Predicting e-commerce Returns Behavior

Presented by Marta Scropetta

Deep Dive on e-commerce Returns

Context

e-Commerce industry loses around **$48 billion** per year due to fraudulent returns*. Nonetheless, there is a lack on fraudulent cases data availability.

10.2% Lost Revenue

Analyzing returns customer behavior pattern represent an important first step to prevent fraud issues in this sector.

*S. E. Kohan, Forbes, 2023

Project Timeline

Interest & Idea

The work experience I have in the e-commerce sector made me realize the importance of the topic

Data on Fraud vs Returns

The lack of fraud related data led me stick to analyze the refund information only

Data Collection Challenge

There is a general challenge in finding data related to biggest Western e-commerce

Data Sources Check

After checking different sources, I downloaded the data of a Pakistani e-commerce from Kaggle

Cleaning & Analysis

Once I got the raw data ready, I started with the cleaning, EDA, and, finally, the Models Application

Exploratory Data Analysis

1 ***General Inspection***

- Dataset: 584,524 transaction records (2016-2018)
- Information on status, price, quantity, discount, payment, category, etc. per order
- <0,5% missing values (only some variables are impacted)
- Light data inconsistencies

2 ***Cleaning***

- Wide Variables sub-section
- Re-grouping for better understanding → e.g. culture-related payment methods

3 ***Very First Relevant Insights***

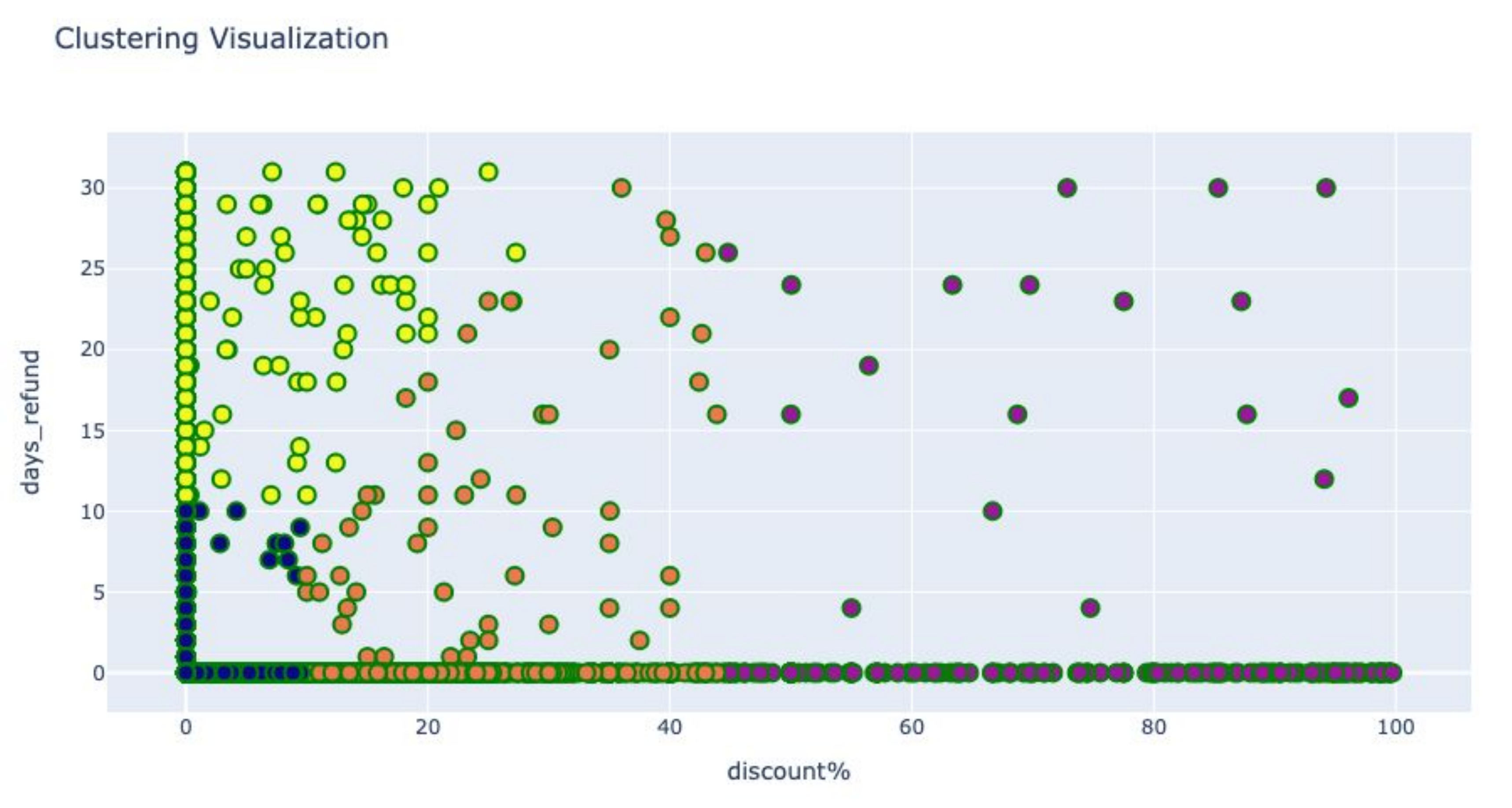
- >20% Orders are refunded
- Large % customers asked refunds → no clear return pattern
- Correlation:
  - Shipping fee & Quantity: positive
  - Refund Status & Refund Days: null
  - Refund Status & Discount: null

