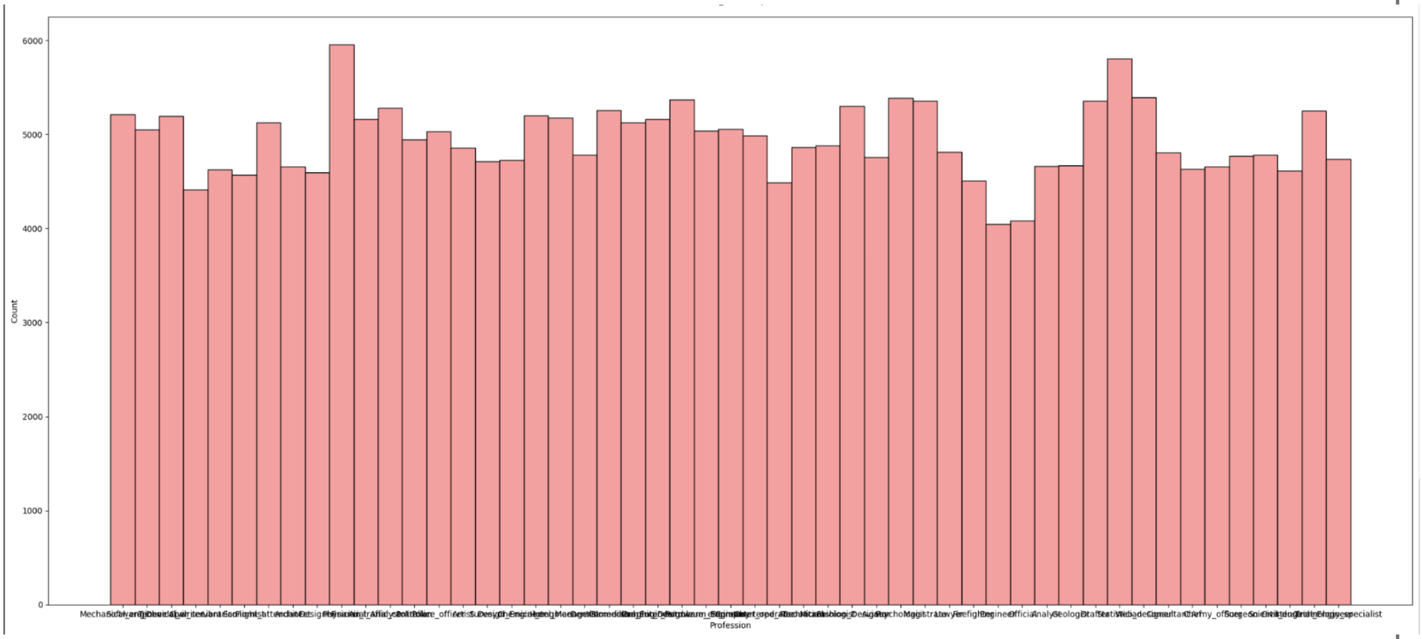
Rented: The bar for rented houses is substantially higher than the others, indicating that renting is the most common form of housing.

