Aim:

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.

Introduction:

The Loan risk prediction model was based on Logistic regression technique of supervised learning in the presence of given dataset.

The risk factor was evaluated and the prediction was generated to determine whether the client is High risk or Low risk if provide loan on the basis of Risk\_Flag.

Code:

<https://colab.research.google.com/drive/1yqKh4XvlkAfOT6leZaI2CV32-nhtNEZi?usp=sharing>

Libraries used:

1. Numpy
2. Pandas
3. Matplotlib.pyplot
4. Seaborn
5. Sklearn
6. Tensorflow
7. Plotly
8. Json
9. Csv

Data Visalization:

Libraries like Matplotlib.pyplot, Seaborn, Sklearn, Tensorflow and Plotly are used for the visualization of data in the forms of:

1. Confusion Matrix: A confusion matrix in machine learning summarizes the predictions of a classification model as follows:

* True Positive (TP): Model predicts positive (high-risk) correctly.
* True Negative (TN): Model predicts negative (low-risk) correctly.
* False Positive (FP): Model predicts positive (high-risk) incorrectly.
* False Negative (FN): Model predicts negative (low-risk) incorrectly.

It helps evaluate the model's performance beyond accuracy, especially in scenarios with imbalanced classes or varying costs of errors.

1. Classification Report: A classification report in machine learning provides precision, recall, F1-score, and support for each class. It summarizes model performance as follows:

* Precision: Proportion of true positive predictions among all positive predictions.
* Recall: Proportion of true positives correctly identified by the model.
* F1-score: Harmonic mean of precision and recall, balancing both metrics.
* Support: Number of occurrences of each class in the dataset.

1. ROC Curve: An ROC (Receiver Operating Characteristic) curve in machine learning visually evaluates the trade-off between true positive rate (sensitivity) and false positive rate (1-specificity) for different threshold values. Key points include:

* True Positive Rate (TPR): Ratio of true positives to all actual positives.
* False Positive Rate (FPR): Ratio of false positives to all actual negatives.
* **Threshold**: Decision boundary for class prediction, affecting TPR and FPR.

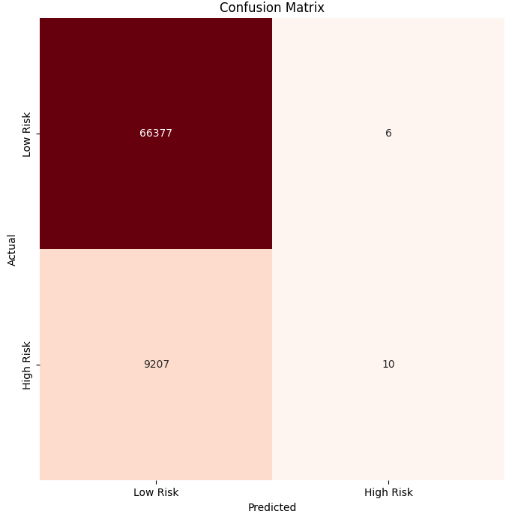
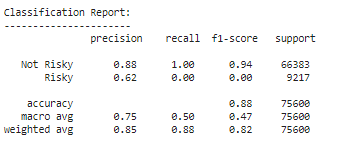
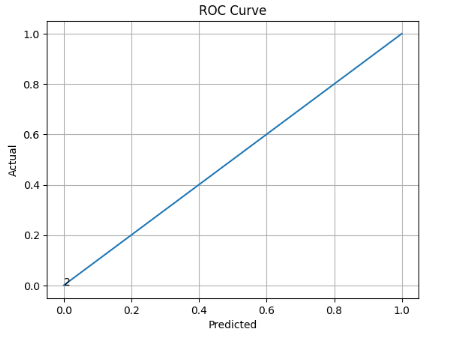
To visualize the model’s predictively and accuracy.

The accuracy\_score method of sklearn.metrics package is used to describe the model’s behavior across all the classes of the dataset.

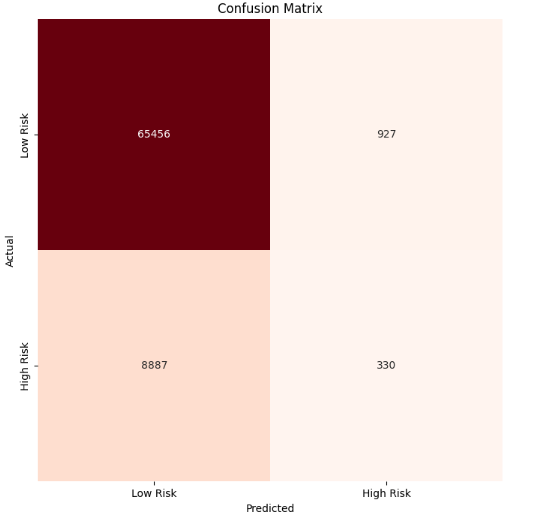
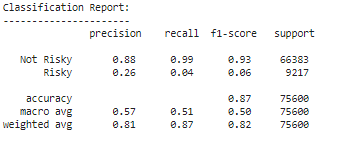
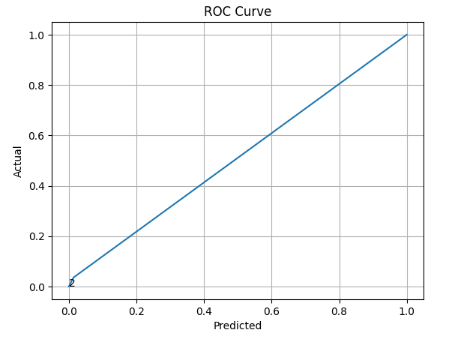
On the basis of observations based on the Classification report we can study that the

