Capstone Project

Diana Procel

Introduction

Dataset: E_commerce

Rows: 10,999

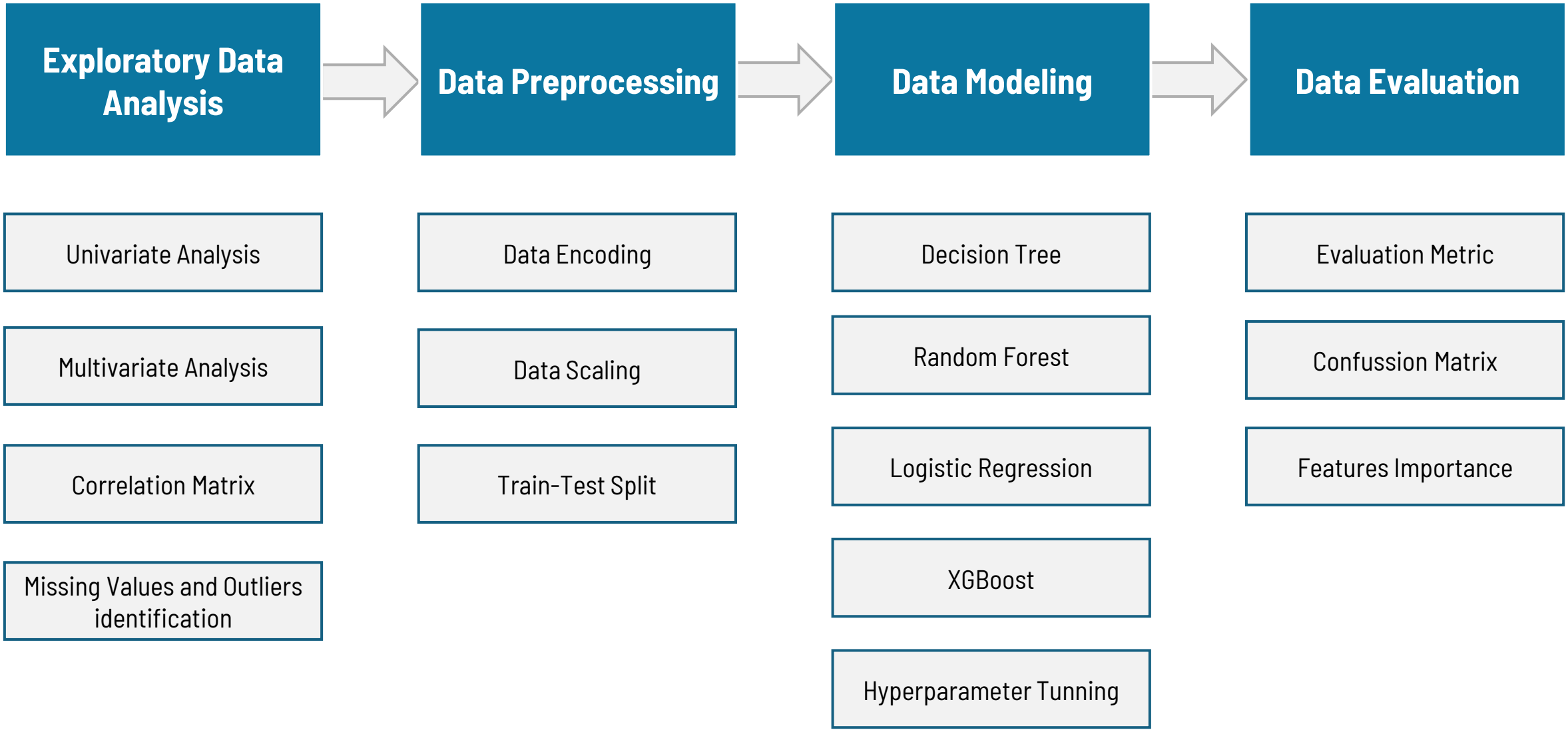
Columns: 12

Reached on time is the target variable, where 1 Indicates that the product has NOT reached on time and 0 indicates it has reached on time.

