MACHINE LEARNING INTERNSHIP
QUALIFICATION TASK

Title: Bank Loan classification

Introduction:

In this report, I am presenting the methodologies, findings, and insights obtained from the bank loan classification project. I describe the dataset used for training and testing our models, the preprocessing steps undertaken, and the machine learning algorithms employed. I also discuss the performance metrics of the trained models, their strengths and weaknesses, and the implications of accurate loan predictions. Finally, I am able to accurately classify the loan classification process.

Problem statement:

The problem at hand is to develop a machine learning model that accurately predicts the approval or rejection of bank loan applications.

Objective:

The objective is to create a model that can analyze loan applications using a range of relevant variables such as applicant demographics, income, credit score, employment history, and loan amount. By training the model on a labeled dataset comprising past loan applications with known outcomes, we aim to leverage machine learning algorithms to develop a robust classification system. The model will then be capable of accurately predicting whether a given loan application is likely to be approved or rejected.

Data Description:

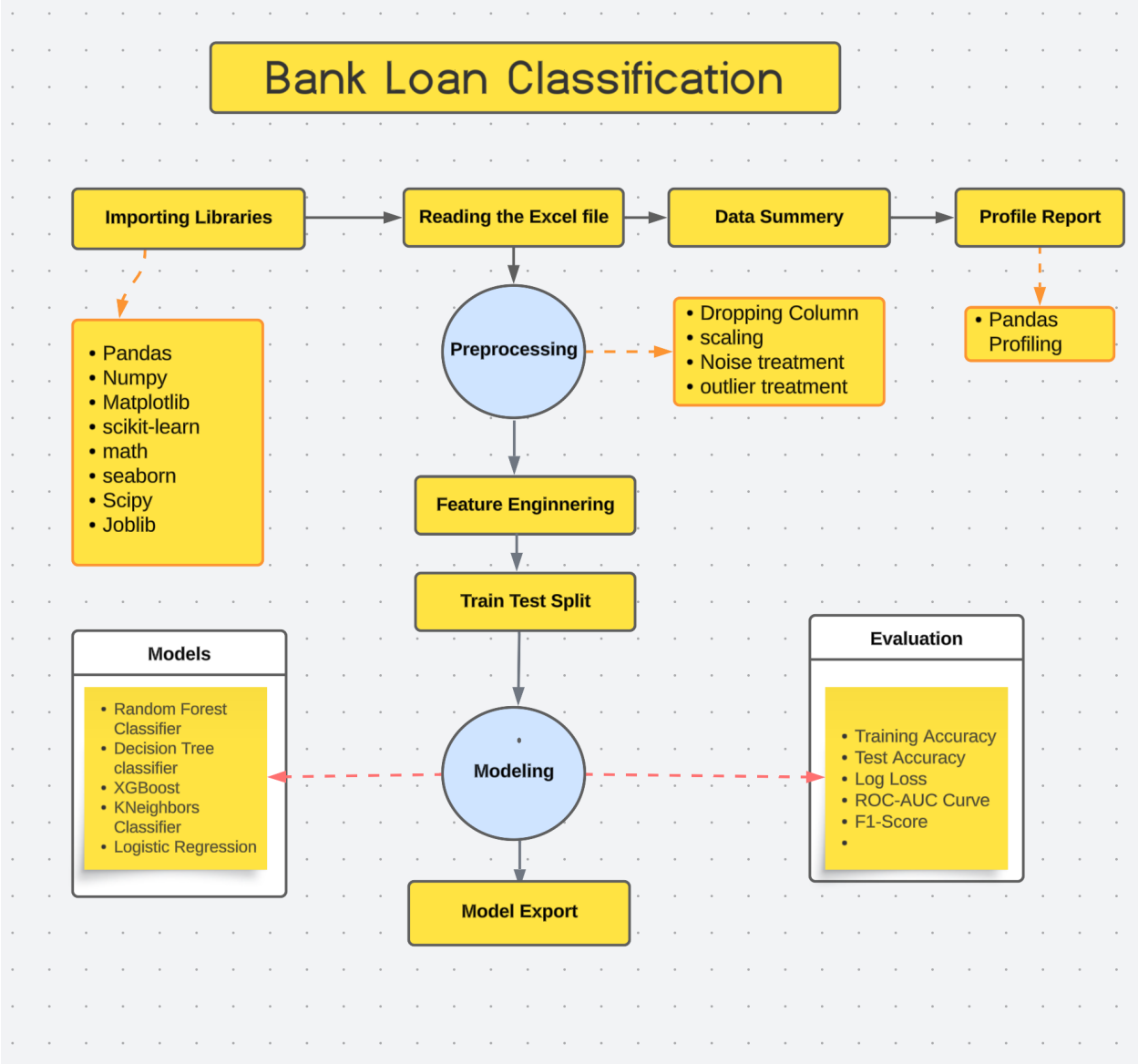
- ID: ID of the customer
- Age: Age of the customer
- Gender: M for Male, F for Female and O for Others
- Experience: Amount of work experience in years
- Income: Amount of annual income (in thousands)
- Home Ownership: Home Owner, Rent and Home Mortgage.
- Zip Code: Postal code in which the client lives
- Family: Number of family members
- CCAvg: Average monthly spending with the credit card (in thousands)
- Education: Education level (1: bachelor's degree, 2: master's degree, 3:advanced/professional degree)

