Credit Level Prediction: A Machine Learning Approach

Problem Statement

Low Credit Level

The customers that are likely to default on their credit obligations. It is for this reason that they may have a short credit history, erratic earnings or have been in a position to make any payments in time.

Medium Credit Level

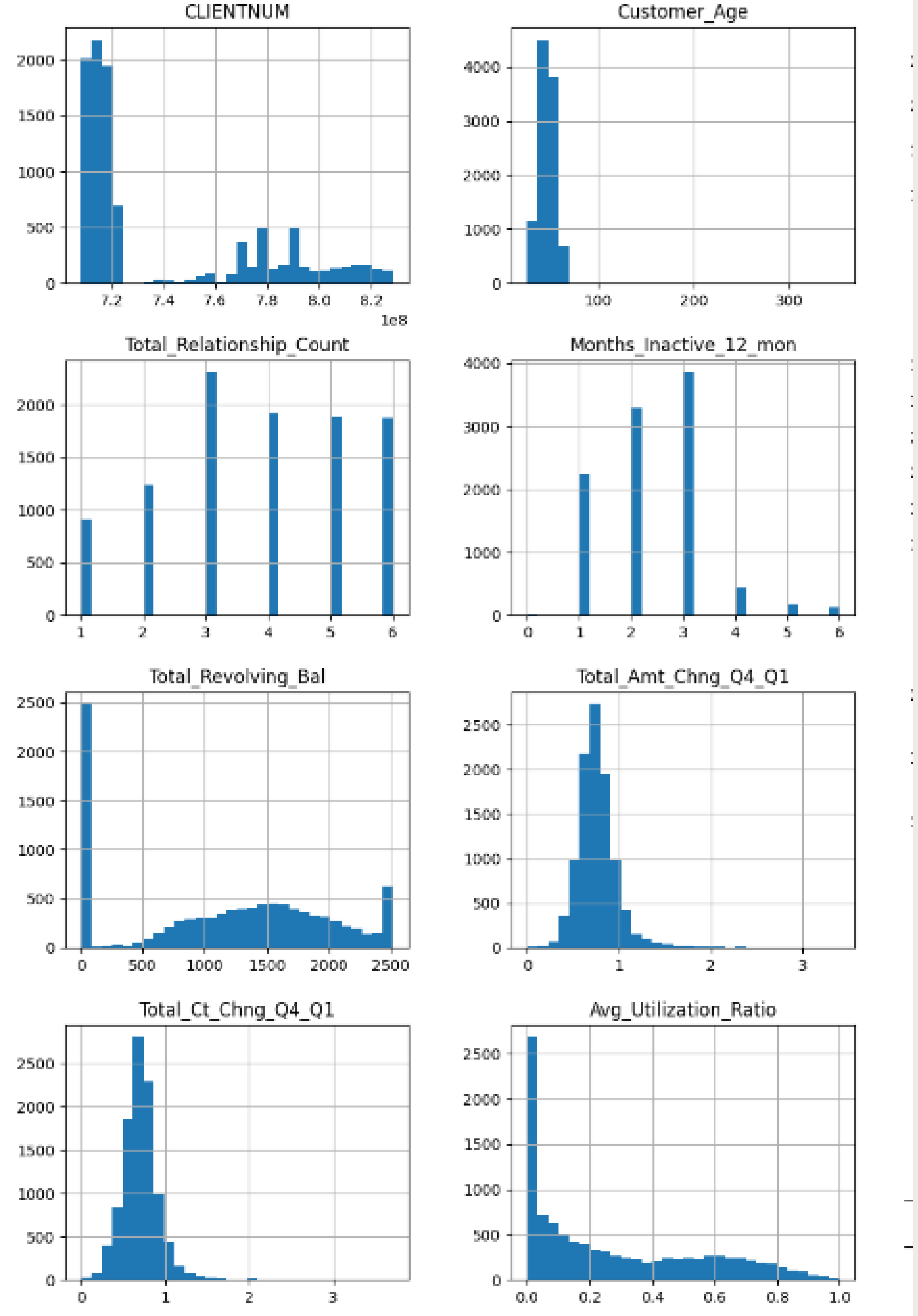
Customers with moderate creditworthiness. They may have a poor credit record, a steady income source and a satisfactory credit performance score.

High Credit Level

Some examples of target customers include customer with a good credit rating, good credit score and behavioral patterns. They have fixed and stable income, high repayment rate and low usage rate.

