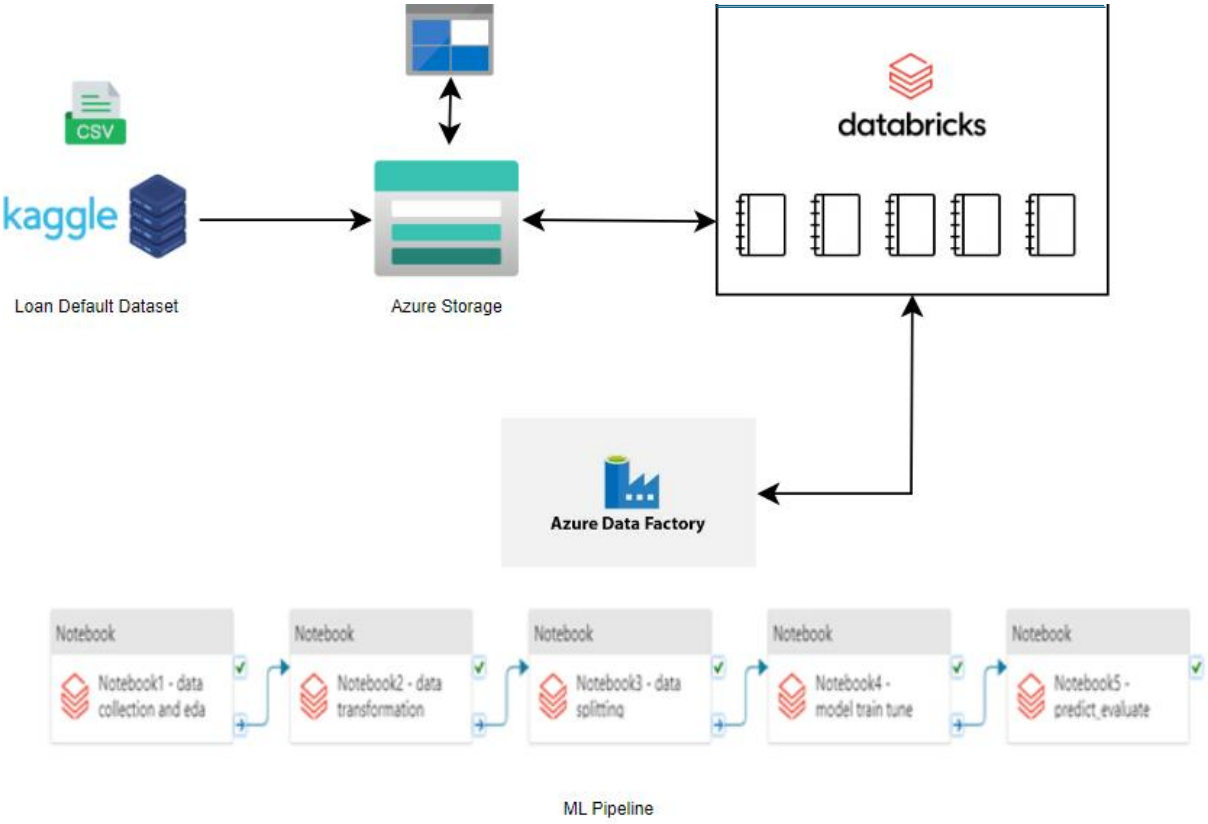
Project Topic: Loan Default Prediction

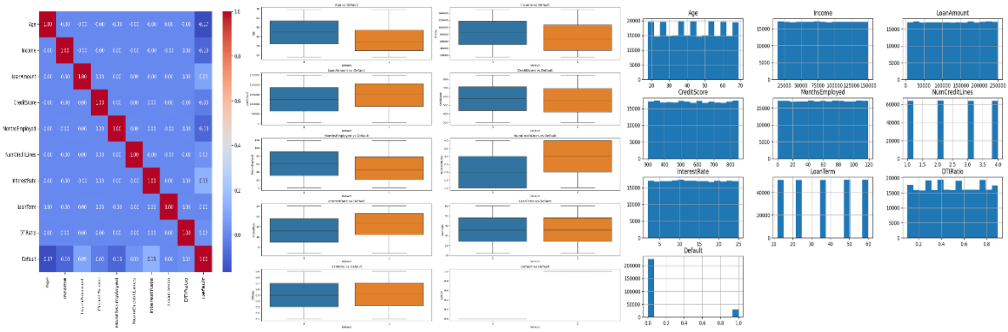
Project Overview:

