Problem Statement:

Based on the given financial data create a ML model to predict if the client is high risk or low risk if we were to provide them loan. We need to predict the column Risk_Flag and it contains value 1 if the client is high risk else it will be 0.

Dataset Variable Description:

1) Income: Annual salary
2) Age: Age of the person
3) Experience: Work Experience
4) Married/Single: Married or Single
5) House_Ownership: rented/owned/ norent_noown
6) Car_Ownership: Owns car or not (yes/no)
7) Profession: Type of profession
8) City: City
9) State: State
10) CURRENT_JOB_YRS: How many years the person is working on current job.
11) CURRENT_HOUSE_YRS: How many years the person is living in current house.
12) Risk_Flag: Target Variable, 0-Good customer for loan, 1- Risky customer to give loan

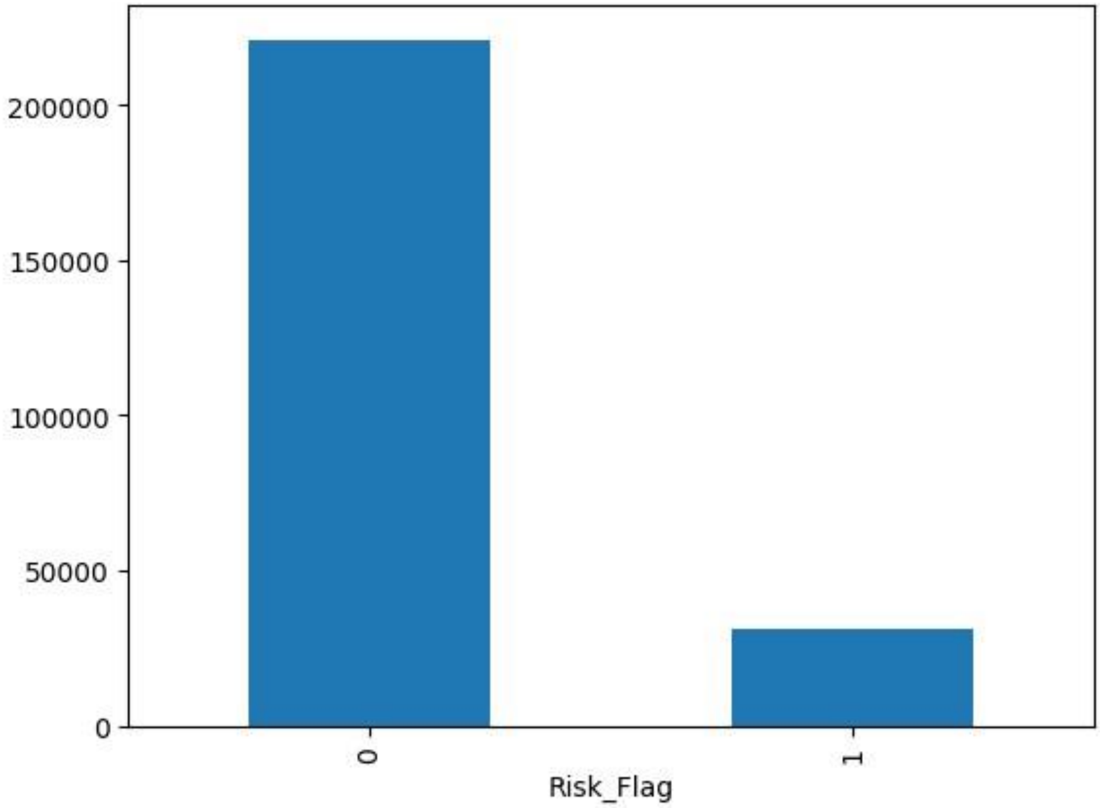
Methodology:

1) Checked for missing values in dataset.
2) Analyzed unique values of each variable and identified categorical and continuous variables.
3) Checked for any Outliers in dataset by plotting boxplots.
4) Conducted Univariate analysis for each variable. Checked for frequency distribution and also conducted ANOVA analysis for checking the impact of categorical variable wrt target variable.
5) Encoding of categorical variables.
6) Scaling of data.
7) Model-1 → Logistic Regression Classifier was built. Metrics evaluated. (LR)
8) Model-2 → Random Forest Classifier was built. Metrics evaluated. (RF)
9) Model-3 → SMOTE on imbalanced dataset and then Random Forest Classifier was built Metrics evaluated. (SMOTE+RF)
10) Important Features was identified.