Variables:

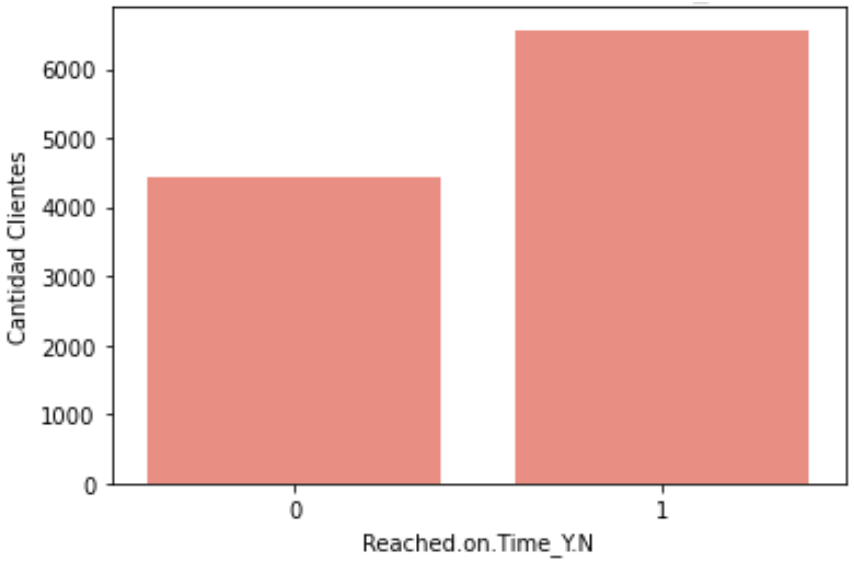
ID	Warehouse_block	Mode_of_Shipment
Customer_care_calls	Customer_rating	Cost_of_the_Product
Prior_purchases	Product_importance	Gender
Discount_offered	Weight_in_gms	Reached.on.Time_Y.N

