**Predicting the Likelihood of a Completed Purchase Using Classification model**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Use Case** | **Skill Covered** | **Expected Outcome** | **Complexity** | **Industry** | **Total Marks** |
| Build Classification models | * EDA (missing value and outlier treatment, univariate and multivariate analysis) * Feature Engineering (one-hot, label encoding, Scaling) * Building classification models and evaluating them * Choosing the best model to suit the business problem | Understand the Significance of Classification models (Logistic Regression, Tree models) for solving business problems | Complex | Retail | 100 |

In retail, predicting the likelihood that a customer will complete a purchase is crucial for optimizing sales strategies, managing inventory, and personalizing customer experiences. Given a dataset of product sales, customer demographics, and promotional campaigns, the task is to **predict whether a transaction will result in a completed purchase (binary outcome: Yes/No)**. The goal is to develop **classification models** that estimate the probability of a customer completing their order based on various features, such as product pricing, customer demographics, and promotional offers.

**Problem Statement**: Develop **various classification models** that predict the likelihood of a purchase being completed, given the customer and product features.

The task is to predict the likelihood that a transaction will result in a **completed purchase** based on features such as:

* **Customer ID**: A unique identifier for each customer.
* **Age**: The customer's age, which could influence buying patterns.
* **Gender**: The gender of the customer (Male/Female).
* **Loyalty Member**: Whether the customer is a member of the loyalty program (Yes/No), which could indicate higher engagement or likelihood to complete a purchase.
* **Product Type**: The category or type of product purchased (e.g., Electronics, Apparel), which might correlate with purchase completion rates.
* **SKU**: Unique identifier for the product (Stock Keeping Unit).
* **Rating**: The average product rating (1-5), which could influence the customer's decision to complete the purchase.
* **Order Status**: Indicates whether the order was completed, pending, or canceled. For this model, the target will be whether the **order was completed** (Yes/No).
* **Payment Method**: The method of payment used (e.g., Credit Card, PayPal), which may influence the probability of a transaction being completed.
* **Unit Price**: The price per unit of the product, as higher-priced items may have lower conversion rates.
* **Quantity**: The number of units purchased in the order.
* **Purchase Date**: The date when the purchase was made, which might capture time-related patterns in purchase behavior.
* **Shipping Type**: The shipping method chosen by the customer (e.g., Standard, Expedited), which could be linked to the likelihood of completing the order.
* **Add-ons Purchased**: Whether the customer bought additional items (e.g., accessories, warranties) along with the main product (Yes/No).
* **Add-on Total**: The total cost of any add-ons purchased, which contributes to the final price and may influence purchase completion.

The goal is to predict the **probability** that a customer will complete a transaction, which will be modelled as a binary outcome:

**Approach Overview:**

1. **Data Preprocessing**:
   * Clean the data, handling missing values, encoding categorical variables, and ensuring that only completed orders are used in the target variable.
   * Create a binary target variable: **1** for completed orders, **0** for non-completed orders.
2. **Feature Engineering**:
   * Engineer features from the raw data, such as creating binary variables for **Add-ons Purchased** or **Loyalty Member** status.
   * Normalize or scale continuous features like **Unit Price** and **Quantity** if necessary.
3. **Model Training**:
   * Split the dataset into training and testing sets.
   * Train the classification models on the training data to predict the binary outcome of purchase completion.
4. **Model Evaluation**:
   * Use metrics such as **accuracy**, **precision**, **recall**, and **AUC-ROC** to evaluate model performance.
5. **Prediction**:
   * Use the trained classification models to predict the likelihood of purchase completion for future transactions.
6. **Conclusion**:

* Choose the best model based on efficiency in solving the business problem

**Task 1: Load the dataset and perform preliminary EDA (Exploratory Data Analysis) with key observations and insights- (weightage - 20 marks)** (AE)

|  |  |  |
| --- | --- | --- |
| **T1.1** | Load the Sales dataset. | (weightage - 2 marks) |
| **T1.2** | Check the shape and data types. | (weightage - 1 mark) |
| **T1.3** | Remove variables that are not required, namely, Customer ID, SKU and Purchase Date | (weightage - 2 marks) |
| **T1.4** | Draw box plots to check for outliers for numeric variables, namely: Age, Rating, Total Price, Quantity and Add-on Total  Run ‘describe’ function to get the descriptive statistics of the aforementioned variables | (weightage - 3 marks) |
| **T1.5** | Do outlier treatment. Take lower and upper bound based on Quartiles and 1.5 times IQR and then cap the outliers with the lower bound and upper bound values | (weightage - 6 marks) |
| **T1.6** | Run the ‘describe’ function on the treated data and note down the variables for which the ‘max’ value has now changed post the outlier treatment | (weightage - 2 marks) |
| **T1.7** | Find the missing values. Note down the number of missing values for variable ‘Gender’ | (weightage - 2 marks) |
| **T1.8** | Drop the rows in which ‘Gender’ has missing values, recheck for missing values and note down the variable(s) that still has missing values | (weightage - 2 marks) |

**Task 2: Carry out extensive data preparation and feature engineering (weightage - 15 marks)** (ME)

|  |  |  |
| --- | --- | --- |
| **T2.1** | Do further univariate and multivariate analysis and convert the target variable into 0 and 1. | (weightage -5 marks) |
| **T2.2** | Split the dataset into train and test. | (weightage -3 marks) |
| **T2.3** | Create dummy variables and scale the numerical features | (weightage-7 marks) |

**Task 3: Build and evaluate the models (weightage - 40 marks)**  (ME)

|  |  |  |
| --- | --- | --- |
| **T3.1** | Build classification models (Logistic Regression, Decision Tree, Random Forest and at least two Boosting models is the minimum) | (weightage-20 marks) |
| **T3.2** | Check for the model evaluation parameters and do fine-tuning when necessary to make models free of errors | (weightage-20 marks) |

**Task 4: Summarize the findings of the analysis and draw conclusions with PPT / PDF. (weightage - 25 marks) (ME)**

**Final Submission guidelines:**

1. Download the Jupyter notebook in the format of html.
2. Upload it in the lumen (UNext LMS)
3. Summarized PPT/ PDF prepared in Task 4 to be uploaded in the lumen (UNext LMS) and **it must contain why and how the best model was selected.**

 ========================================================