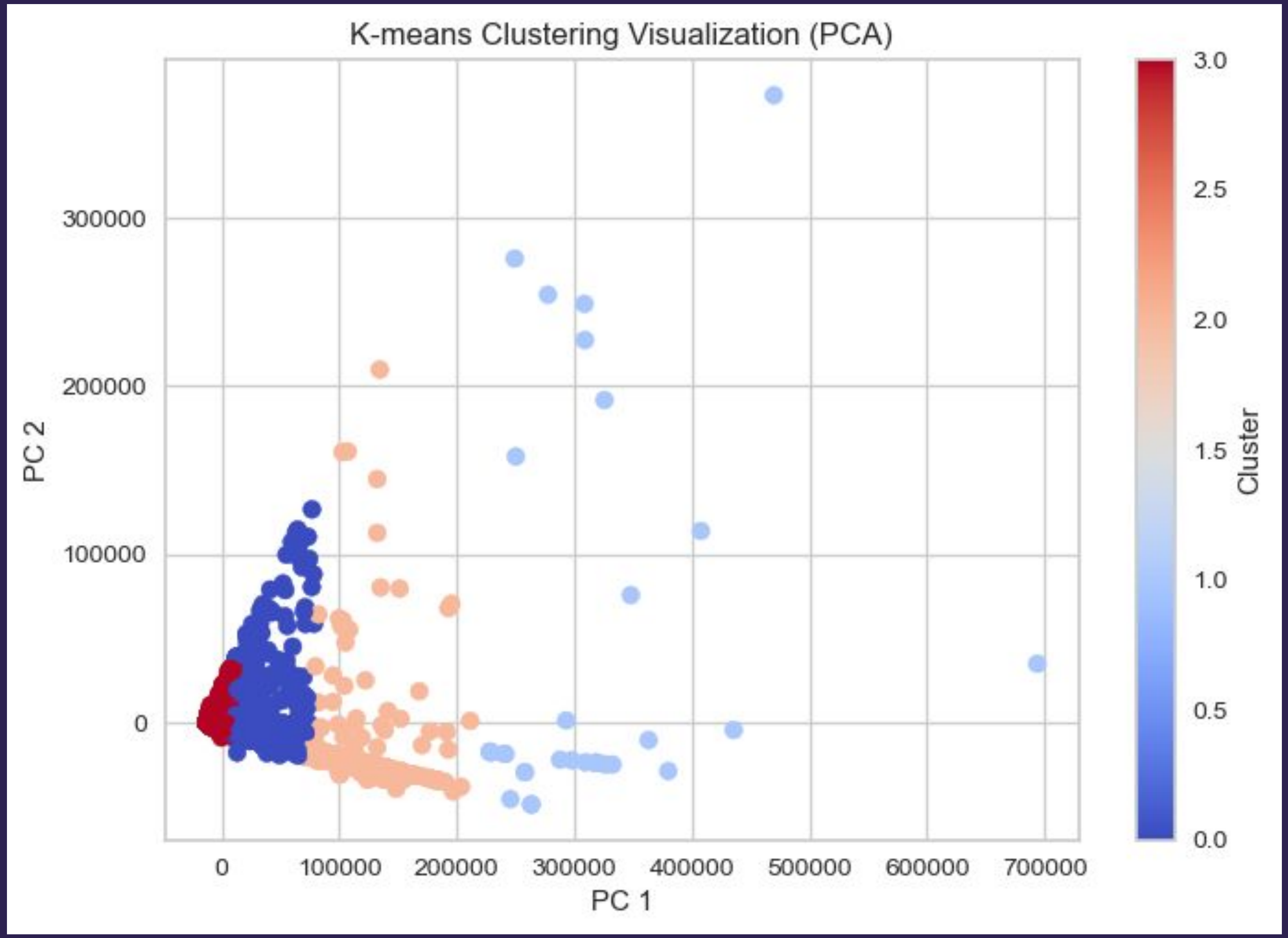
Exploratory Data Analysis

4

Data Clustering

- The magnitude of the dataset prevented the clustering analysis → Smaller Samples creation
- DBSCAN not effective anyways





Deep Dive on e-commerce Returns

Models Application

KNN

Advantage: intuitive for non-linear patterns (e.g. changing purchase trends)

→ 91,92% Accuracy

Random Forest

Advantage: resilient to outliers and good handling of high-dimensional data

→ 99,98% Accuracy

Xgboost

Advantage: effective in capturing interactions between features (e.g. orders details)