Methodology

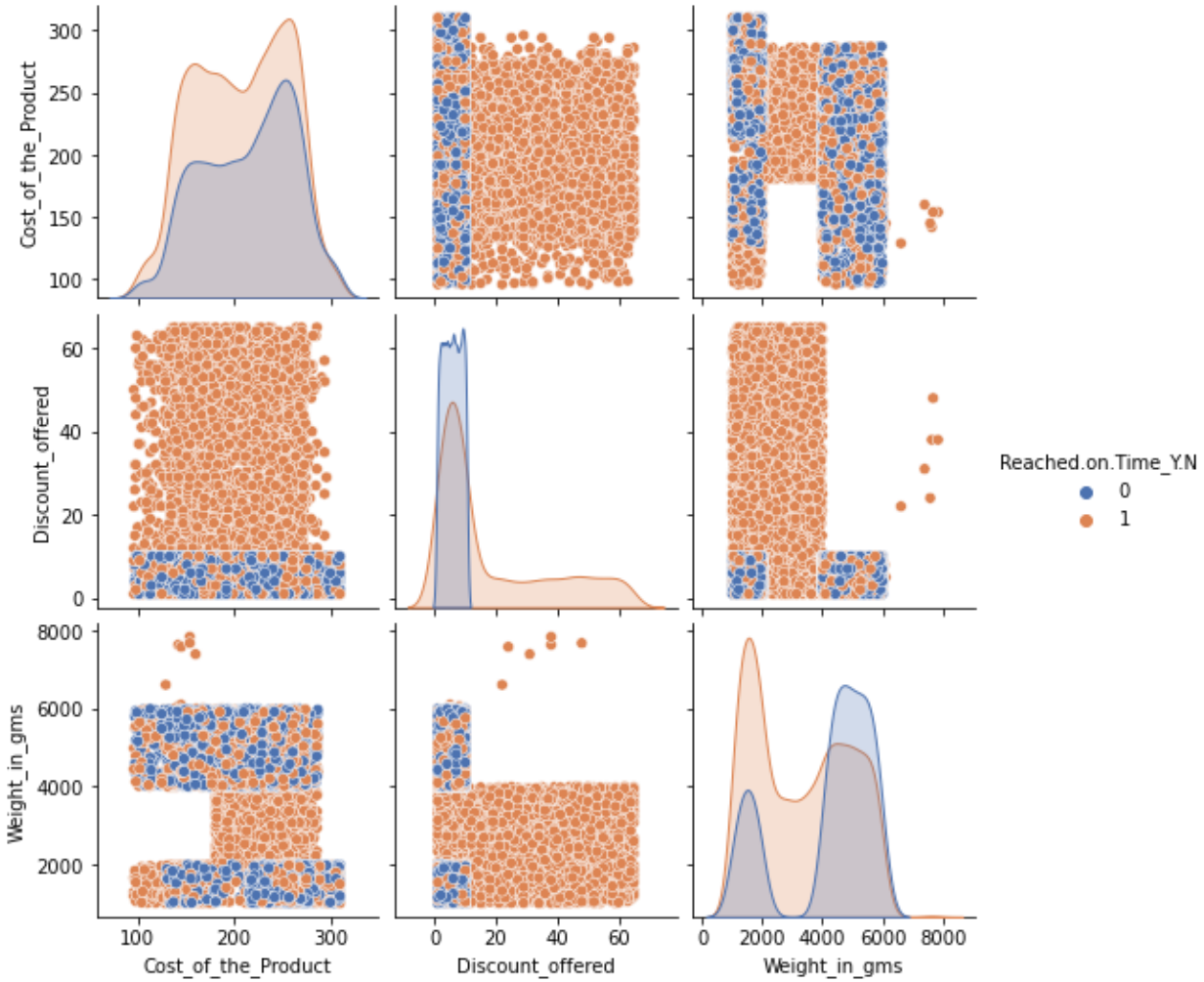


Exploratory Data Analysis

Distribution of Target

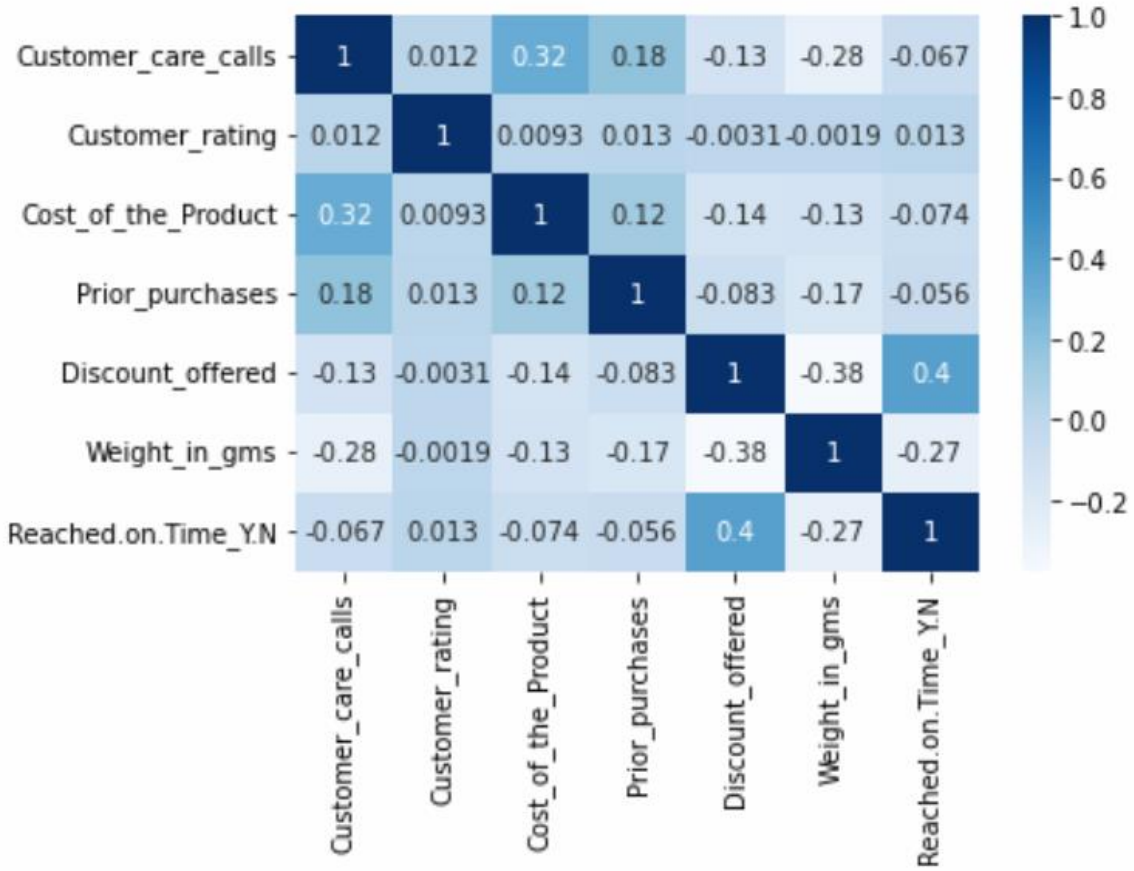


Pairplot

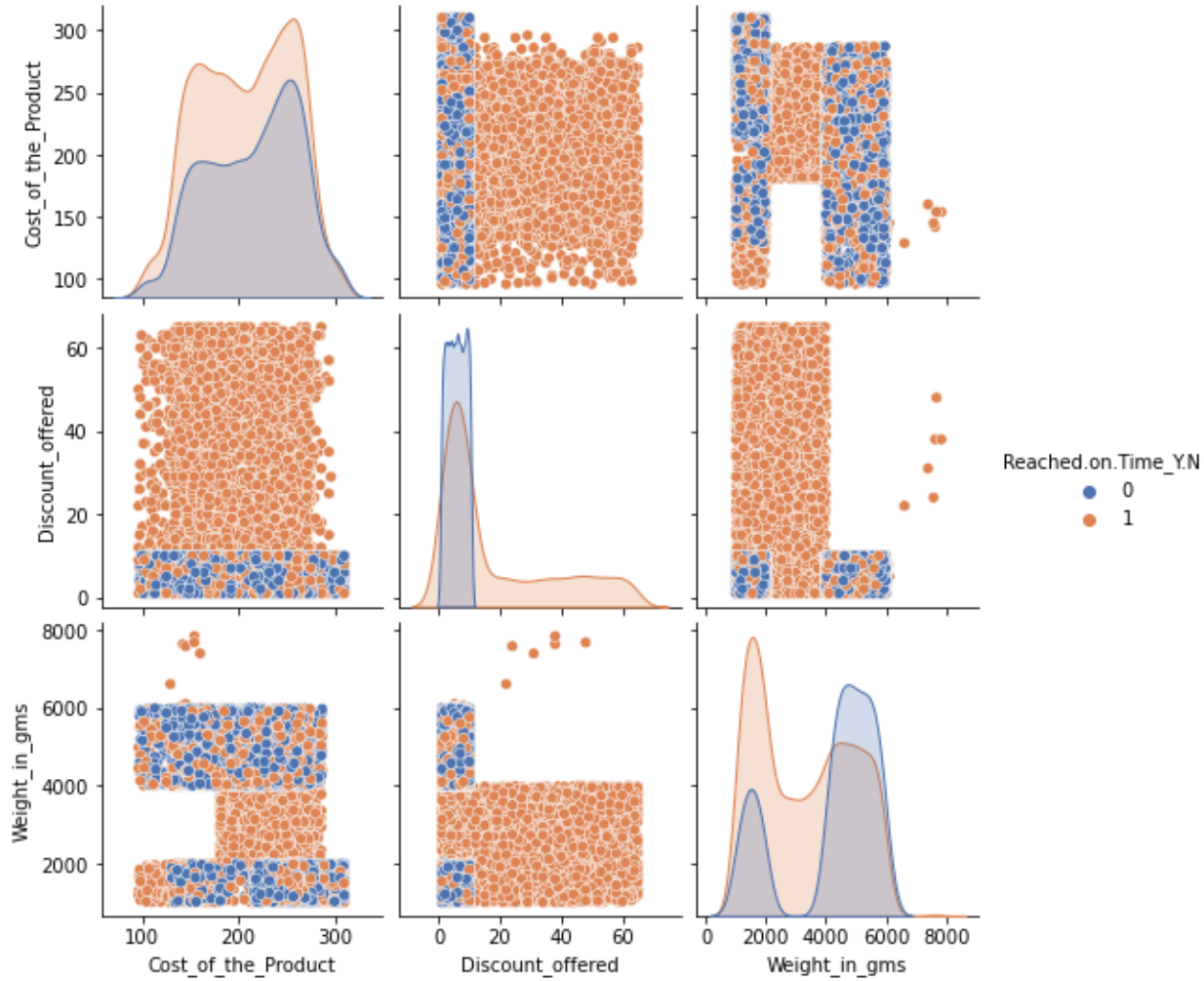


Exploratory Data Analysis

Correlation Matrix



Pairplot



Exploratory Data Analysis

Conclusions:

- The dependent variable is not uniformly distributed, so for evaluating a future classification model, accuracy would not be the correct metric.