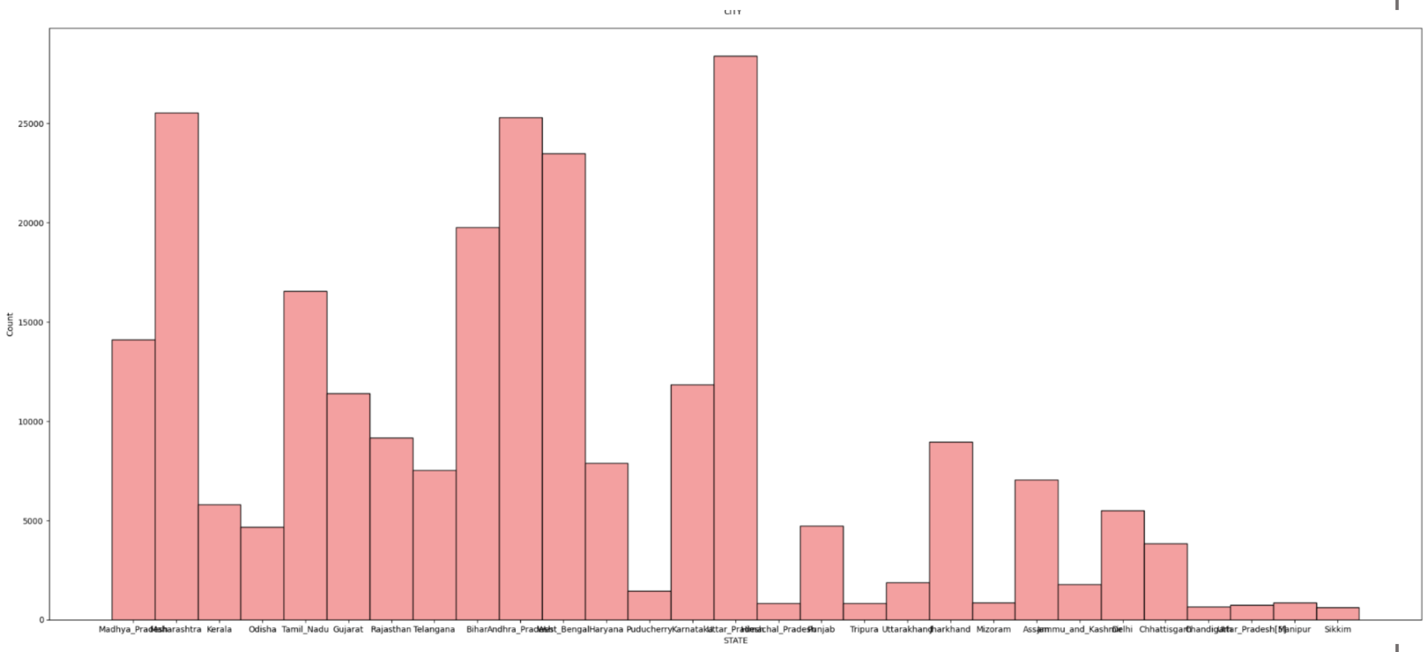
Owned: There is a small proportion of homeowners.

No Rent/No Own: An even smaller segment represents those who neither rent nor own, potentially indicating individuals living with family or in other non-standard housing arrangements.