Loan default prediction is a concern for financial institutions as it directly impacts their profitability and risk management strategies. By guessing these defaults correctly, banks can take early action to lower risks, make better loan decisions, and get better financial results. This project aims to build a predictive model capable of accurately identifying potential loan defaults before they occur.



1. **Data Collection:**
    - Connect to data sources- csv files from Kaggle to gather relevant data.
    - Dataset contains the following features - Age, Income, LoanAmount, CreditScore, MonthsEmployed, NumCreditLines , InterestRate, LoanTerm, DTIRatio, Education, EmploymentType, HasMortgage, HasDependents, LoanPurpose, HasCoSigner, Default.
    - The label is Default (0/1).
2. **Data Cleaning and Preprocessing:**
    - Use libraries like Pandas and NumPy for data manipulation.
    - Check for null values and missing values. Check for ranges of each feature values and the distribution in the data.
    - Understand the domain significance of each feature.
    - Check for the count of data points present in each class.

3. **Exploratory Data Analysis:**
    - Utilize visualization tools like Matplotlib and Seaborn.
    - Identify key features and correlations.
    - Create plots to see the relationships between the features and the label.
    - Generate the heatmap to depict the correlation between the dependent and independent variables.



4. **Feature Transformation:**
    - Identify what are the numeric and categorical columns and choose appropriate data encoding techniques for the same.
    - Choose what features are relevant to the prediction of the label and what can improve the model accuracy.
    - In this case all the categorical columns are encoded using one hot encoding. This is because the number of unique values is low i.e. We have low cardinality.
5. **Model Building/Tuning:**
    - Implement multiple models using Scikit-Learn, XGBoost, or similar libraries – Random Forest, Decision Trees, KNN, AdaBoost, Naïve bayes, SVM, Logistic regression.
    - Use GridSearchCV for hyperparameter tuning and tweak the model parameters to improve accuracy.
    - Used parameter ‘balanced’ to handle the label imbalance present in the data.
    - Choose the model with the best accuracy and save it as a pickle file in a storage container (Model Artifactory).
    - From all the models, AdaBoost gave the best accuracy, so it was chosen for the next steps.
6. **Model Predictions and Evaluation:**
    - Use cloud platforms (AWS, Azure) for deployment and scaling.
    - Use the stored trained model in the container blob storage and use it to make the predictions on the test dataset.
    - Generate relevant metrics like accuracy, precision, confusion matrix etc.
7. **Pipelining and Automation:**
    - Set up automated training pipeline using Azure Data Factory.
    - Regularly update the model with new data and monitor performance.

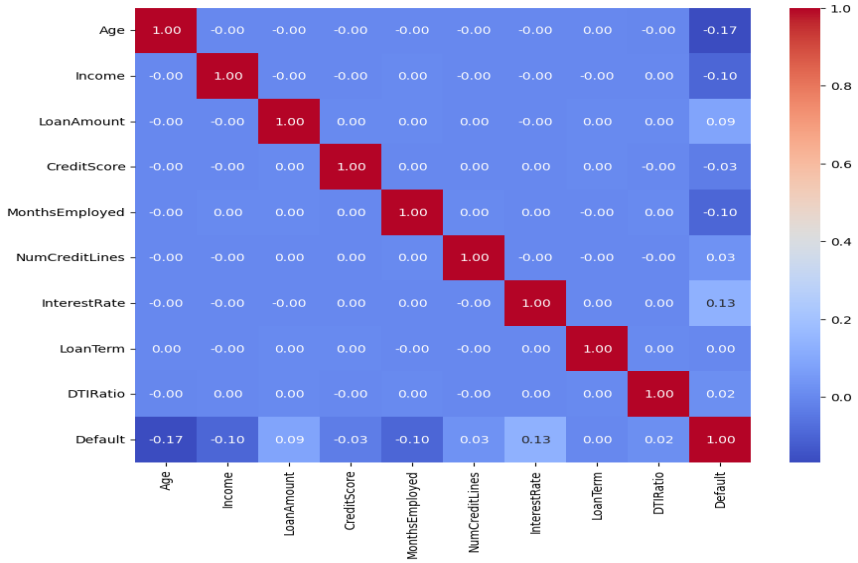
Expected Outcomes:

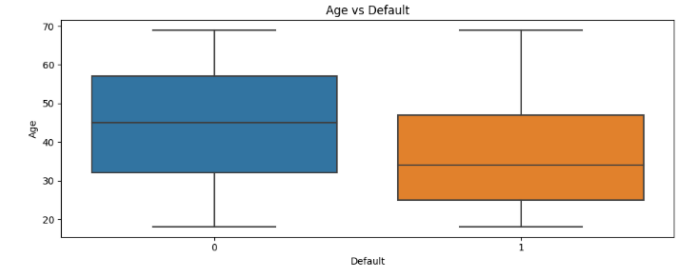
- Reduced Loan Defaults: By predicting which loans are likely to default, the institution can take proactive measures to mitigate risk, such as offering revised payment plans, increasing monitoring, or even denying high-risk loans.
- Enhanced Risk Management: With a predictive model, the bank can better manage the risk profile of its loan portfolio.
- Increased Financial Stability

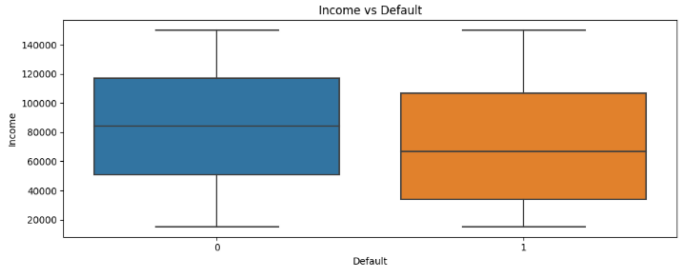
Tools and Technologies:

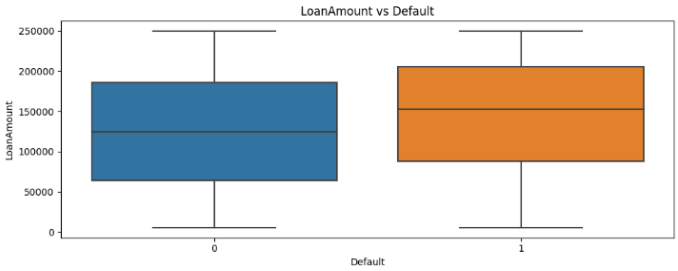
- **Programming Languages:** Python, SQL, Spark
- **Libraries:** Pandas, NumPy, Scikit-Learn, XGBoost, Matplotlib, Seaborn, Pickle
- **Deployment:** Azure Databricks, Azure Data Factory, Azure storage container
- **Cloud Platforms:** Azure

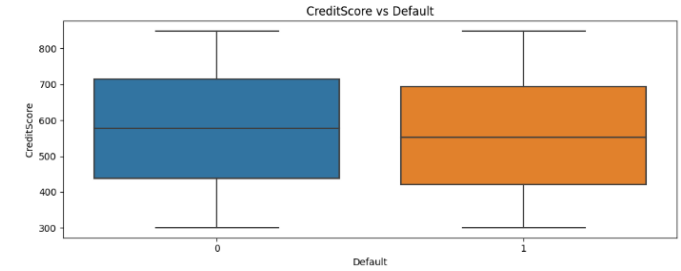
Appendix:

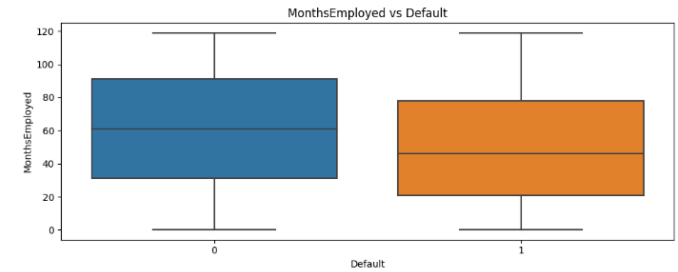

Age 1.00 -0.00 -0.00 -0.00 -0.00 -0.00 -0.00 0.00 -0.00 -0.17
Income -0.00 1.00 -0.00 -0.00 0.00 -0.00 -0.00 -0.00 0.00 -0.10
LoanAmount -0.00 -0.00 1.00 0.00 0.00 0.00 -0.00 0.00 0.00 0.09
CreditScore -0.00 -0.00 0.00 1.00 0.00 0.00 0.00 0.00 -0.00 -0.03
MonthsEmployed -0.00 0.00 0.00 0.00 1.00 0.00 0.00 -0.00 0.00 -0.10
NumCreditLines -0.00 -0.00 0.00 0.00 0.00 1.00 -0.00 -0.00 -0.00 0.03
InterestRate -0.00 -0.00 -0.00 0.00 0.00 -0.00 1.00 0.00 0.00 0.13
LoanTerm 0.00 -0.00 0.00 0.00 -0.00 -0.00 0.00 1.00 0.00 0.00
DTIRatio -0.00 0.00 0.00 -0.00 0.00 -0.00 0.00 0.00 1.00 0.02
Default -0.17 -0.10 0.09 -0.03 -0.10 0.03 0.13 0.00 0.02 1.00
Age
Income
LoanAmount
CreditScore
MonthsEmployed
NumCreditLines
InterestRate
LoanTerm
DTIRatio
Default
1.0
0.8
0.6
0.4
0.2
0.0