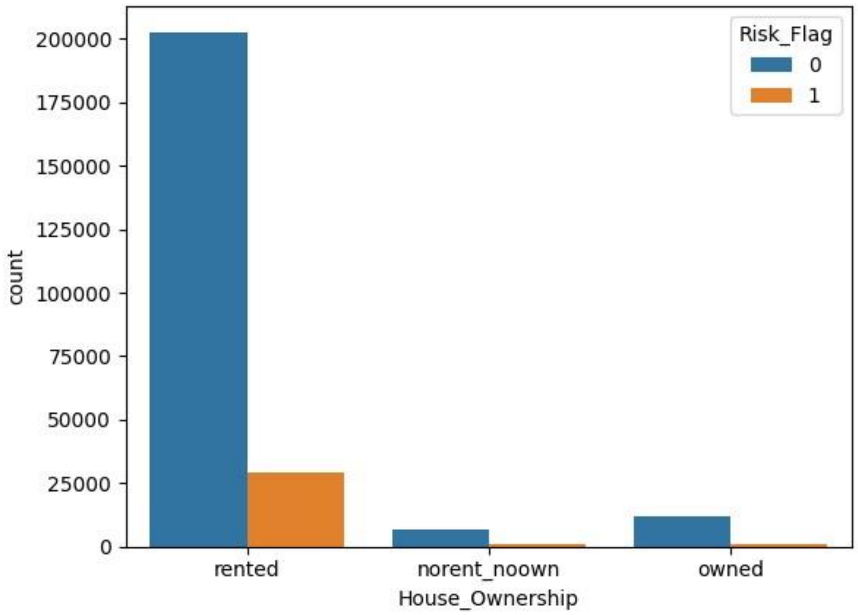
EDA Insights and Data Visualization:

- Distribution of classes is unbalanced in dataset.

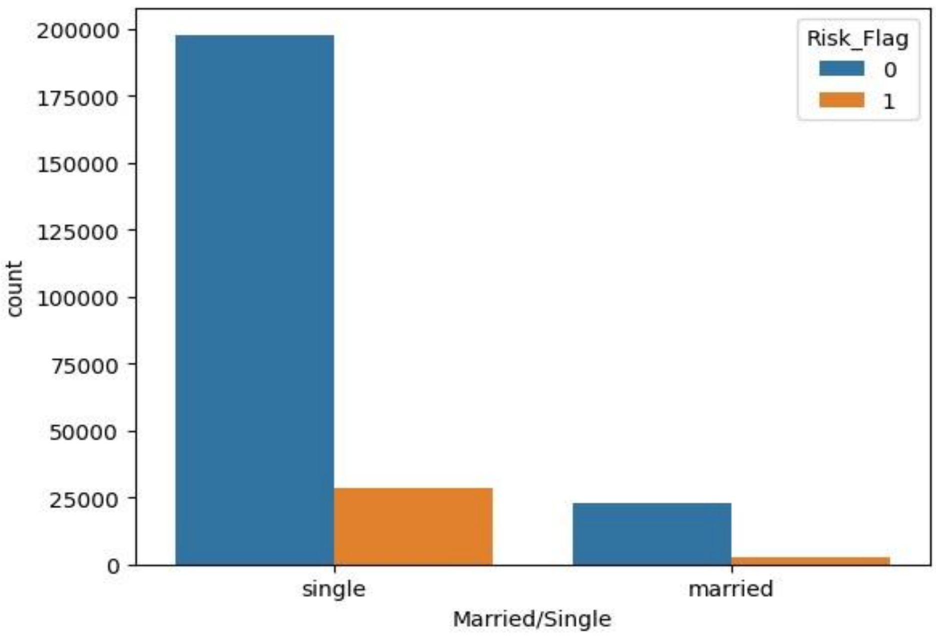
0- Good customers to give loan	**87.7%**
1- Risky Customers to give loan	**12.3%**



- Distribution of House_ownership with respect to target variable. It suggests that the customers living in rented house are riskier since there is a possibility for these customer to pay rent as well as pay loan amount which can be a burden to them. Thus a possibility that offering loan to rented customers is risky.



- Distribution of Married/Single with respect to the target variable. Riskier customers are observed from customers who are single. Thus suggesting a possibility that married customers can pay the loan amount because of their dependents who might be working.



Metrics Used:

1. Precision

Definition: Precision is the ratio of true positive predictions to the total predicted positives. It answers the question: "Out of all customers predicted to be risky, how many actually were risky?"

Formula:

$$\text{Precision} = \frac{TP}{TP+FP}$$

- **TP (True Positives)**: Number of customers correctly predicted as risky (Risk_Flag = 1).
- **FP (False Positives)**: Number of customers incorrectly predicted as risky when they were actually good (Risk_Flag = 0).

Interpretation: High precision means that when the model predicts a customer as risky, it is usually correct. This is important for minimizing the number of good customers incorrectly labeled as risky.