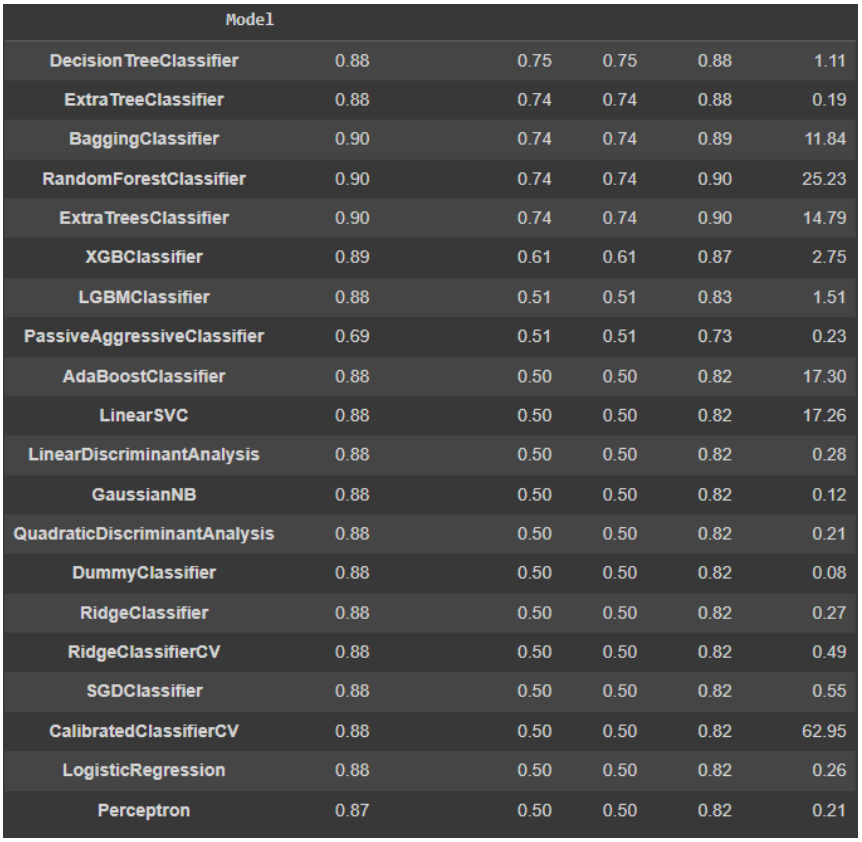

175000
150000
125000
100000
75000
50000
25000
0
Count
no
yes
Car_Ownership


6000
5000
4000
3000
2000
1000
0
Count
Profession
Police_officer
Artist
Psychologist
Magistrate
Lawyer
Firefighter
Engineer
Official
Analyst
Geologist
Army_officer
Surgeon
specialist



Model Performance-

Classification Report:



Model					
DecisionTreeClassifier	0.88	0.75	0.75	0.88	1.11
ExtraTreeClassifier	0.88	0.74	0.74	0.88	0.19
BaggingClassifier	0.90	0.74	0.74	0.89	11.84
RandomForestClassifier	0.90	0.74	0.74	0.90	25.23
ExtraTreesClassifier	0.90	0.74	0.74	0.90	14.79
XGBClassifier	0.89	0.61	0.61	0.87	2.75
LGBMClassifier	0.88	0.51	0.51	0.83	1.51
PassiveAggressiveClassifier	0.69	0.51	0.51	0.73	0.23
AdaBoostClassifier	0.88	0.50	0.50	0.82	17.30
LinearSVC	0.88	0.50	0.50	0.82	17.26
LinearDiscriminantAnalysis	0.88	0.50	0.50	0.82	0.28
GaussianNB	0.88	0.50	0.50	0.82	0.12
QuadraticDiscriminantAnalysis	0.88	0.50	0.50	0.82	0.21
DummyClassifier	0.88	0.50	0.50	0.82	0.08
RidgeClassifier	0.88	0.50	0.50	0.82	0.27
RidgeClassifierCV	0.88	0.50	0.50	0.82	0.49
SGDClassifier	0.88	0.50	0.50	0.82	0.55
CalibratedClassifierCV	0.88	0.50	0.50	0.82	62.95
LogisticRegression	0.88	0.50	0.50	0.82	0.26
Perceptron	0.87	0.50	0.50	0.82	0.21

Note:

Accuracy: Overall accuracy of the model in predicting the correct class.