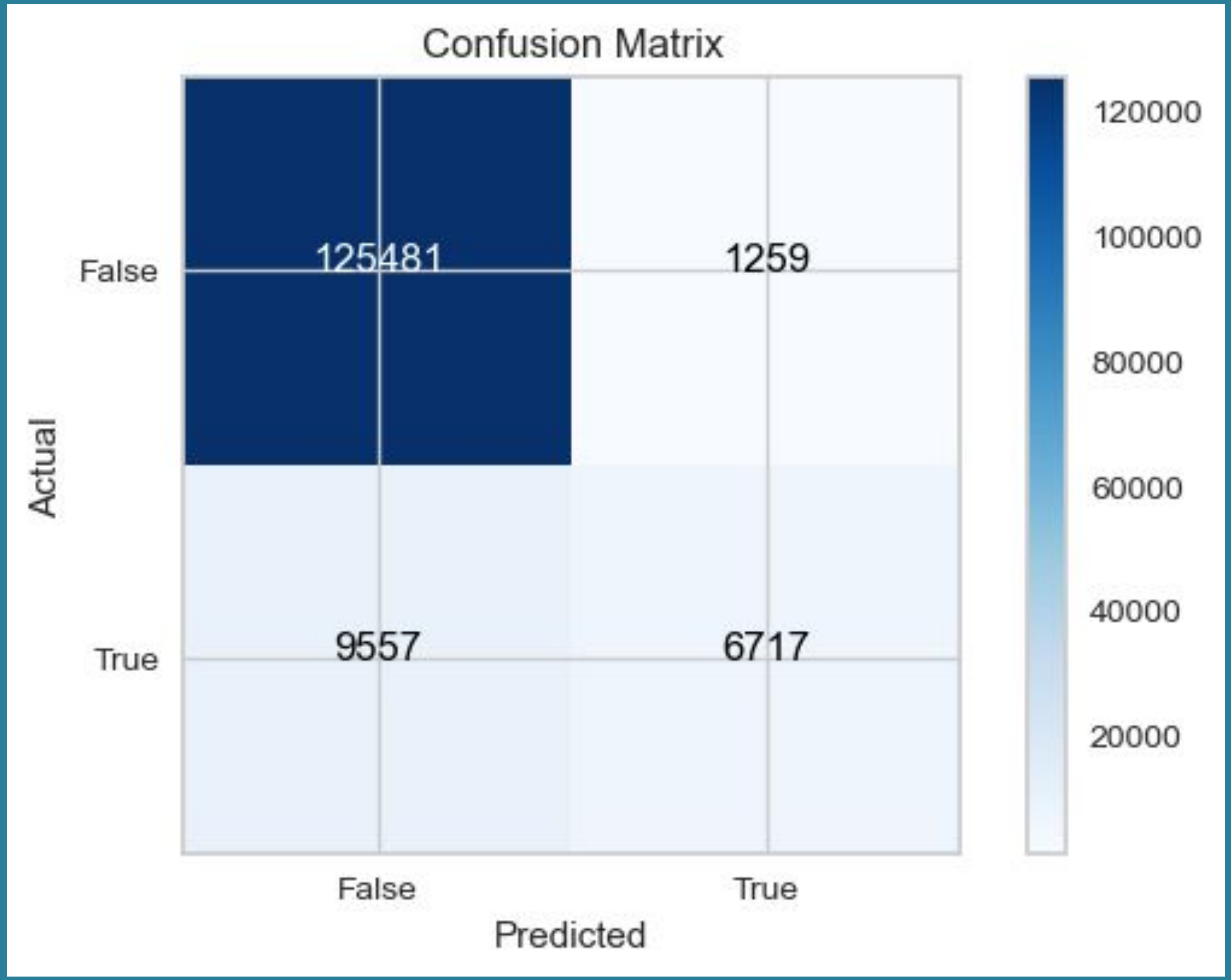
→ 99,99% Accuracy

Logistic Regression