2. Recall

Definition: Recall is the ratio of true positive predictions to the total actual positives. It answers the question: "Out of all the actual risky customers, how many were correctly identified by the model?"

Formula:

$$\text{Recall} = \frac{TP}{TP+FN}$$

- **TP (True Positives)**: Number of customers correctly predicted as risky (Risk_Flag = 1).
- **FN (False Negatives)**: Number of customers incorrectly predicted as good when they were actually risky (Risk_Flag = 1).

Interpretation: High recall means that the model is able to identify a large portion of the risky customers. This is important for catching as many risky customers as possible, even if it means labeling some good customers as risky.

3. F1-Score

Definition: The F1-score is the harmonic mean of precision and recall. It balances the two metrics and is useful when you need a single metric to evaluate the performance of your model.

Formula:

$$\text{F1-Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

Interpretation: The F1-score provides a single measure of model performance when there is a trade-off between precision and recall. A high F1-score indicates a good balance between precision and recall.

4. ROC-AUC Score

Definition: The ROC-AUC score is the area under the Receiver Operating Characteristic (ROC) curve. The ROC curve plots the true positive rate (recall) against the false positive rate (1 - specificity) at various threshold settings.

Interpretation: The ROC-AUC score ranges from 0 to 1. A score of 0.5 indicates no discrimination (the model is no better than random), while a score of 1 indicates perfect discrimination by the model. A higher ROC-AUC score means that the model is better at distinguishing between good and risky customers.

Model-1: Logistic Regression Classifier

TRAINING DATA

	precision	recall	f1-score	support
0	0.91	0.57	0.70	154703
1	0.17	0.62	0.27	21697
accuracy			0.58	176400
macro avg	0.54	0.60	0.48	176400
weighted avg	0.82	0.58	0.65	176400

ROC-AUC Score:0.6

TEST DATA

	precision	recall	f1-score	support
0	0.91	0.58	0.71	66301
1	0.17	0.60	0.26	9299
accuracy			0.58	75600
macro avg	0.54	0.59	0.48	75600
weighted avg	0.82	0.58	0.65	75600

ROC-AUC Score:0.59

- This model performs badly on predicting Class-1 which is Risky customer for giving loan.
- Out of all customers predicted to be risky, the model is only able to predict 17% of customers who are actually risky (Precision). This means that when the model predicts that a customer is risky, it is right only 17% of the time.
- Low F-1 score for class-1 indicates that there is poor balance between recall and precision.

Model-2: Random Forest Classifier

TRAINING DATA

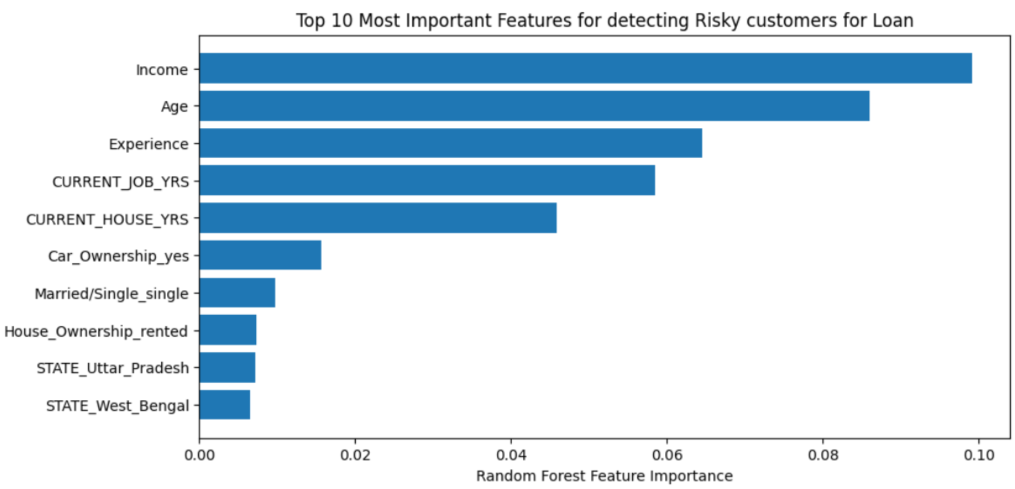
	precision	recall	f1-score	support
0	0.99	0.92	0.95	154703
1	0.61	0.95	0.74	21697
accuracy			0.92	176400
macro avg	0.80	0.93	0.85	176400
weighted avg	0.95	0.92	0.93	176400

ROC-AUC Score:0.93

TEST DATA

	precision	recall	f1-score	support
0	0.97	0.91	0.94	66301
1	0.56	0.77	0.65	9299
accuracy			0.90	75600
macro avg	0.76	0.84	0.79	75600
weighted avg	0.92	0.90	0.90	75600

ROC-AUC Score:0.84



- Precision, Recall and F-1 score have improved over the logistic regression classifier model on test data.
- Out of all customers predicted to be risky, the model is only able to predict 56% of customers who are actually risky (Precision). This means that when the model predicts that a customer is risky, it is right only 56% of the time.
- The top 5 features influencing for approving loan are: Income, Age, Experience, Current_job_years and current_house years.