The accuracy score varies as per the classification threshold. I have taken three different threshold value and following are the outputs:

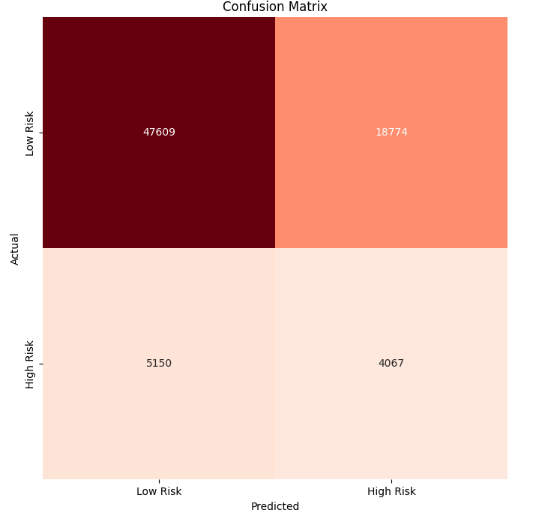
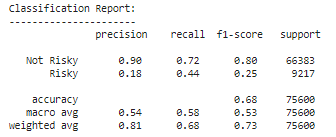
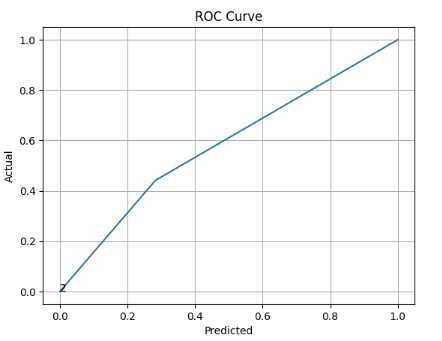
When Classification threshold=0.5 the Accuracy score= 87.81%

When Classification threshold=0.35 the Accuracy score= 87.02%

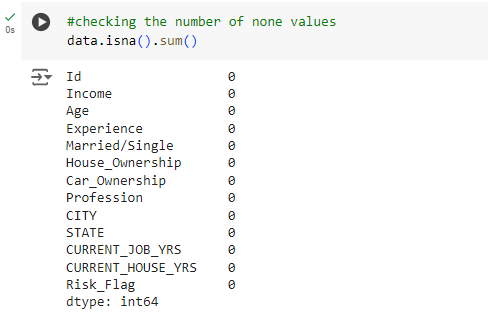
  

When Classification threshold=0.2 the Accuracy score= 68.35%

Data Exploration Insights:

1. Missing Values: Identifying features with missing values and handling them by either imputation or deletion as missing data can significantly impact model performance.



The above command and output shows that no attribute had any missing value thus the given dataset does not contain any NaN or Missing values.

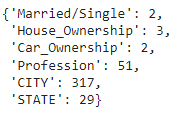
1. Encoding data: The given dataset contained multiple input data types like string, integer, etc…

* Thus to identify the columns containing data type values other than integer were identified and then encoded.

{column:list(data[column].unique()) for column in data.select\_dtypes('object').columns}

Command was used to fetch all the data other than integers. The output provided the list which was complex to read therefore

{column:len(data[column].unique()) for column in data.select\_dtypes('object').columns}

This command provided the length of all the attributes. 

* The attributes with length 2 were encoded by binary encoding technique, the attribute with length 3 was encoded by ordinal encoding and other attributes with high amount of values or more length were encoded by one hot encoding.
* While binary encoding the data of Married/Single and Car\_Ownership was classified into positive and negative set as 1 and 0 respectively.
* In ordinal encoding the data of Housing \_Ownership was arranged into an order of:

norent\_noown < rent < own.

* In One Hot encoding, columns with multiple categories were divided into sub categories with the prefixes of columns name. Dummies were created for these sub categorical data.

All of these encoding processes came under the pre-processing part of the model.

1. Data Distribution: Data was distributed on the basis of numerical and categorical features. This distribution allowed in understanding the distribution of high-risk vs. low-risk clients. A balanced dataset is ideal but often not the case in real-world scenarios.
2. Visualization: Utilization of visual tools (Confusion Matrix, Classification Report, ROC curve etc.) to understand the relationship between features and the target variable. Visual inspection can reveal patterns that might not be captured by numerical summaries alone.
3. Model Evaluation: Use appropriate metrics (accuracy, precision, recall, F1-score) to evaluate model performance, considering the specific business context (e.g., minimizing false negatives might be more critical than false positives in loan risk prediction).

 Model Performance:

The performance evaluation of a loan prediction model typically involves assessing how well the model predicts whether a client is high-risk or low-risk for a loan based on historical data.

The key Metric for evaluation:

1. Accuracy: It measures overall correctness of predictions and is calculated as:

**Accuracy = Number of Correct Prediction / Total number of predictions**

1. Precision and recall:

* Precision measures the proportion of true positive predictions (high risk) among all positive predictions.

**Precision=True positives / (True positives + False positives)**

* Recall (sensitivity) measures the proportion of true positives that were correctly identified.

**Recall=True positives / (True positive + False negative)**

1. F1- Score: It combines precision and recall into single metric, balancing both measures:

**F1-Score= 2 \* (Precision \* Recall) / (Precision + Recall)**

1. ROC- Curve: ROC Curve plots the true positive rate (recall) against the false positive rate. It helps visualize the trade-off between sensitivity and specificity.

Main deciding factors associated with risk:

When determining risk in the context of loan approval or similar decisions, several key factors play crucial roles:

1. Risk-Flag
2. Income
3. Profession
4. Owned Capital
5. Experience
6. Marital Status
7. Location
8. Current Job and House Years