REPORT ON

LOAN PREDICTION

By

Nitin

https://www.linkedin.com/in/nitin-b57867168/

Index

Introduction

The objective of this project was to develop a machine learning model to predict whether a loan application should be approved or rejected based on various applicant details, such as income, credit history, employment status, and demographic factors. The goal was to create an efficient and accurate predictive system to streamline the loan approval process for financial institutions.

Problems:

Q-1. What problem are you solving?

The problem at hand is predicting whether a loan application should be approved or rejected based on various applicant details, such as income, credit history, employment status, and demographic factors. This classification problem is crucial for financial institutions aiming to streamline their loan approval process in real time.

Q-2. Why is it worth solving?

Loan approval decisions are critical in the banking and financial sector, impacting both lenders and applicants. An efficient and accurate predictive model can help:

- Reduce manual processing time and decision-making errors.
- Minimize the risk of loan defaults by identifying high-risk applicants.
- Ensure fair and consistent loan approval decisions.
- Improve customer experience by providing quicker responses.

Q-3. What is the source of your data and what kinds of data are you using?

The dataset used for this study originates from a publicly available dataset on Kaggle https://www.kaggle.com/datasets/devzohaib/eligibility-prediction-for-loan , specifically designed for loan prediction classification tasks. The data consists of:

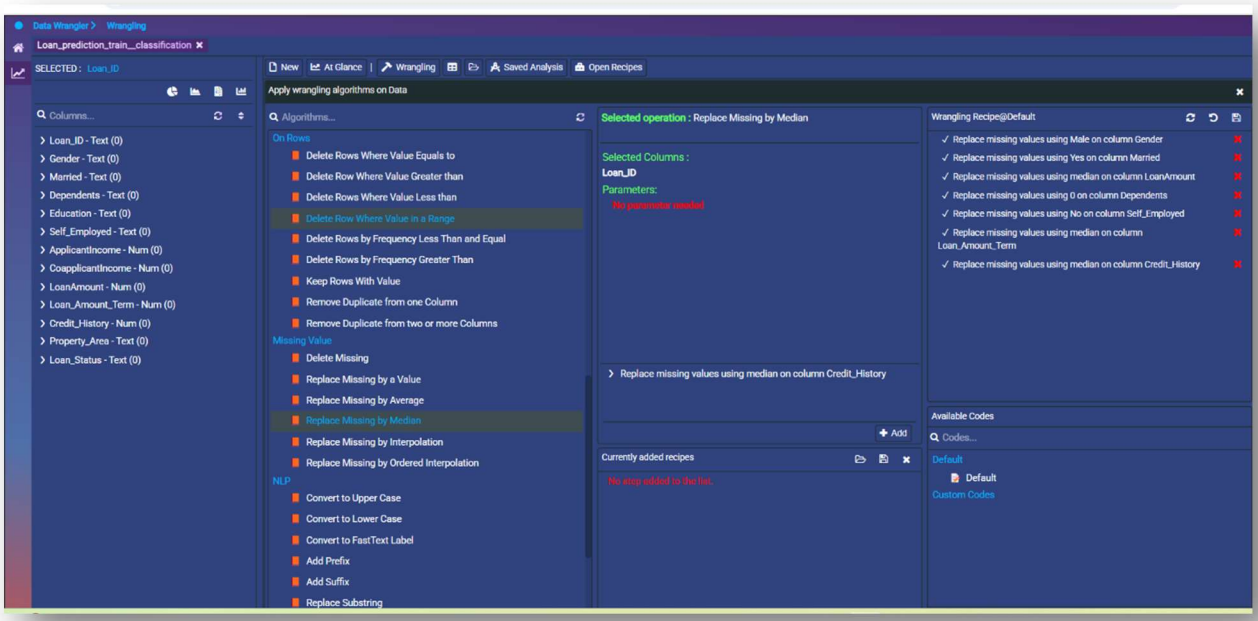
- **Categorical variables**:
    - **i.** Nominal: Property Area
    - **ii.** Binary: Gender, Marital Status, Education, Self-Employed, Credit History, Loan Status (target variable: Y/N)
- **Numerical variables**: Dependents, Applicant Income, Coapplicant Income, Loan Amount and Loan Term.