- Mortgage: Value of home mortgage, if any (in thousands)
- Securities Account: Does the customer have a securities account with the bank?
- CD Account: Does the customer have a certificate of deposit account (CD) with the Bank?
- Online: Does the customer use the internet banking facilities?
- CreditCard: Does the customer use a credit card issued by the bank?
- Personal Loan: Did this customer accept the personal loan offered in the last campaign?
- (Target Variable)

Methodology:

- **Data Preprocessing:**
  - Perform data preprocessing steps, including data cleaning, handling missing values, and encoding categorical variables.
  - Create a preprocessing pipeline that includes a StandardScaler to scale the numerical features appropriately.
  - Split the dataset into training and testing sets.
- **Model Selection:**
  - Evaluate multiple classification models, including Random Forest Classifier, Decision Tree Classifier, XGBoost, KNeighborsClassifier, and Logistic Regression.
  - Train each model on the preprocessed training data and evaluate their performance using the specified metrics: Training Accuracy, Test Accuracy, Log Loss, ROC-AUC Curve, and F1-Score.
- **Model Evaluation and Comparison:**
  - Analyze the performance metrics obtained from the trained models.
  - Compare the results of each model based on their Training Accuracy, Test Accuracy, Log Loss, ROC AUC Score, and F1 Score.
  - Identify the best-performing model based on the evaluation metrics.
- **Export Preprocessing Pipeline and Selected Model:**
  - Export the preprocessing pipeline, including the StandardScaler, to ensure consistent data preprocessing for future predictions.
  - Save the selected model, XGBoost, along with the preprocessing pipeline using a serialization library like Joblib.
- **Performance on Unseen Data:**
  - Utilize the exported preprocessing pipeline and the selected XGBoost model to predict loan approvals on unseen data.
  - Evaluate the model's performance on the unseen data using accuracy and F1 Score as metrics.

Chart :



How to Use and Predict with the Trained Model:

To utilize the trained model for predicting loan approvals, follow the steps outlined below:

1. **Data Preparation:**
    - Make sure your input data includes the following columns: ID, Age, Experience, Income, ZIP Code, Family, CCAvg, Education, Mortgage, Personal Loan, Securities Account, CD Account, Online, and CreditCard.
    - Remove any columns that are not present in the above list, as they were not used during model training.
2. **Load Preprocessing Pipeline and Model:**
    - Load the preprocessing.pkl file, which contains the preprocessing pipeline, including the StandardScaler, used during training.

- Load the trained model file (e.g., xgboost_model.pkl) that was saved after model training.

3. **Preprocess the Data:**
   - Apply the preprocessing pipeline to the input data to ensure consistent scaling and preprocessing.
   - Use the loaded StandardScaler from the preprocessing pipeline to scale the numerical features in the data.
4. **Predict Loan Approvals:**
   - Feed the preprocessed data into the trained model for loan approval predictions.
   - Obtain the predicted loan approval results based on the model's classification.

Result:

The performance of various machine learning models was evaluated for the task of bank loan classification. The table below summarizes the results obtained from each model:

Model	Train Accuracy	Test Accuracy	Log Loss	ROC AUC Score	F1 Score
XGBoost	1.000	0.989	0.405	0.944	0.923
RandomForest	0.984	0.984	0.589	0.898	0.881
Decision Tree	1.000	0.982	0.663	0.934	0.878
KNeighbors Classifier	1.000	0.957	1.546	0.772	0.661
Logistic Regression	0.957	0.949	1.841	0.730	0.583

Among the evaluated models, XGBoost achieved the highest test accuracy of 0.989, along with a high F1 score of 0.923. It also demonstrated the lowest log loss of 0.405 and the highest ROC AUC score of 0.944. These results indicate that the XGBoost model performs exceptionally well in accurately classifying bank loan approvals.