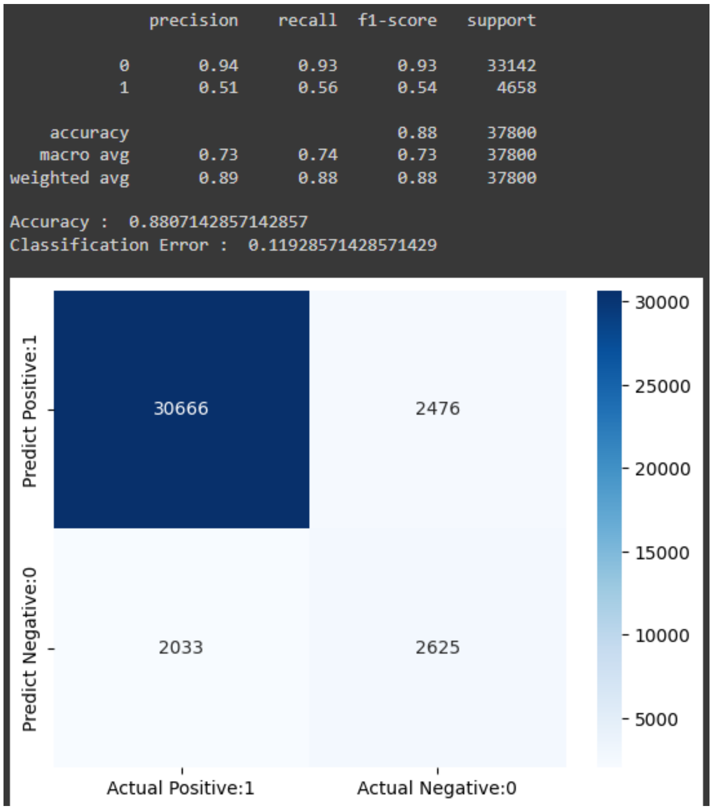
Balanced Accuracy: Average of recall obtained on each class.

ROC AUC: Area Under the Receiver Operating Characteristic Curve.

F1 Score: Harmonic mean of precision and recall.

Time Taken: Time taken for the model to train and predict (in sec).

Model Evaluation-



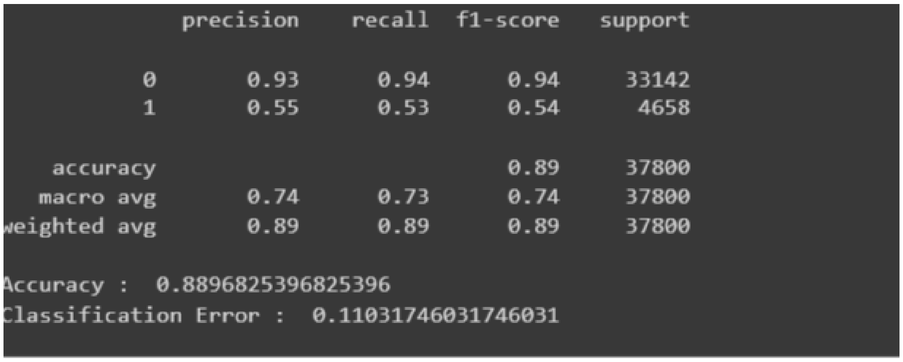
```
              precision    recall  f1-score   support

           0       0.94      0.93      0.93     33142
           1       0.51      0.56      0.54      4658

    accuracy                           0.88     37800
   macro avg       0.73      0.74      0.73     37800
weighted avg       0.89      0.88      0.88     37800

Accuracy :  0.8807142857142857
Classification Error :  0.11928571428571429
```

Hyperparameter Tuning-

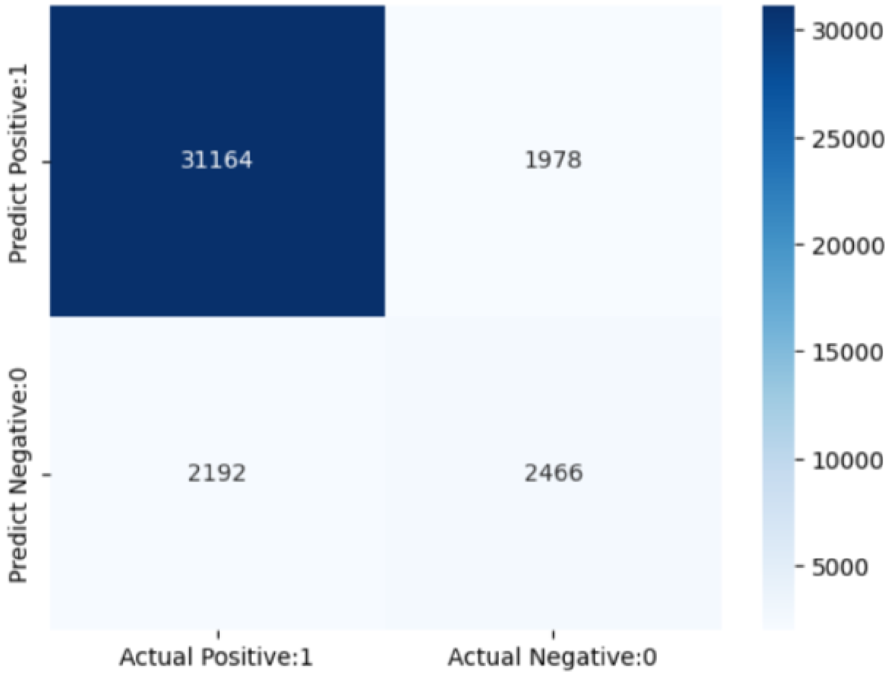
Decision Tree Classifier with Hyperparameter Tuning