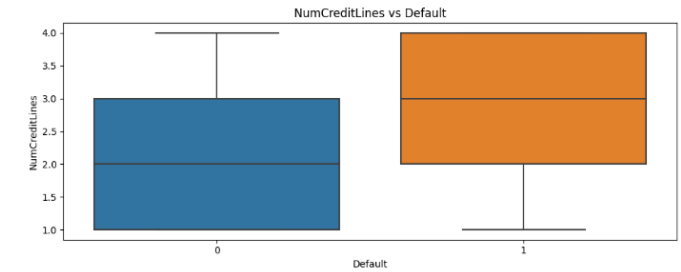

Age vs Default
Age
Default

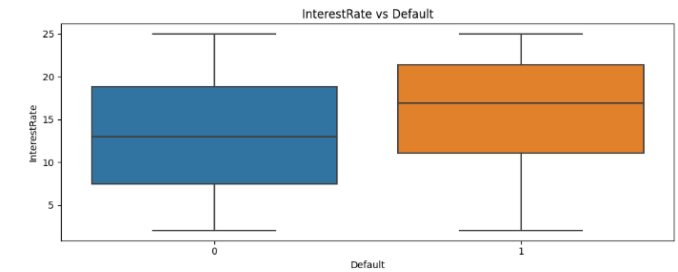

Income vs Default
Income
Default

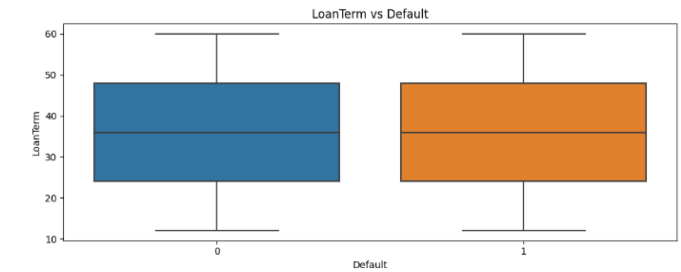

LoanAmount vs Default
LoanAmount
Default

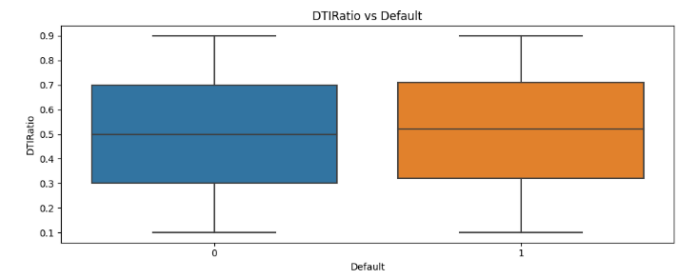

CreditScore vs Default
CreditScore
Default

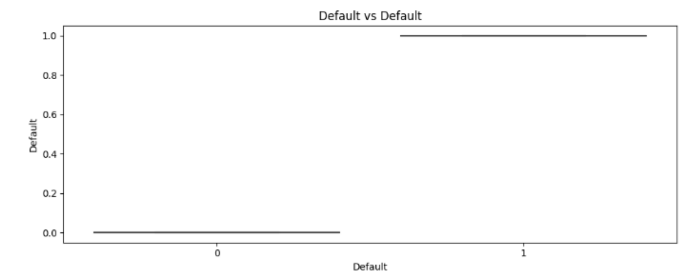

MonthsEmployed vs Default
MonthsEmployed
Default

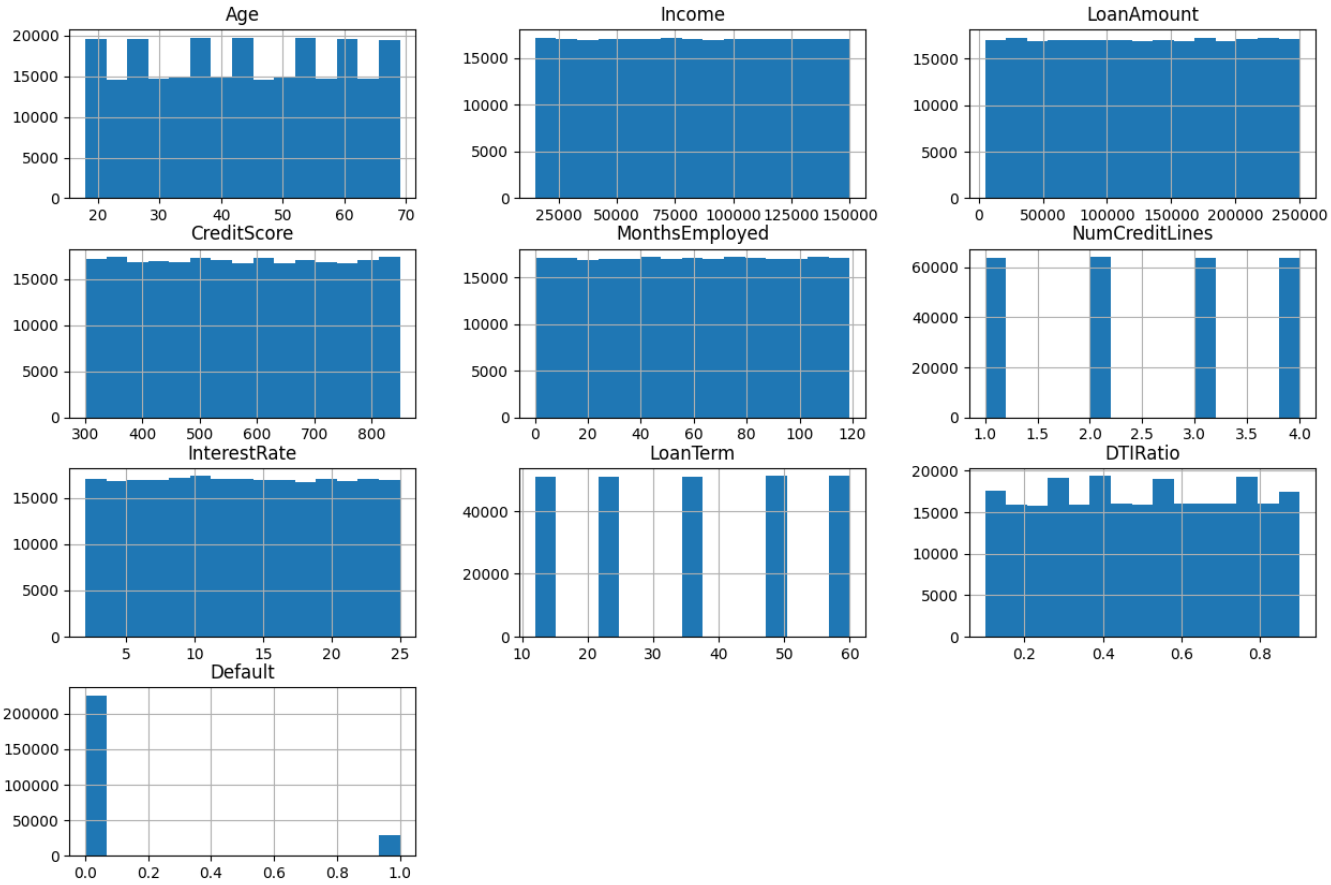

NumCreditLines vs Default
NumCreditLines
Default

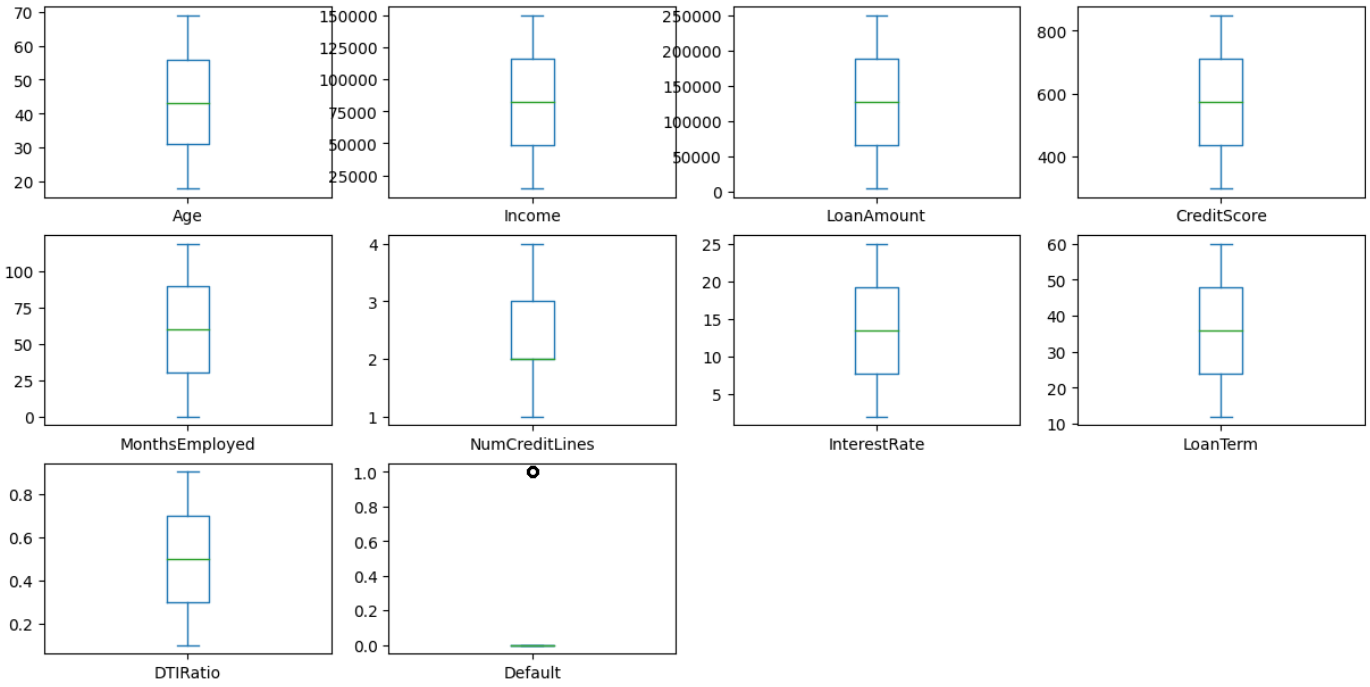

InterestRate vs Default
InterestRate
Default

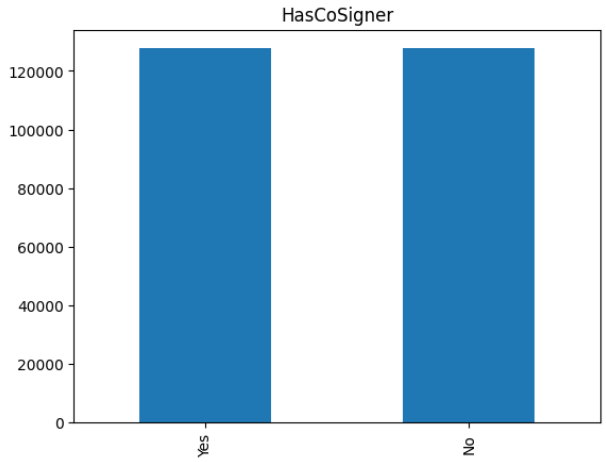

LoanTerm vs Default
LoanTerm
Default

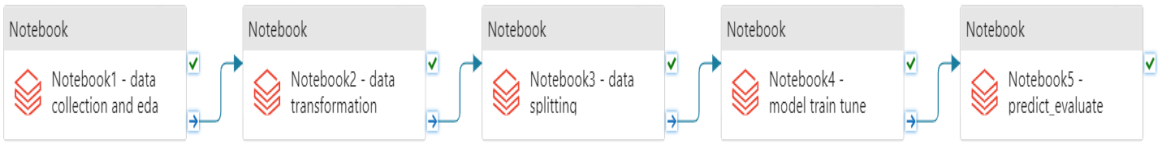

DTIRatio vs Default
DTIRatio
Default


Default vs Default
Default
Default


Age
Income
LoanAmount
CreditScore
MonthsEmployed
NumCreditLines
InterestRate
LoanTerm
DTIRatio
Default


Age
Income
LoanAmount
CreditScore
MonthsEmployed
NumCreditLines
InterestRate
LoanTerm
DTIRatio
Default






```
