Zalando Product Return Prediction

1. Abstract

This project focuses on predicting whether a product purchased from Zalando will be returned, using supervised machine learning. A Random Forest Classifier model was trained on features like product price, delivery duration, coupon usage, and size issues to predict the likelihood of a return. The project aims to reduce return costs and optimize logistics.

2. Introduction

Product returns in e-commerce, especially in fashion retail, create substantial costs in terms of logistics, inventory management, and customer dissatisfaction. Zalando, a leading European fashion retailer, experiences high return rates, and understanding the drivers behind returns is critical. This project builds a predictive model to estimate the probability of a return using customer and order-related features.

3. Business Objective

The objective is to create a classification model that can predict product returns accurately. Such predictions can help Zalando:

- Optimize warehouse operations
- Personalize recommendations
- Adjust marketing strategies
- Improve size guidance tools

4. Dataset Description

The dataset includes the following columns:

- `product_price` : Price of the product
- `delivery_days` : Days taken for delivery
- `used_coupon` : Whether a coupon was used (1 = Yes, 0 = No)
- `size_issue` : Whether the return reason was a size issue (1 = Yes, 0 = No)
- `returned` : Target variable (1 = Returned, 0 = Not Returned)

5. Data Exploration & Visualization

The initial analysis showed the distribution of returned vs. non-returned items:

```
import matplotlib.pyplot as plt
df['returned'].value_counts().plot(kind='bar', title='Return Status Distribution')
plt.xlabel('Returned')
plt.ylabel('Count')
plt.show()
```

Key findings:

- The dataset is moderately imbalanced.
- Return rates are higher for products with size issues.

---

6. Data Preprocessing

No missing values were present in the dataset. All features were numerical or binary and did not require encoding. Normalization was not applied as Random Forests are scale-invariant.

---

7. Feature Selection

The following features were selected based on domain knowledge:

- `product_price`
- `delivery_days`
- `used_coupon`
- `size_issue`

These were used as input variables (X), and `returned` as the target (y).

---

8. Model Building

A Random Forest Classifier was trained with default parameters:

```
from sklearn.ensemble import RandomForestClassifier
model = RandomForestClassifier()
model.fit(X_train, y_train)
```

9. Model Evaluation

Predictions were made on the test set, and evaluation was performed using classification report and confusion matrix:

```
from sklearn.metrics import classification_report, confusion_matrix
print(classification_report(y_test, y_pred))
```

The confusion matrix was visualized using:

```
import seaborn as sns
sns.heatmap(confusion_matrix(y_test, y_pred), annot=True, fmt='d')
```

---

10. Results & Insights

- The model showed good accuracy in identifying returns.
- `size_issue` and `delivery_days` were significant predictors.
- Coupon usage slightly increased the probability of returns.

---

11. Error Analysis

False positives (predicting a return when not returned) were more frequent than false negatives. This implies the model errs on the side of caution, which could be beneficial in logistic planning.

---

12. Next Steps

- Try other models like Logistic Regression or XGBoost.
- Perform hyperparameter tuning.
- Include additional features such as customer demographics, product category, and order history.

---

13. Deployment Options

The model can be deployed as a REST API using Flask or integrated into a web dashboard using Streamlit. It could be embedded in Zalando's internal order processing system.

---

14. Public Usage

A public version of the model could be hosted on a Streamlit or Flask app where users input features and see predictions in real time.

---

15. Tools & Technologies Used

- **Language**: Python 3
- **Libraries**: Pandas, Seaborn, Matplotlib, Scikit-learn
- **IDE**: Google Colab
- **Model**: Random Forest Classifier

---

16. Conclusion & Future Work

This project demonstrates that predicting product returns using machine learning is both feasible and valuable for e-commerce businesses. With more data and refined features, the model’s accuracy can be improved further. Future work could include time-based predictions, product clustering, and integration into CRM systems.