Dataset Overview

Feature	Description	Relevance to Credit Prediction
Age	Customer's age in years	Age is something that can be associated with financial maturity, stable income, and tolerance to risks.
Gender	Customer's gender	Gender can also shape expenditure and savings and investment behaviours in unseen ways.
Income Category	Categorical representation of income level	Income level is a good marker of creditworthiness and credit repay ability among the loan seeking public.
Credit Limit	Maximum amount of credit available to the customer	Higher credit limits normally suggest a higher level of trust and a lower risk factor.
Total Transaction Amount	Sum of all transactions made by the customer	Illustrates the customer’s total expenditure and financial transactions.



Data Exploration and Analysis

1 **Summary Statistics**

Mean, median, standard deviation, and quartiles for numerical features were calculated to understand the central tendency, spread, and presence of outliers.

2 **Histograms**

Histograms provided a visual representation of the frequency distribution of numerical features, revealing patterns and skewness in the data.

3 **Pairplots**

Pairplots illustrated relationships between pairs of features, uncovering potential correlations and identifying patterns that might influence credit level predictions.

Data Cleaning and Preprocessing

Missing Data

For handling of missing values, imputation techniques were used whereby median was used for numerical features and mode for categorical features.

Outlier Handling

We were also able to handle outliers through procedures like trimming or capping so as not to influence the model results.

Data Transformations

Features were normalized and one hot encoded in order to bring all features into similar scale to be used with machine learning algorithms.

Feature Engineering

1

Combined Features

Existing features were combined to create new features that provided more meaningful information. For example, 'Avg_Trans_Amt' was derived from 'Total_Trans_Amt' and 'Total_Trans_Ct'.

2

Interaction Terms

The interaction terms were formed by squaring or multiplying two or more features with the aim of estimating the impact they have on the target variable. This it useful in modeling non-linear relationships.

3

Ratio Features

Ratios were calculated to capture relative relationships between features. For example, 'Utilization_Ratio' was derived from 'Total_Revolving_Bal' and 'Credit_Limit'.

Modeling Strategy

1

Model Selection

This classifier was chosen because Random Forest is known for it high resistance to overfitting, its capacity to work with numerical and categorical data, suitable for classification and is relatively accurate and computationally efficient in comparison with other mathematical models.

2

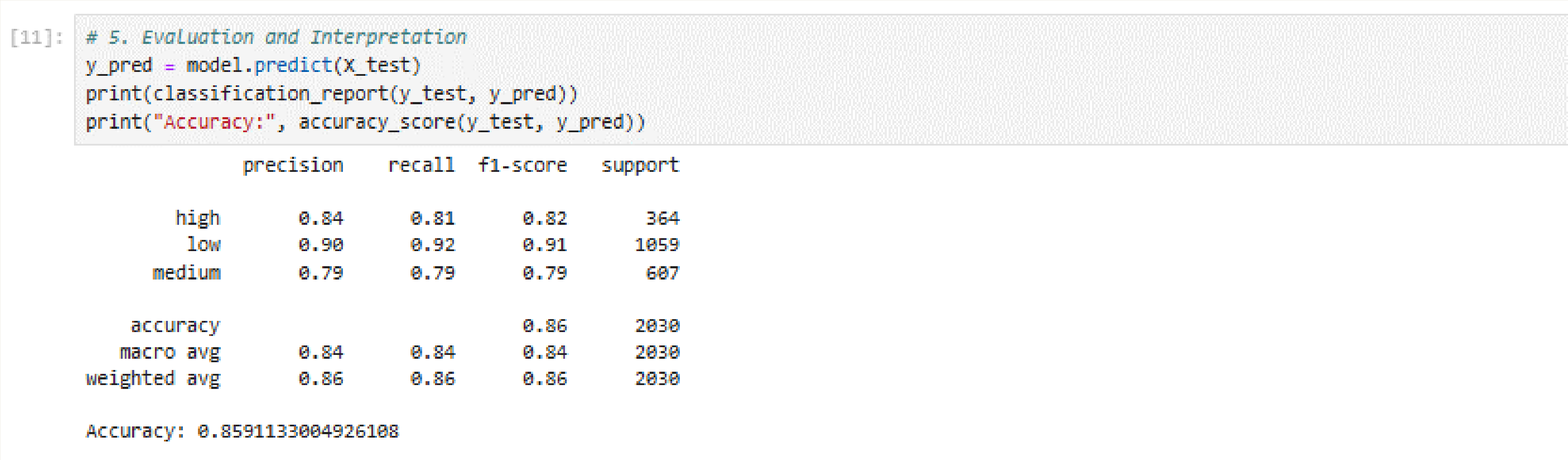
Model Setup and Training

The dataset was then divided into training set which contained 80% of the data and the testing set which contained the remaining 20%. The training process used k-fold cross validation in order to reduce the likelihood of overfitting thereby making the models more stable.