Model 3- Random Forest Classifier with SMOTE

On comparing the metrics of training and testing dataset, it is observed that F1-score is very less for testing dataset. The reason for this is because of imbalanced dataset of the label class.

Usually when dealing with imbalanced datasets accuracy and roc_score are not good measures. When the positive class is more important F-1 score is usually focussed upon.

In order to increase the performance of our ML model, we need to create a balanced dataset of labels and then check metrics.

SMOTE- Synthetic Minority Oversampling Technique.

It's a popular technique used to address class imbalance by generating synthetic samples for the minority class.

For Training data:

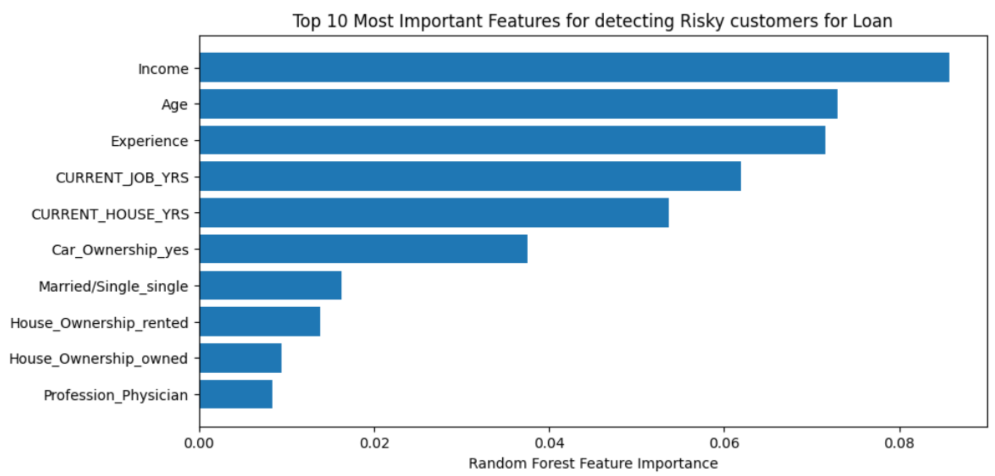
	precision	recall	f1-score	support
0	0.99	0.91	0.95	154858
1	0.92	0.99	0.96	154547
accuracy			0.95	309405
macro avg	0.96	0.95	0.95	309405
weighted avg	0.96	0.95	0.95	309405

ROC-AUC Score:0.95

For Testing data:

	precision	recall	f1-score	support
0	0.97	0.91	0.94	66146
1	0.91	0.97	0.94	66457
accuracy			0.94	132603
macro avg	0.94	0.94	0.94	132603
weighted avg	0.94	0.94	0.94	132603

ROC-AUC Score:0.94



- The model performs excellently on test data for both classes.

- F-1 score of 94% indicates that there is a balance between recall and precision.
- ROC_AUC score of 95% indicates that the model is good at distinguishing between the two classes.
- The top 5 features influencing for approving loan are: Income, Age, Experience, Current_job_years and current_house years.

Model-1	**F1-score: 26%**
Model-2	**F1-score: 65%**
Model-3	**F1-score: 94%**

Hence Model-3 is best model for predicting the classes for loan approval.

Factors affecting risk:

- **Income:** If the person has a low income, then the person will have a hard time repaying the loan. Hence it is the most important factor for assessing loan approval.
- **Age:** If the person's age is too young or too old, then the person might be risky to give out loan since no or less income.
- **Experience:** If the person has very less work experience means that there is a possibility that the person recently joined job and may have less income.
- **Current Job Years:** A customer who has been in their current job for many years may be seen as more stable and reliable. It suggests a steady income, lower risk of unemployment, and an established career, which can positively impact their ability to repay the loan. A customer with a short tenure in their current job might be viewed as less stable. Frequent job changes can indicate a riskier financial situation, potentially leading to a higher likelihood of default.
- **Current House Years:** A customer who has lived in their current house for many years may be seen as more stable and less likely to move. This stability can indicate a lower risk of default, as moving frequently can be associated with financial instability. Frequent moves might be a red flag for lenders, indicating potential financial difficulties, lack of stability, or other underlying issues.

LINK FOR THE PYTHON CODE:

https://colab.research.google.com/drive/15o1o7sqMeMEnMaW-i7WlvayUc-8nPP4A?usp=sharing