- **Missing values**: Gender (13%), Dependent (15%), Self_Employed (32%), LoanAmount (22%), Loan_Amount_Term (14%), and Credit History (50%), contain missing values that require handling during preprocessing.

Methodology: (Supervised Learning – Classification)

Q-1. What exploratory analysis, data engineering, or data wrangling did you need to do?

Exploratory Analysis & Data Wrangling



Handling Missing Values (Imputation)

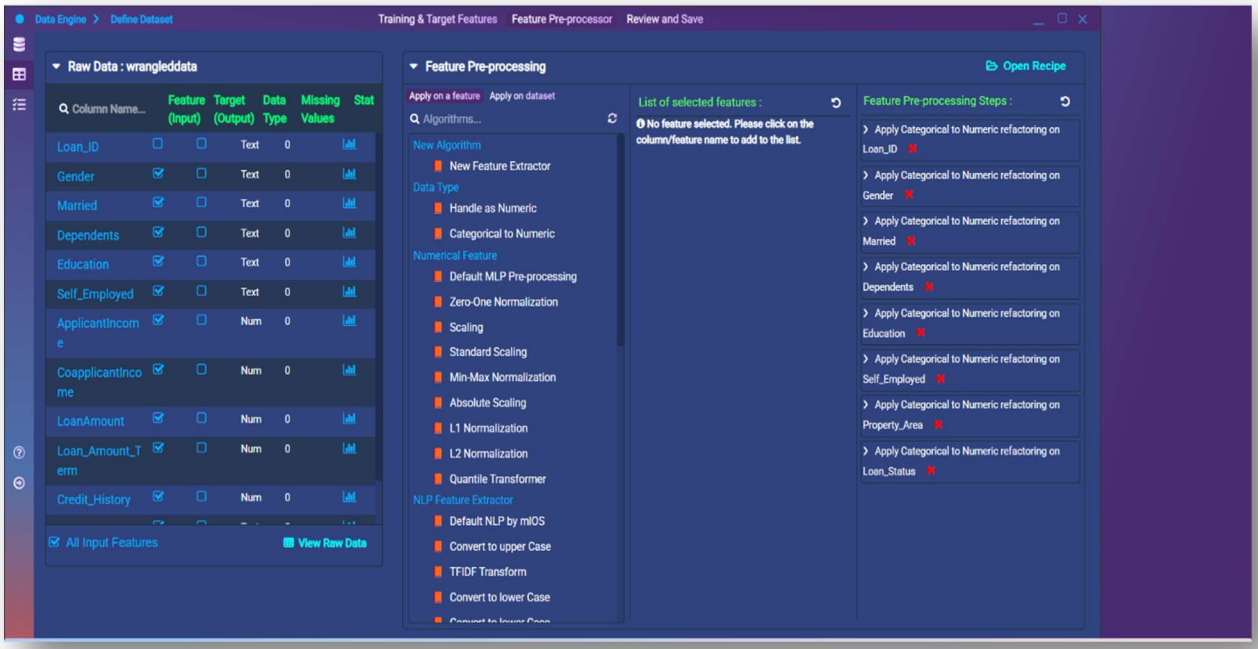
- **Missing values were replaced using different strategies**:
    - **Mode Imputation (Most Frequent Value)** for categorical features:
        - Gender → Filled with "Male"
        - Married → Filled with "Yes"
        - Self_Employed → Filled with "No"
    - **Median Imputation** for numerical features:

- LoanAmount → Filled with the median loan amount.
- Loan_Amount_Term → Filled with the median loan term.
- Credit_History → Filled with the median credit history score.

Justification:

- **Mode for categorical variables** ensures consistency with the most common value.
- **Median for numerical variables** prevents the influence of extreme values (outliers)

Q-2. How did you prepare the data for modeling?



Feature Selection & Defining Inputs/Outputs

- The **input features** (Feature (Input)) were selected from the cleaned dataset (wrangleddata).
- **Loan_Status** (Target variable) is identified for classification modeling.
- **Loan_ID** is excluded since it's just an identifier and doesn't contribute to predictions.

Why?

- Selecting relevant features ensures better model performance and avoids overfitting.

2. Handling Missing Values

- **All missing values have been handled**, as indicated by the **"Missing Values: 0"** column.
- Missing values were imputed in the **data wrangling step** (as seen in previous images).

Why?