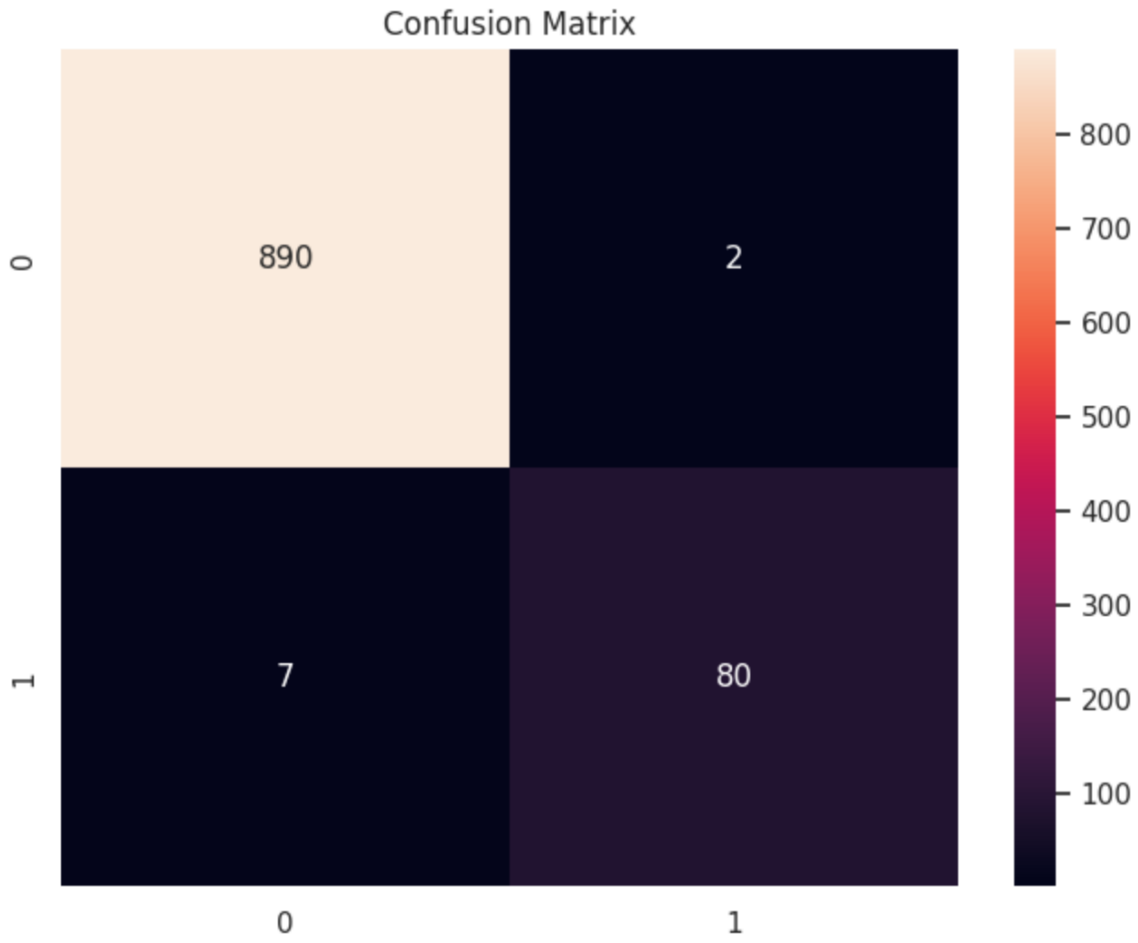
Performance in unseen data :

In [89]:
```
accuracy = metrics.accuracy_score(new_y_test, prediction)
f1_score_rfc2=metrics.f1_score(new_y_test, prediction)


print('Accuracy:', accuracy)
print('F1 Score:', f1_score_rfc2)
```

```
Accuracy: 1.0
F1 Score: 1.0
```

Image of confusion matrices:




```
<Figure size 1200x900 with 0 Axes>
```

Discussion:

The superior performance of the XGBoost model can be attributed to its ability to handle complex relationships within the data and effectively capture nonlinear patterns. The ensemble nature of XGBoost, combining multiple decision trees, enables it to effectively handle both numerical and categorical features, resulting in improved prediction accuracy.

The high training accuracy scores obtained by all models suggest that they were able to learn the training data well. However, it is important to consider the test accuracy as a more reliable measure of the models' generalization capabilities. XGBoost not only achieved the highest test accuracy but also demonstrated excellent performance in terms of the F1 score, which indicates a balance between precision and recall.

The logistic regression model, although having a lower accuracy and F1 score compared to XGBoost, still performed reasonably well. It is worth noting that logistic regression is a simpler model compared to XGBoost, which may explain the slightly lower performance. However, logistic regression can still be useful in scenarios where interpretability and model simplicity are crucial.

Conclusion:

In this project, we explored the task of bank loan classification using various machine learning models. Based on the evaluation results, the XGBoost model emerged as the top performer, exhibiting high accuracy and F1 score on the test dataset. The logistic regression model also showed promising results.

The development of a preprocessing pipeline, including the use of a StandardScaler, ensured consistent data scaling and preprocessing, contributing to the models' performance. The selected XGBoost model, along with the preprocessing pipeline, was exported and can now be utilized for predicting loan approvals on unseen data.

Summary:

Based on the evaluation results, the XGBoost model exhibited the highest accuracy and F1 Score on the test dataset. Therefore, it was selected as the best-performing model for loan classification. The preprocessing pipeline, including the StandardScaler, was saved to ensure consistent preprocessing of new data during predictions.

- ○