3

Validation Approach

Performance of the model was evaluated based on the results of the training and testing phase to determine the possible modifications for the model



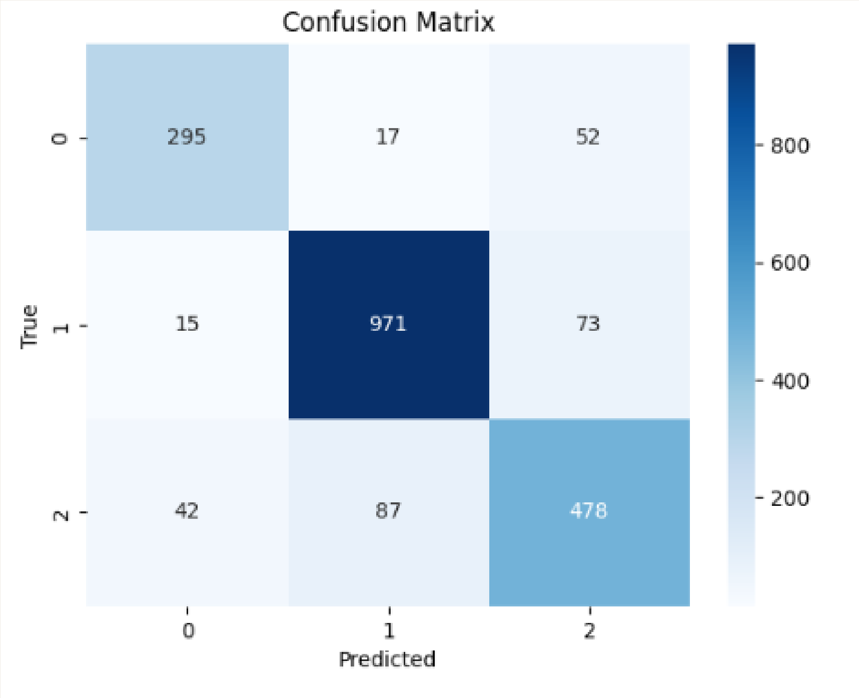
```
[11]: # 5. Evaluation and Interpretation
      y_pred = model.predict(X_test)
      print(classification_report(y_test, y_pred))
      print("Accuracy:", accuracy_score(y_test, y_pred))
```

```
              precision    recall  f1-score   support

        high       0.84      0.81      0.82       364
         low       0.90      0.92      0.91      1059
      medium       0.79      0.79      0.79       607

    accuracy                           0.86      2030
   macro avg       0.84      0.84      0.84      2030
weighted avg       0.86      0.86      0.86      2030

Accuracy: 0.8591133004926108
```



Model Evaluation

Accuracy

The proportion of instances which are correctly classified.

Precision

The ratio of the instances that belong to the specific class to the total number of instances that were predicted to be of that class.

Recall

The percentage of the number of cases of a particular class that has been classified correctly in relation to the total number of cases of that particular class.

F1-Score

The F-measure of precision and recall, which provides the average of both the parameters.

Feature Importance Analysis

1

Average Utilization Ratio

Measures the ratio of credit card balance to the total credit limit. It reflects how much credit is being used compared to the available credit. A high utilization ratio may indicate higher credit risk, affecting a customer's ability to repay.