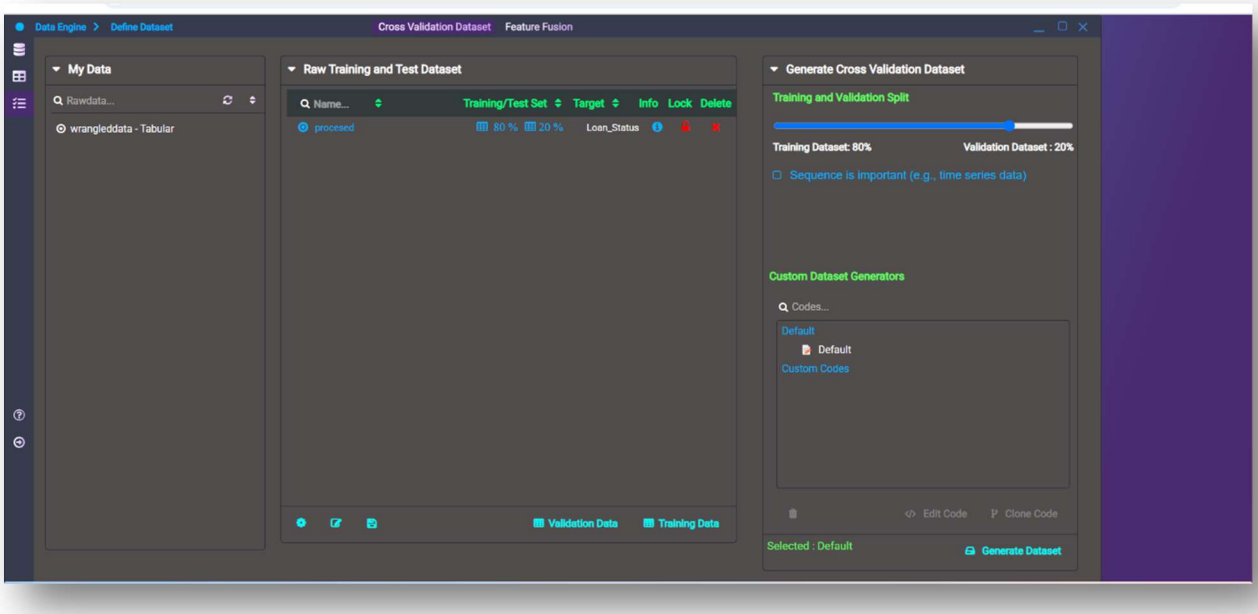
- Ensuring no null values prevents errors in training machine learning models.

3. Categorical to Numeric Transformation

- The images show that categorical columns were converted into numerical format using encoding techniques.
    - **Applied categorical-to-numeric transformations on:**
        - Gender
        - Married
        - Dependents
        - Education
        - Self_Employed
        - Property_Area
        - Loan_Status (Target Variable)

Why?

- Machine learning models require numerical inputs, so categorical features must be converted



1. Data Import and Selection:

- You imported the dataset labeled as wrangleddata - Tabular.

- The dataset appears to have undergone some preprocessing, as it's labeled "wrangleddata," indicating cleaned or transformed data.

2. Target Variable Assignment:- The target variable for the modeling task is set as Loan_Status, which is probably a classification label indicating loan approval status.

3. Training and Validation Split: - You defined an 80% training and 20% validation split. This ensures that most of the data is used for training while reserving a portion for evaluating the model's performance.

4. Sequence Importance (Unchecked): - The option for "Sequence is important (e.g., time series data)" is not selected. This suggests that the dataset is not time-dependent, or the order of rows does not affect model performance.