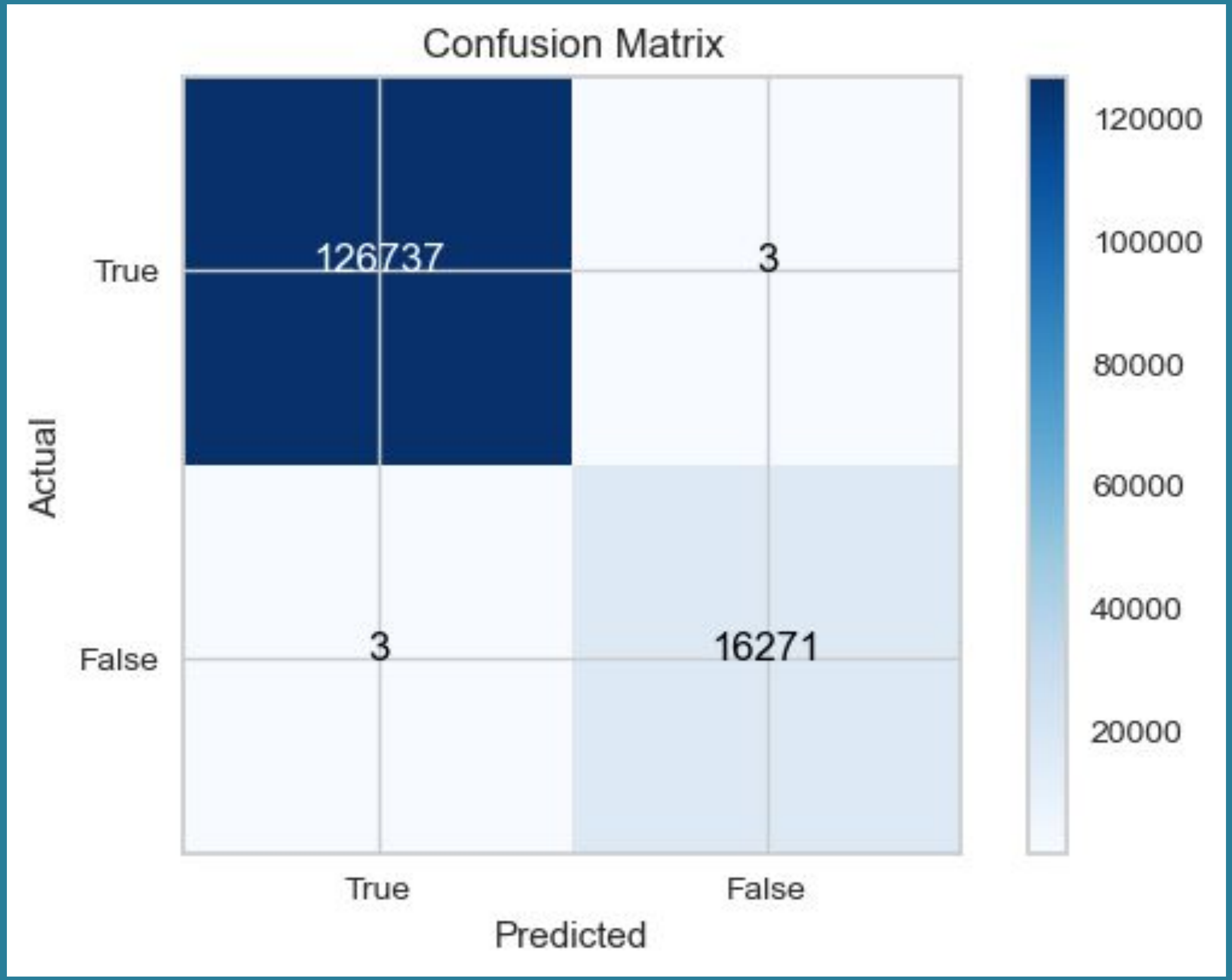
Advantage: robust for large-scale datasets (number of orders transactions)