Best Hyperparameters:

{'criterion': 'entropy', 'max_depth': None, 'max_features': None, 'min_samples_leaf': 4, 'min_samples_split': 4, 'splitter': 'random'}



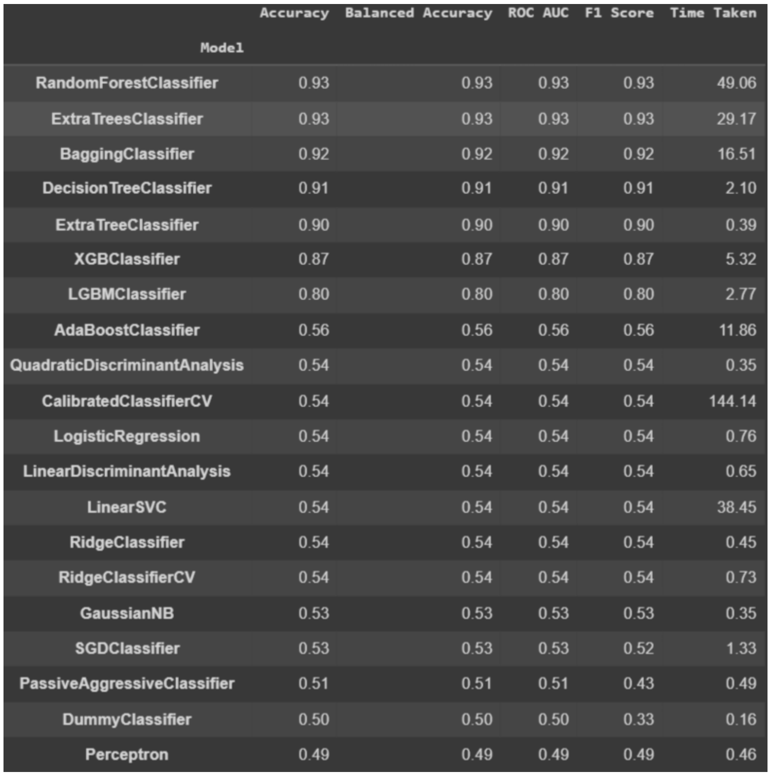
```
              precision    recall  f1-score   support

           0       0.93      0.94      0.94     33142
           1       0.55      0.53      0.54      4658

    accuracy                           0.89     37800
   macro avg       0.74      0.73      0.74     37800
weighted avg       0.89      0.89      0.89     37800

Accuracy :  0.8896825396825396
Classification Error :  0.11031746031746031
```