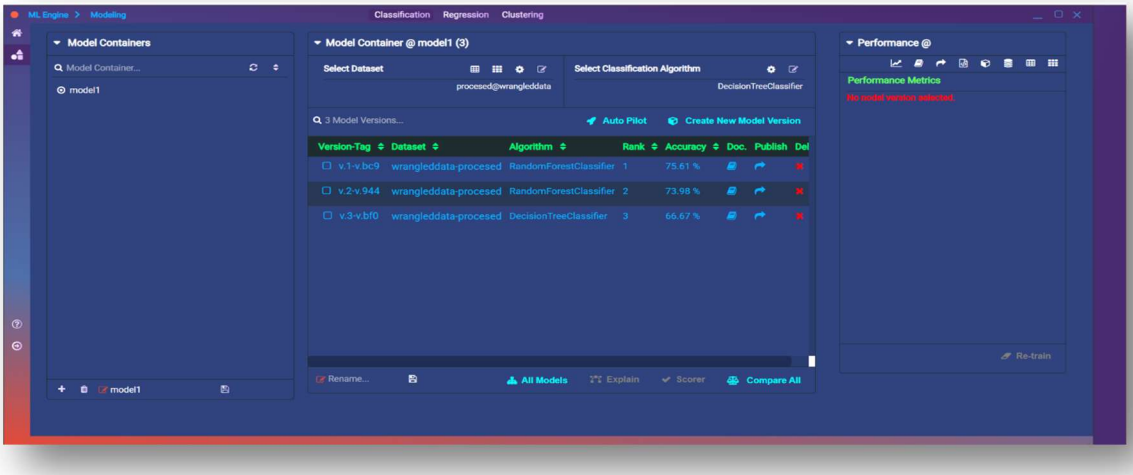
5. Cross-Validation Dataset Generation: - You used the “Default Dataset Generator” to split the data, ensuring standard cross-validation preparation.

If further steps were taken or specific feature engineering methods were applied, let me know for more detailed documentation!

Data Preparation for Modeling

Post-cleaning, categorical variables were encoded appropriately, and numerical variables were normalized. This ensured consistency and compatibility with the selected machine learning models.

Q-3. What was your modeling process? Specifically, which algorithms and parameters did you use and why?



Modeling Process:
Dataset: Used the processed dataset wrangleddata after preprocessing.

Algorithms:

Random Forest Classifier:

Version 1 (v1-bc9): Best performance with 75.61% accuracy.

Version 2 (v2-944): Slightly lower accuracy at 73.98%.

Parameter: n_estimators = 200.

Decision Tree Classifier (v3-bf0): Lowest accuracy at 66.67%.

Reasoning*:*

Random Forest: Chosen for its robustness, ability to handle mixed data types, and reduced overfitting through ensemble learning.

Decision Tree: Used for comparison due to its simplicity but performed worse.

Conclusion: The Random Forest Classifier (v1-bc9) was the best model with the highest accuracy, making it the preferred choice.

Result:

Q-1. What were your results?

- The best-performing model was the **Random Forest Classifier (v1-bc9)** with an accuracy of **75.61%**.
- The second version of Random Forest had **73.98% accuracy**.
- The **Decision Tree Classifier (v3-bf0)** had the lowest accuracy at **66.67%**.

Q-2. How did you evaluate the performance of your model? What metrics did you use?

- **Accuracy** was the main metric used to evaluate the models.
- The **Random Forest Classifier** was chosen for its robustness, ability to handle mixed data types, and reduced overfitting through ensemble learning.
- The dataset was split into **80% training and 20% validation** to assess performance.

Conclusions

- The Random Forest Classifier (v1-bc9) was the most effective model for loan prediction.
- The Decision Tree Classifier underperformed compared to Random Forest, showing its limitations in handling the dataset.

Q-1. What improvements would you like to make in the future?

- Fine-tuning hyperparameters in the **Random Forest model** to potentially increase accuracy.
- Exploring additional features or external datasets to enhance predictive power.
- Testing **other advanced models** like XGBoost or Neural Networks.
- Improving the **handling of missing values** to minimize data loss.

Q-2. How do you think the solution could be used in real life?

- The model can be integrated into a **loan approval system** to automate decision-making.
- It can help banks and financial institutions **reduce manual processing time** and improve efficiency.
- The model can be used for **risk assessment**, ensuring fair and consistent loan approvals.

Q-3. What value do you think the solution will have to the client?

- Faster loan processing times, improving customer experience.
- More **accurate risk assessment**, reducing the chance of defaults.
- Ensuring **fair lending practices** by removing human biases in decision-making.

Q-4. What did you learn through this project?

- How to handle **missing data effectively** using imputation techniques.
- How to preprocess **categorical and numerical variables** for machine learning.
- The importance of **choosing the right model** for classification problems.
- How **Random Forest** outperforms **Decision Trees** in predictive accuracy.
- The impact of **data quality and feature selection** on model performance.