**Data Analysis Task:**

**Predictive Model for Lowering Product Returns at the E-Store**

In this report, we suggest a data analysis and preprocessing process that sets the foundation for building predictive models aimed at addressing the product returns problem within the e-store of a Nordic fashion brand.

Brief summary of the key steps you've taken:

1. Step 1: Data Exploration and Understanding:

* In this initial phase, we loaded the dataset into Jupyter Notebook, examined its structure, and assessed its size, basic statistics, missing values, unique values in categorical columns, and visualized data distributions using histograms and bar plots.

1. Step 2: Data Preprocessing

* In the data preprocessing step, we addressed missing values, converted date columns to datetime objects, split the 'order\_products' column into individual features, and applied one-hot encoding to the 'gender' variable.

1. Step 3: Data Analysis and Visualization

* Explored the distribution of returns by gender, weekday, and order time.

1. Step 4: Feature Selection and Engineering

* Calculated feature correlations with the target variable ('returned') and selected relevant features based on a specified threshold (0.02).
* Conducted feature engineering to calculate the product return rate.

1. Step 5: Splitting the Dataset

* Split the dataset into training (70%) and testing (30%) sets for model evaluation.

1. Step 6: Building Predictive Models

* Trained multiple models for binary classification tasks, such as Logistic Regression, Random Forest, and K-Nearest Neighbors (KNN).
* Evaluate, and compared different machine learning models to find the best solution.

The goal of this data analysis task was to address the issue of product returns at the Nordic fashion brand's e-store by creating predictive models and analyzing their potential use. Various models were constructed and evaluated. The following table summarizes the performance of each model:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Model Training** | **Time (s)** | **Accuracy** | **Precision** | **Recall** | **F1 Score** | **ROC AUC** |
| Random Forest | 1741.10 | 0.62 | 0.52 | 0.10 | 0.16 | 0.52 |
| Logistic Regression | 47.36 | 0.62 | 0.44 | 0.00 | 0.01 | 0.50 |
| Gradient Boosting | 1406.36 | 0.62 | 0.55 | 0.04 | 0.07 | 0.51 |
| K-Nearest Neighbors | 826.74 | 0.56 | 0.41 | 0.31 | 0.35 | 0.52 |
| Logistic Regression + Resampling | 26.52 | 0.51 | 0.40 | 0.54 | 0.46 | 0.51 |
| Logistic Regression + Normalization | 34.59 | 0.61 | 0.40 | 0.02 | 0.03 | 0.50 |
| KNN + Normalization | 128.50 | 0.56 | 0.41 | 0.31 | 0.35 | 0.52 |

**Interpretation:**

The table provides a quick overview of how each model performs across multiple metrics. While models like Random Forest and Logistic Regression show similar accuracy, they exhibit variations in precision, recall, and F1 Score. Gradient Boosting and K-Nearest Neighbors demonstrate differences in precision, recall, and ROC AUC.

**Recommendation:**

While some models demonstrated acceptable accuracy, it's crucial to consider precision, recall, and F1 Score in the context of product return prediction. The low recall in most models suggests that they struggle to correctly identify cases of actual product returns. **Given this, the predictive models, in their current states, may not be suitable as the sole basis for an automatic prediction plugin.** **However, they could still be valuable for informing human decision-making.** For instance, customer service representatives could benefit from these predictions to handle returns more efficiently.

We can enhance the model by applying refinements to feature engineering or reviewing the feature selection. If the model evaluation gives better results, we can suggest integrating the predictive model into the checkout process. In that case (prediction is high), we might provide suggestions to reduce returns such as alternative product recommendations or sizing guidance.

**Analysis:**

The analysis of the product returns problem reveals several important insights. The dataset is a valuable resource for predictive modeling of product returns, but it does present certain challenges. Notably, a significant number of missing values are observed in the "date of birth" column, which can impact the accuracy and completeness of our predictive models. The age of the customers can be a potential feature for the predictive model; however, this feature was not considered in this work, but further strategies for handling these missing values need to be explored.