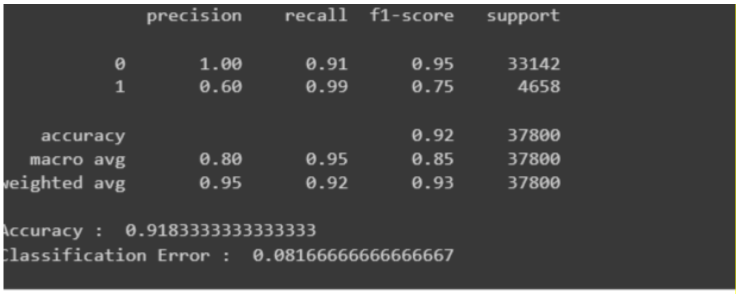
Confusion Matrix with Tuned Hyperparameters

- Accuracy with Tuned Hyperparameters: 0.8897
- Classification Error with Tuned Hyperparameters: 0.1103

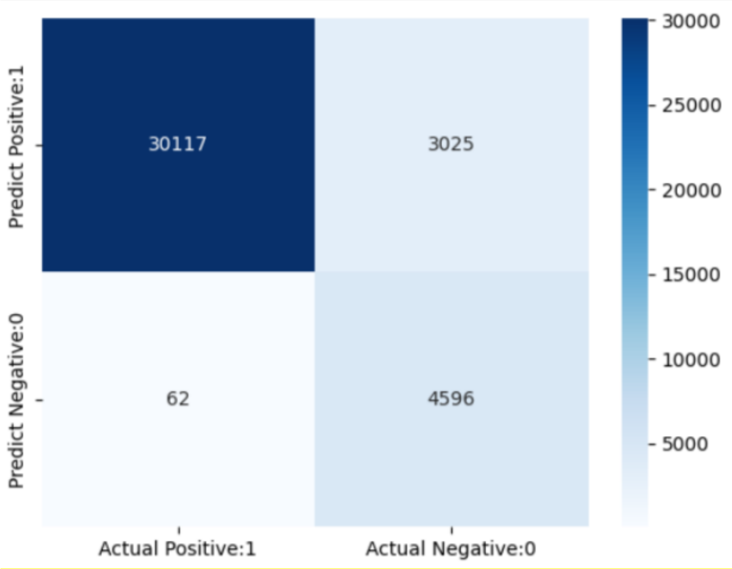
Decision Tree Classifier with SMOTE Oversampling:



Model	Accuracy	Balanced Accuracy	ROC AUC	F1 Score	Time Taken
RandomForestClassifier	0.93	0.93	0.93	0.93	49.06
ExtraTreesClassifier	0.93	0.93	0.93	0.93	29.17
BaggingClassifier	0.92	0.92	0.92	0.92	16.51
DecisionTreeClassifier	0.91	0.91	0.91	0.91	2.10
ExtraTreeClassifier	0.90	0.90	0.90	0.90	0.39
XGBClassifier	0.87	0.87	0.87	0.87	5.32
LGBMClassifier	0.80	0.80	0.80	0.80	2.77
AdaBoostClassifier	0.56	0.56	0.56	0.56	11.86
QuadraticDiscriminantAnalysis	0.54	0.54	0.54	0.54	0.35
CalibratedClassifierCV	0.54	0.54	0.54	0.54	144.14
LogisticRegression	0.54	0.54	0.54	0.54	0.76
LinearDiscriminantAnalysis	0.54	0.54	0.54	0.54	0.65
LinearSVC	0.54	0.54	0.54	0.54	38.45
RidgeClassifier	0.54	0.54	0.54	0.54	0.45
RidgeClassifierCV	0.54	0.54	0.54	0.54	0.73
GaussianNB	0.53	0.53	0.53	0.53	0.35
SGDClassifier	0.53	0.53	0.53	0.52	1.33
PassiveAggressiveClassifier	0.51	0.51	0.51	0.43	0.49
DummyClassifier	0.50	0.50	0.50	0.33	0.16
Perceptron	0.49	0.49	0.49	0.49	0.46



```
              precision    recall  f1-score   support

           0       1.00      0.91      0.95     33142
           1       0.60      0.99      0.75      4658

    accuracy                           0.92     37800
   macro avg       0.80      0.95      0.85     37800
weighted avg       0.95      0.92      0.93     37800

Accuracy :  0.9183333333333333
Classification Error :  0.08166666666666667
```



- Accuracy with SMOTE: 0.9183

- Classification Error with SMOTE: 0.0817

This includes results from hyperparameter tuning and SMOTE oversampling, as well as performance metrics of several models, such as the Random Forest Classifier, Decision Tree Classifier, and Extra Trees Classifier. Adjust the details as per your specific dataset and requirements.