→ 97,32% Accuracy

Models & Confusion Matrices



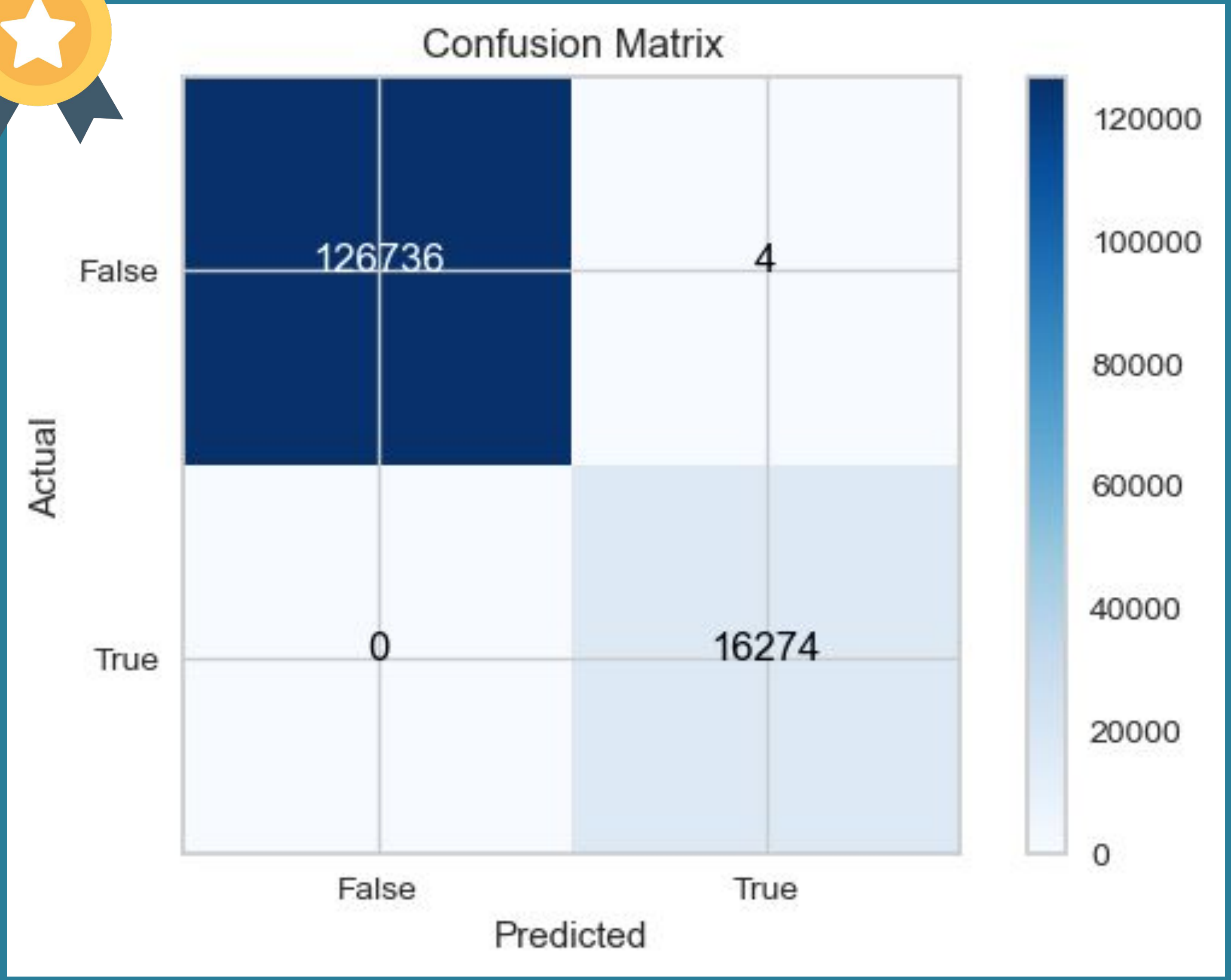
KNN



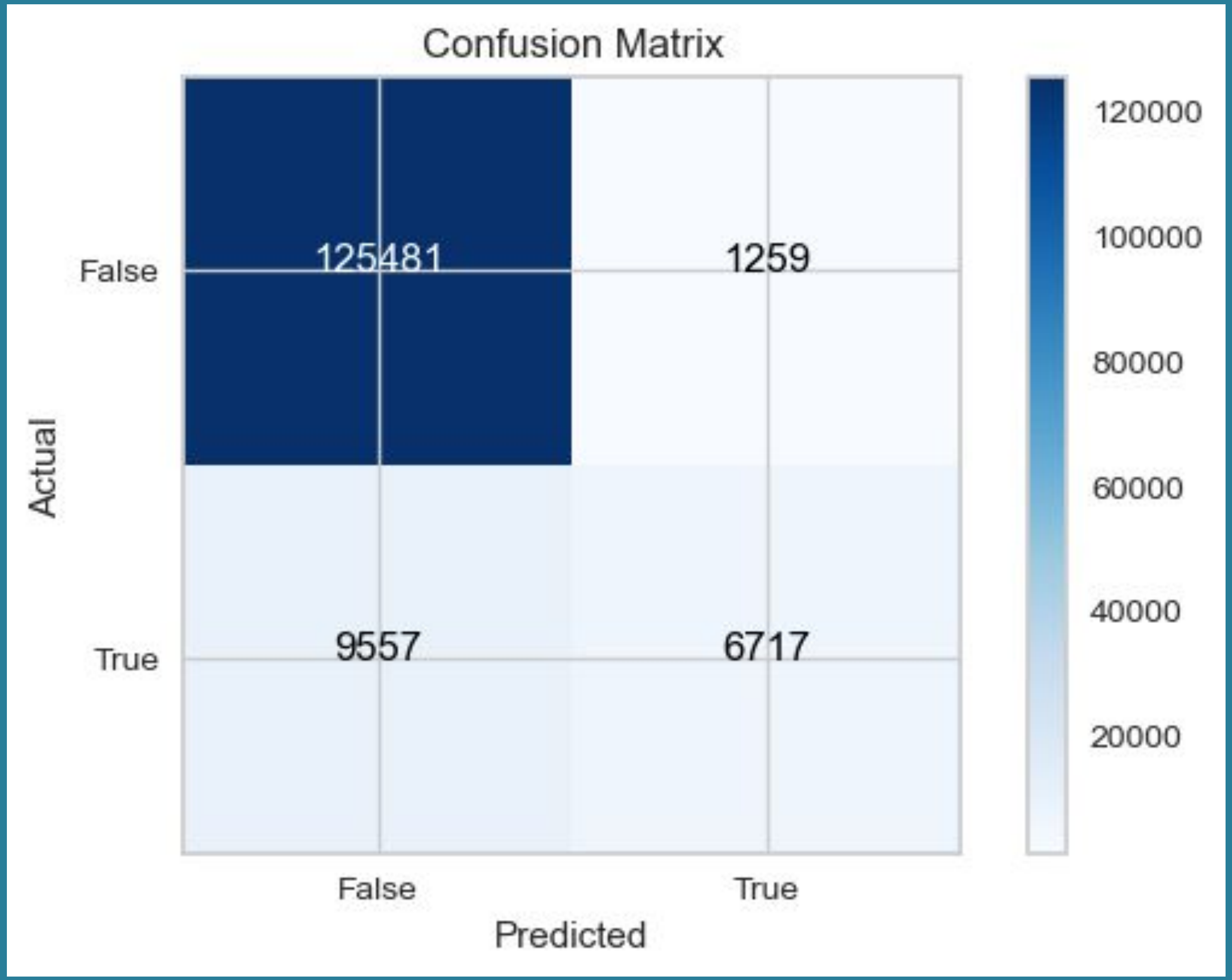
Random Forest

Deep Dive on e-commerce Returns

Models & Confusion Matrices



Xgboost



Logistic Regression

Conclusion & Implications

The Learning Model chosen to Predict the Order Status is the **Xgboost** as:

- Given the context, it is best to spot all returned items rather than risking to not correctly predict some of them
- It shows a slightly higher Accuracy

Overall, an improved detection on the Status of an Order Transaction implies:

1. **Fraud Detection and Prevention**: suspicious patterns or anomalies in refund requests an be identified
2. **Enhanced Decision-Making**: product offerings, pricing strategies, and marketing campaigns
3. **Reliable Planning**: accurate forecast on future returns volumes

Next Steps

Given the increasing losses caused by the larger amount of fraudulent returns, the suggested **next steps** are the following:

1. Deep Dive on the data Inconsistencies, as found for in the Returns Dataset
2. Exploitation of the Learning Model provided in this analysis to test different datasets (e.g. change in country, amount of orders)
3. Improving the gathering of data on fraudulent e-commerce return cases
4. Further analysis on customer profiles or product features that show larger fraudulent returns

Thank You!