2

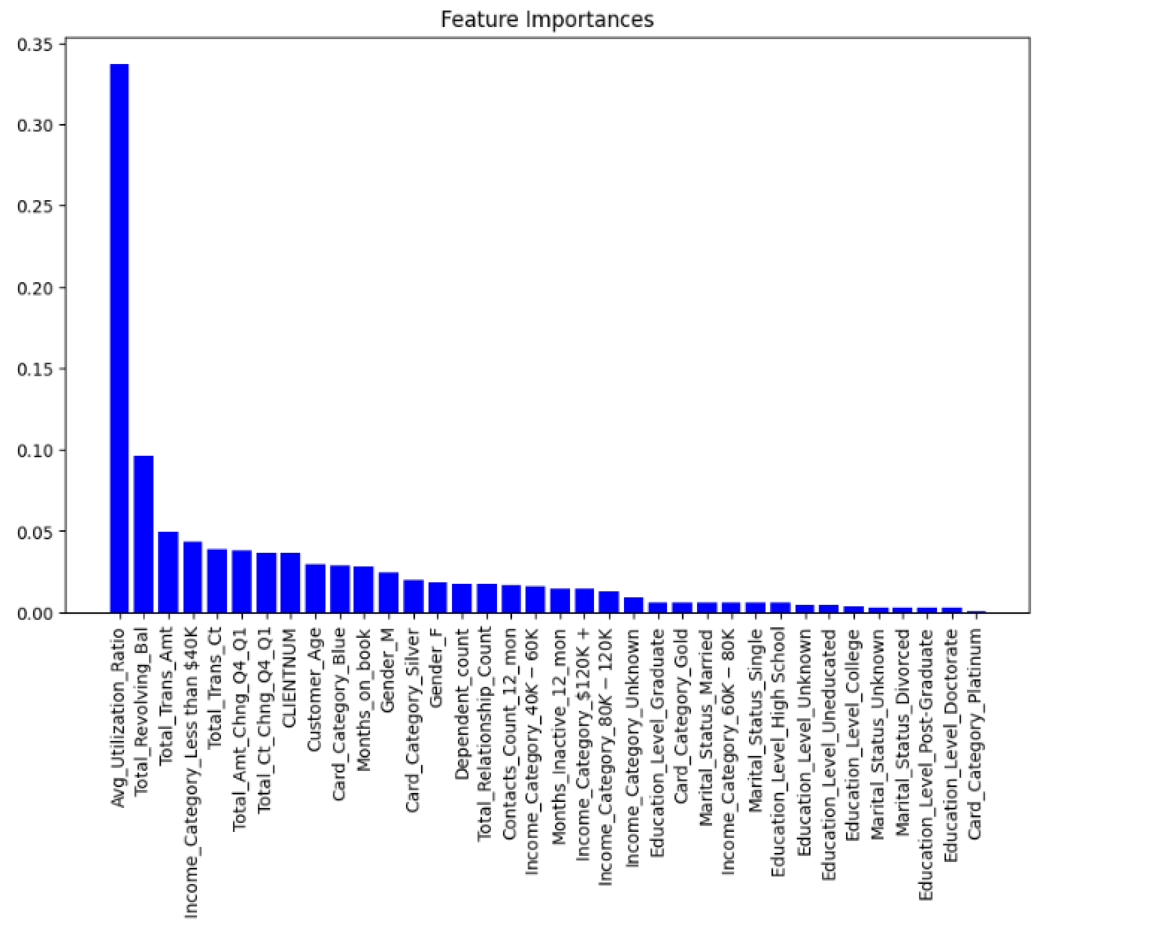
Total Revolving Balance

Having high revolving balances might suggest that customers rely more on credit and might pose more risk and thus pose a threat to predictions towards median or low credit scores.

3

Total Transaction Amount

Represents the total value of transactions (purchases, payments, etc.) made within a certain time frame. High transaction amounts could signify high income or financial activity, while low amounts might suggest lower income or spending power.



Challenges and Learnings

Skewed Data and Overfitting

Balance and Simplicity

Cross-Validation and Ensembles

Conclusion

1 **Model Accuracy**

The accuracy of the proposed model was approximately 86 %; thus, the model can be used to classify the customers into the correct credit level efficiently.

2 **Model Limitations**

The performance of the model depends on the quality of the data, though the model can get trained from time to time to capture changes in the economic environment and customer behavior.

3 **Future Directions**

More work in this line can include endeavours to use real time data, research other methods of algorithms that can be used for the assessment and the inclusion of other unstructured sources of data for an enhanced risk evaluation.

Recommendations and Future Work

1 **Real-Time Data Integration**

Integrating real-time data feeds into the model training process would allow for continuous updates, ensuring its relevance and accuracy in reflecting current market conditions.

2 **Alternative Algorithms**

Exploring alternative algorithms, such as deep learning techniques, could potentially lead to improved performance or faster predictions suitable for real-time analysis.

3 **Unstructured Data Integration**

Incorporating unstructured data sources, such as customer interaction logs, social media activity, and sentiment analysis, could provide a more comprehensive understanding of customer behavior and risk profiles.

Special Focus: Innovations and Creativity

Custom Outlier Detection

Dynamic Imputation Technique

Impact of Innovations

Foundation for Future Innovations

Thank You