Accuracy: 0.8858037987076561
Confusion Matrix:
 [[44943   196]
 [ 5636   295]]
Classification Report:
              precision    recall  f1-score   support

           0       0.89      1.00      0.94     45139
           1       0.60      0.05      0.09      5931

    accuracy                           0.89     51070
   macro avg       0.74      0.52      0.52     51070
weighted avg       0.86      0.89      0.84     51070
```

```
#  Column                 Non-Null Count   Dtype
---  ------                 --------------   -----
 0  Age                    255347 non-null  int32
 1  Income                 255347 non-null  int32
 2  LoanAmount             255347 non-null  int32
 3  CreditScore            255347 non-null  int32
 4  MonthsEmployed         255347 non-null  int32
 5  NumCreditLines         255347 non-null  int32
 6  InterestRate           255347 non-null  float64
 7  LoanTerm               255347 non-null  int32
 8  DTIRatio               255347 non-null  float64
 9  Default                255347 non-null  int32
 10  Education_High School       255347 non-null  Sparse[float64, 0]
 11  Education_Master's          255347 non-null  Sparse[float64, 0]
 12  Education_PhD               255347 non-null  Sparse[float64, 0]
 13  EmploymentType_Part-time    255347 non-null  Sparse[float64, 0]
 14  EmploymentType_Self-employed  255347 non-null  Sparse[float64, 0]
 15  EmploymentType_Unemployed   255347 non-null  Sparse[float64, 0]
 16  MaritalStatus_Married       255347 non-null  Sparse[float64, 0]
 17  MaritalStatus_Single        255347 non-null  Sparse[float64, 0]
 18  HasMortgage_Yes             255347 non-null  Sparse[float64, 0]
 19  HasDependents_Yes           255347 non-null  Sparse[float64, 0]
 20  LoanPurpose_Business        255347 non-null  Sparse[float64, 0]
 21  LoanPurpose_Education       255347 non-null  Sparse[float64, 0]
```

22	LoanPurpose_Home	255347 non-null	Sparse[float64, 0]
23	LoanPurpose_Other	255347 non-null	Sparse[float64, 0]
24	HasCoSigner_Yes	255347 non-null	Sparse[float64, 0]