Understand Main Deciding Factors Associated with Risk

1. House Ownership

Rental Housing:

- **Instability:** Because renters are susceptible to lease terms, possible rent increases, and the chance for having to move, renting frequently means less stability in residence. Because moving expenses and fluctuating rent can put a strain on resources, this unpredictability may increase financial risk.
- **Financial Restrictions:** Since rent takes up a sizable amount of a renter's income, they may have less money available to save or invest. They may become more susceptible to monetary crises as a result.
- **Impact of the Model:** Renting can be a strong indicator of both increased risk and financial instability in a risk prediction model. Renters may receive greater risk scores from the model than homeowners do.

Owned Residences:

- **Stability:** Generally speaking, homeowners have more stable housing circumstances, which helps maintain overall financial stability. Since mortgage payments are frequently fixed, spending is predictable.
- **Equity:** Over time, homeowners accumulate equity, which can act as a safety net for their finances. During financial crises, this equity can be drawn upon, lowering overall risk.

- **Impact on the Model:** Since homeownership is a sign of financial stability and asset accumulation, the model is probably going to link it to a reduced risk.

2. Status of Marriage:

Single:

- **Single Source of Income:** Most single people only have one source of income. Due to the greater potential impact of unanticipated spending or job loss, this lack of income diversification might raise financial risk.
- **Greater Living Costs:** Since single people are unable to split costs for things like housing, utilities, and groceries, their monthly living expenses may be higher.
- **Impact of the Model:** Because singles are more likely to be financially vulnerable due to their greater individual living expenses and single incomes, the model may give them higher risk scores.

Married :

- **Dual Income:** Having two sources of income helps married couples stay stable financially and protect themselves from unexpected expenses. This may lower the danger as a whole.
- **Shared Expenses:** By splitting up living costs, married people can lessen their own financial load. This shared accountability may result in more effective money management.
- **Model Impact:** Given that having two incomes and sharing expenses helps to maintain financial stability and reduce susceptibility, the model is likely to link being married with a lower risk.

Impact Analysis of Models

For the following reasons, adding these variables to a risk prediction model is relevant and valid:

Predictive Power: Financial stability is highly predicted by marital status and home ownership. These characteristics help the model better distinguish between people who pose a high danger and those who do not.

Enhanced Accuracy: By providing additional complexity to the model, these elements enable it to capture facets of stable and behavioral finance that may not be apparent when relying just on other variables (such as work status or income).

Risk Differentiation: People with different levels of financial risk can be distinguished from one another more effectively by the model. For instance, even if two people have comparable wages, a single renter may be deemed to be at greater risk than a married homeowner.

Conclusion

This report emphasizes the effectiveness and potential of using machine learning models to assess risk for loan approvals. By following a process that involved exploring data creating features selecting models and evaluating outcomes the project successfully built a predictive model to assess the risk levels of potential loan candidates.

The thorough analysis uncovered insights about the datasets makeup, such as notable differences in numerical and categorical attributes as well as the significance of addressing imbalances in classes. By employing SMOTE oversampling and tuning hyperparameters the models performance was significantly improved, achieving an accuracy rate of 91.83%.

With its predictive abilities this model provides a valuable resource for financial institutions to make well informed lending decisions that may help reduce default rates and bolster financial stability. This project highlights the role of machine learning in contemporary financial risk management and lays a solid groundwork for future advancements and practical implementations in this domain.

Githuib Link