Introduction

This report details the steps taken during the data analysis and modelling process, including basic checks, exploratory data analysis (EDA), data pre-processing, and model building. Additionally, it presents the performance of multiple machine learning models and recommends the best model for production use.

1. Basic Checks

The initial step involved loading the dataset and performing basic checks to understand its structure and characteristics. This included:

- Checking the shape of the dataset.

- Displaying the first few rows to get a sense of the data.

- Summarizing the data to understand the distribution of numerical and categorical features.

2. Checking Missing Values

Next, the dataset was checked for missing values. Missing values can significantly impact the performance of machine learning models. The following steps were taken:

- Identified columns with missing values.

- Decided on appropriate techniques to handle missing data, such as imputation or removal, based on the percentage of missing values and the importance of the features.

3. Exploratory Data Analysis (EDA)

a. Univariate Analysis

Univariate analysis was conducted to understand the distribution of individual features. This included:

- Plotting histograms and boxplots for numerical features.

- Plotting bar charts for categorical features.

b. Bivariate Analysis

Bivariate analysis was conducted to understand the relationships between pairs of features. This included:

- Scatter plots for numerical feature pairs.

- Box plots and bar charts for categorical vs. numerical features.

- Correlation matrix to identify the strength of relationships between numerical features.

4. Conversion of Categorical Columns into Numericals

To prepare the data for machine learning algorithms, categorical columns were converted into numerical values. This was done using techniques such as:

- Label encoding for ordinal categorical features.

- One-hot encoding for nominal categorical features.

Identify Columns with Object (String) dtype

Columns with object (string) data type were identified and converted to numerical values as described above.

5. Splitting Dataset into X & y

The dataset was split into feature matrix (X) and target vector (y):

- (X): Contains all the features.

- (y): Contains the target variable.

6. Splitting Data into Train & Test

The dataset was further split into training and testing sets to evaluate model performance. Typically, a 70-30 or 80-20 split was used:

- Training set: Used to train the models.

- Testing set: Used to evaluate the models' performance.

7. Model Building

Multiple machine learning models were built and evaluated to identify the best performing model. The models include Decision Tree, Random Forest, Linear Regression with hyper parameter tuning, and K-Nearest Neighbors (KNN).

a. Decision Tree

- **Accuracy:** 81.77%

- **Challenges:** Prone to overfitting, especially with complex datasets.

- **Techniques Used:** Pruning to reduce overfitting.

b. Random Forest

- Accuracy: 87.83%

- Challenges: Computationally intensive, especially with a large number of trees.

- Techniques Used: Ensemble method to improve accuracy and reduce overfitting compared to a single decision tree.

c. Linear Regression with Hyper parameter Tuning

- **Accuracy**: 87.57%

- **Challenges**: Sensitive to outliers and multicollinearity.

- **Techniques Used**: Regularization techniques (Ridge or Lasso) to improve performance.

d. K-Nearest Neighbors (KNN)

- **Accuracy (Test):** 86.99%

- **Accuracy (Train):** 95.29%

- **Challenges:** Sensitive to the choice of (k), computationally expensive with large datasets.

Techniques Used: Hyper parameter tuning to find the optimal \(k \).

Challenges Faced

Data Imbalance

- **Challenge**: Imbalanced classes can lead to biased models.

- **Solution:** Used techniques such as SMOTE (Synthetic Minority Over-sampling Technique) to balance the classes.

Missing Values

- Challenge: Missing data can lead to biased or invalid models.

- **Solution:** Used imputation techniques to fill missing values or removed columns/rows with excessive missing data.

Feature Engineering

- Challenge: Identifying and creating meaningful features can significantly impact model performance.

- Solution: Created new features based on domain knowledge and existing features to improve model accuracy.

Model Performance and Recommendation

Based on the performance metrics, the Random Forest model achieved the highest accuracy of 87.83%, followed closely by the Linear Regression model with hyper parameter tuning at 87.57%. Despite the slight difference in accuracy, Random Forest is recommended for production use due to its robustness and ability to handle non-linear relationships and interactions between features.

Conclusion

This report summarizes the data pre-processing, exploratory analysis, and model building steps, along with the challenges faced and solutions implemented. The Random Forest model is recommended for production due to its superior performance and robustness.