- The variable discount_Offered has a skewed distribution to the right, so it has outlier. And to work with this variable we should consider using some outlier imputation technique or column transformation, such as logarithmic transformation.

- Within the text variables, the variable Product_importance can be considered as an ordinal variable, so we can encode it keeping its order relation with Ordinal Encoding.

- No missing values

Important Insights:

- All deliveries greater than 17 in discount_offered arrived on time. (This explains the correlation of 0.4 that these two variables have.
- All deliveries that weighed between 2000 and 4000 gms were on time.
- Deliveries with priority 1 and 2, are the ones with better % of deliveries that arrived on time than the other priorities.

Data Preprocessing

Data Transformation

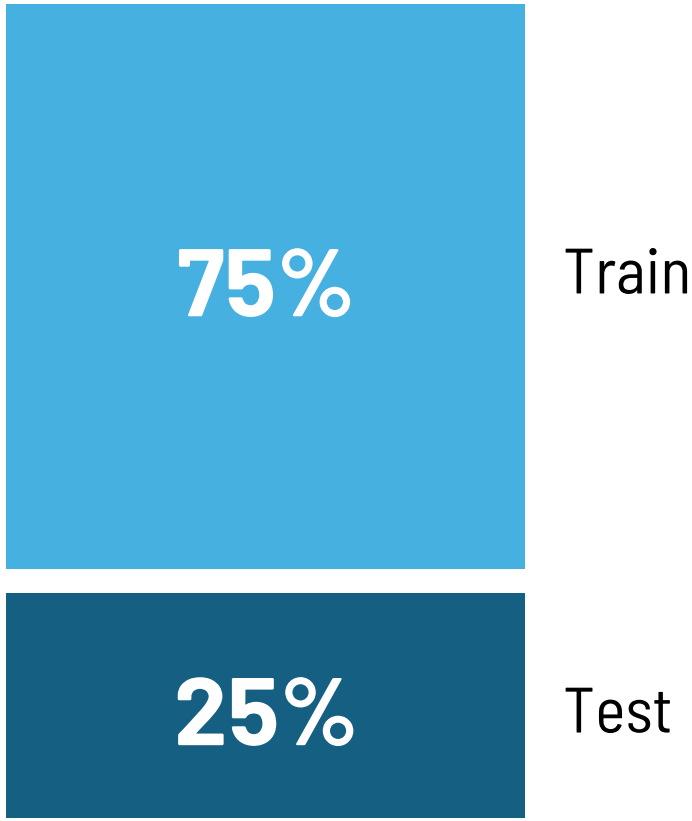
For String variables:

- Ordinal Encoding
- One hot Encoding

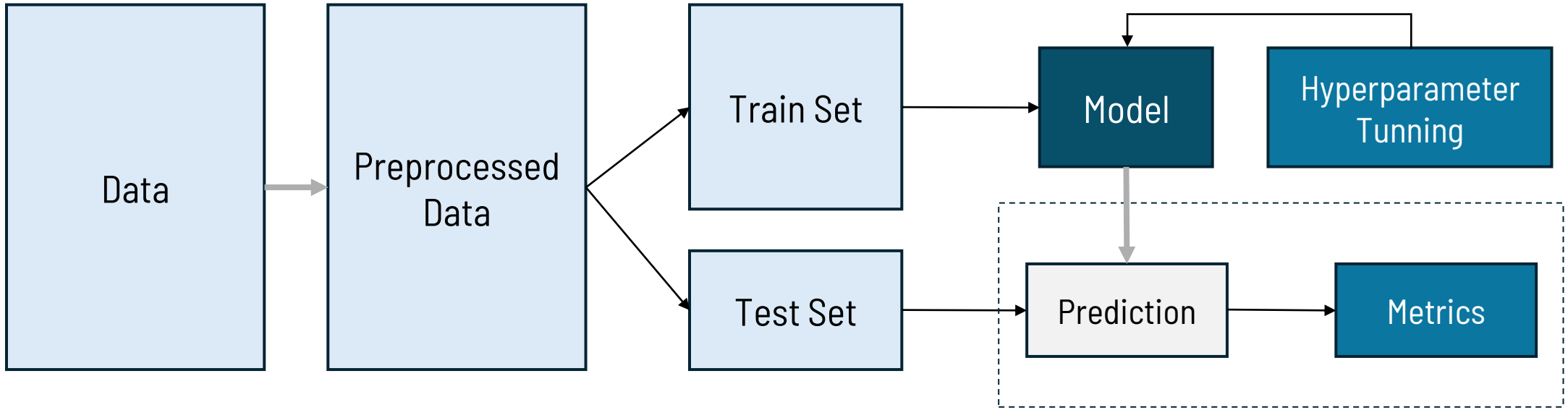
For Numerical variables:

- Min-Max Scaler

Train-Test Split



Data Modeling



Models we tested:

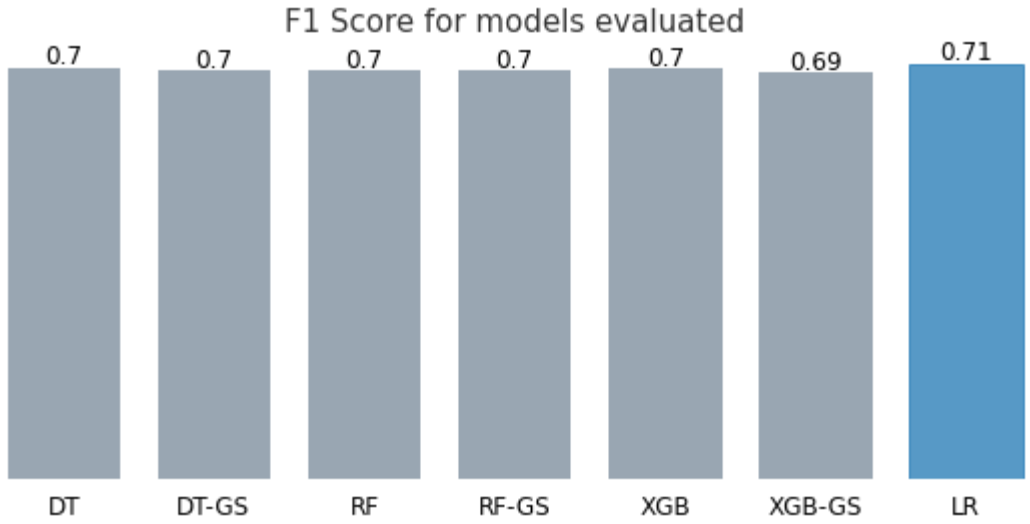
- Decistion Tree
- Random Forest
- XGBoost
- Logistic Regression

Hyper Parameter Tunning Method:

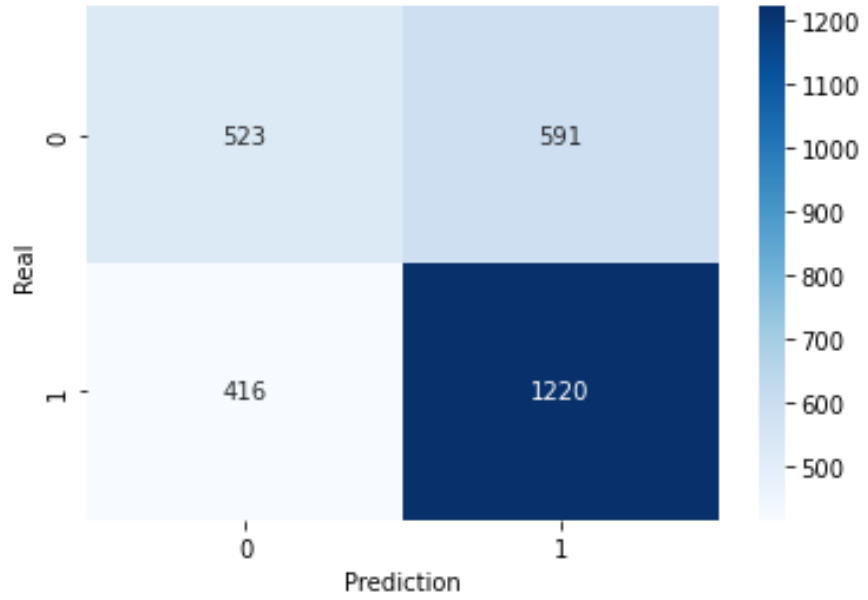
- Grid Search CV

Data Evaluation

Evaluation Metric Selected



Confussion Matrix



According to the unbalance distribution of the target, I decided to use **F1 Score** as the Evaluation Metric

Best Model: **Logistic Regression**

Features Importance