****

**Assessment Report**

on

**“Predicting Product Return”**

submitted as partial fulfillment for the award of

**BACHELOR OF TECHNOLOGY**

**DEGREE**

SESSION 2024-25

in

**CSE(AI)**

By

Name : Akash Kumar Gupta

Roll Number : 202401100300025

Section: A

**Under the supervision of**

“Mr. BIKKI KUMAR”

**KIET Group of Institutions, Ghaziabad**

**May, 2025**

**1. Introduction**

Predicting whether a product will be returned is a common classification problem in e-commerce analytics. Accurate predictions can help companies reduce return rates, improve customer satisfaction, and optimize logistics. This report outlines the approach used to build a machine learning model that predicts product return status based on purchase context and customer reviews.

**2. Problem Definition**

The main objective is to develop a classification model that can:

1. **Classify** whether an item will be returned (Yes or No) based on features like purchase behavior, item details, and reviews.
2. **Evaluate** model performance using metrics such as **accuracy**, **precision**, **recall**, and a **confusion matrix heatmap**.

**3. Data Description**

The dataset contains the following types of features:

* **Customer Info**: Age, location, etc.
* **Purchase Context**: Quantity, delivery time, price, etc.
* **Product Features**: Product ID, category, etc.
* **Reviews**: Sentiment score or textual rating
* **Target Variable**: returned (1 if returned, 0 if not)

**4. Approach**

To solve the classification task, the following steps were performed:

**4.1 Data Preprocessing**

* Cleaned column names
* Encoded categorical variables
* Scaled numeric features using StandardScaler
* Split data into training and testing sets (80/20)

**4.2 Model Training**

* Chose **Logistic Regression** for baseline classification.
* Trained on the processed training data.

**4.3 Evaluation Metrics**

The model is evaluated on:

* **Accuracy**: Overall correct predictions
* **Precision**: Fraction of predicted returns that were correct
* **Recall**: Fraction of actual returns that were correctly predicted
* **Confusion Matrix**: Visualizes TP, FP, TN, FN

**5. Data Preprocessing**

The dataset is cleaned and prepared using the following steps:

* Missing Values Handling:  
  Numerical columns with missing values are filled with their respective column means.
* Categorical Encoding:  
  All categorical variables are converted into numerical format using one-hot encoding.
* Feature Scaling:  
  All numeric features are normalized using StandardScaler to ensure all features contribute equally.
* Train-Test Split:  
  The dataset is split into 80% training and 20% testing subsets to train and evaluate the model properly.

**6. Model Implementation**

A Logistic Regression model is selected due to its:

* Efficiency for binary classification tasks
* Simplicity and interpretability

The model is trained on the processed training dataset and then used to predict the return status (Returned or Not Returned) on the test set.

**7. Evaluation Metrics**

To evaluate model performance, the following metrics are used:

* Accuracy: Measures the percentage of correct predictions over total predictions.
* Precision: Indicates the proportion of correctly predicted product returns out of all predicted returns.
* Recall: Reflects the proportion of actual product returns that were correctly identified.
* F1 Score: Harmonic mean of precision and recall, balancing both metrics.
* Confusion Matrix:  
  A visual heatmap using Seaborn helps analyze:
  + True Positives (correctly predicted returns)
  + False Positives (products predicted to be returned but weren’t)
  + True Negatives
  + False Negatives

**8. Results and Analysis**

* The model achieved reasonable performance in predicting whether a product will be returned.
* The confusion matrix heatmap provided insights into the nature of incorrect predictions.
* Precision and recall scores revealed the trade-off between correctly identifying returns vs. minimizing false positives.
* The F1 score offered a balanced view of model effectiveness in both precision and recall.

**9. Conclusion**

The logistic regression model successfully classified product return status with satisfactory evaluation metrics. This project showcases how machine learning can aid e-commerce businesses in predicting returns, enhancing operational efficiency, and reducing associated costs.

Potential Improvements:

* Incorporating NLP-based sentiment analysis of customer reviews.
* Handling class imbalance with oversampling (e.g., SMOTE) or penalized models.
* Trying more advanced models like Random Forests or Gradient Boosting.

**10. References**

* scikit-learn Documentation
* pandas Documentation
* Seaborn Visualization Library
* Research papers on e-commerce return prediction and classification models

