HACKATHON ORSY®-SHELF
Accelerate sales for Würth Germany

Jan Ecke, Benjamin Harsch, Simon Rehm

January 17, 2024

Applied Data Science Lab (5000-671)
University of Hohenheim

Problem Statement

Problem Statement

What are typical characteristics of an ORSY-Shelf customer, and how do we identify new potential customers based on these characteristics?

To answer these questions, we need to train a machine-learning model, which:

1. Identifies relevant features
2. Predicts potential customers based on certain features
3. Delivers robust results

Data

Descriptive Data

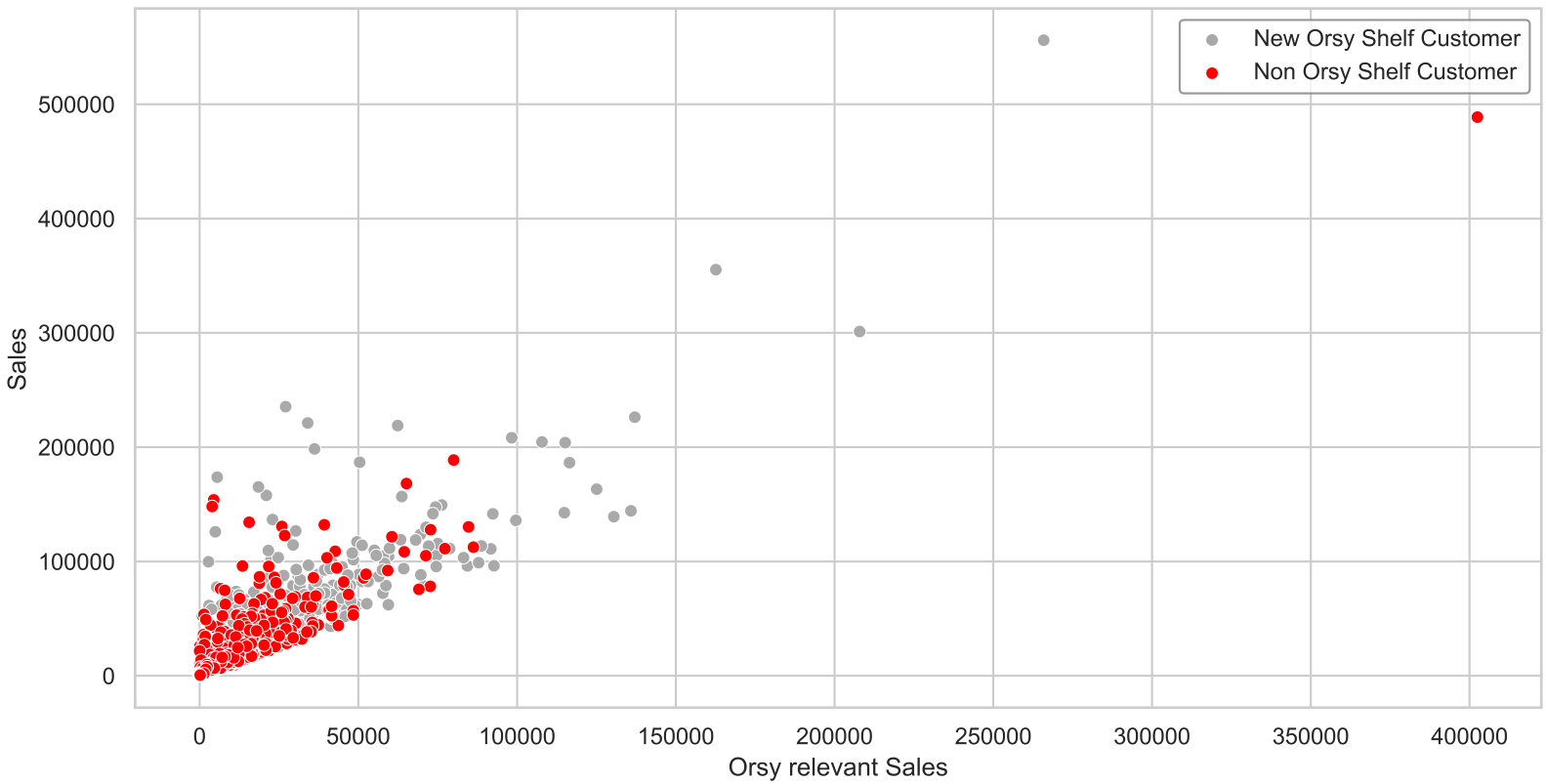


Figure 1: Scatterplot of Sales and ORSY relevant Sales for ORSY-Shelf and Non-ORSY-Shelf Customers

Data Processing

After cleaning the data, e.g., removing irregularities and constructing new variables, we processed the data in four different ways:

1. Full Data Set
2. Filtered data set by relevant features
3. Data set with a condition in place (exclude customers with a current dunning level > 2) to exclude non-suitable customers
4. Filtered and conditioned data set

Further, it is possible to automatically choose from different scalings (e.g., no transformation, standardization, and normalization) of the data. This data transformation is necessary for the various models used in our approach.

Variable Selection

The most simplistic identification criterion of potential features is the correlation. We opted for a threshold of ≥ 0.2, meaning, we only select variables with at least a weak correlation with the target variable.

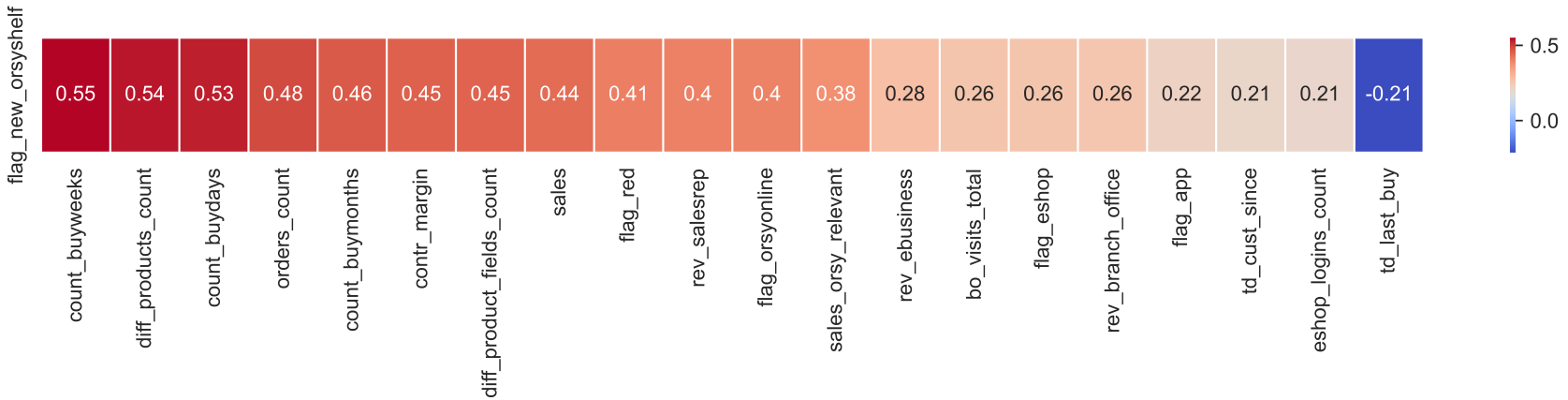


Figure 2: Correlation of the target variable

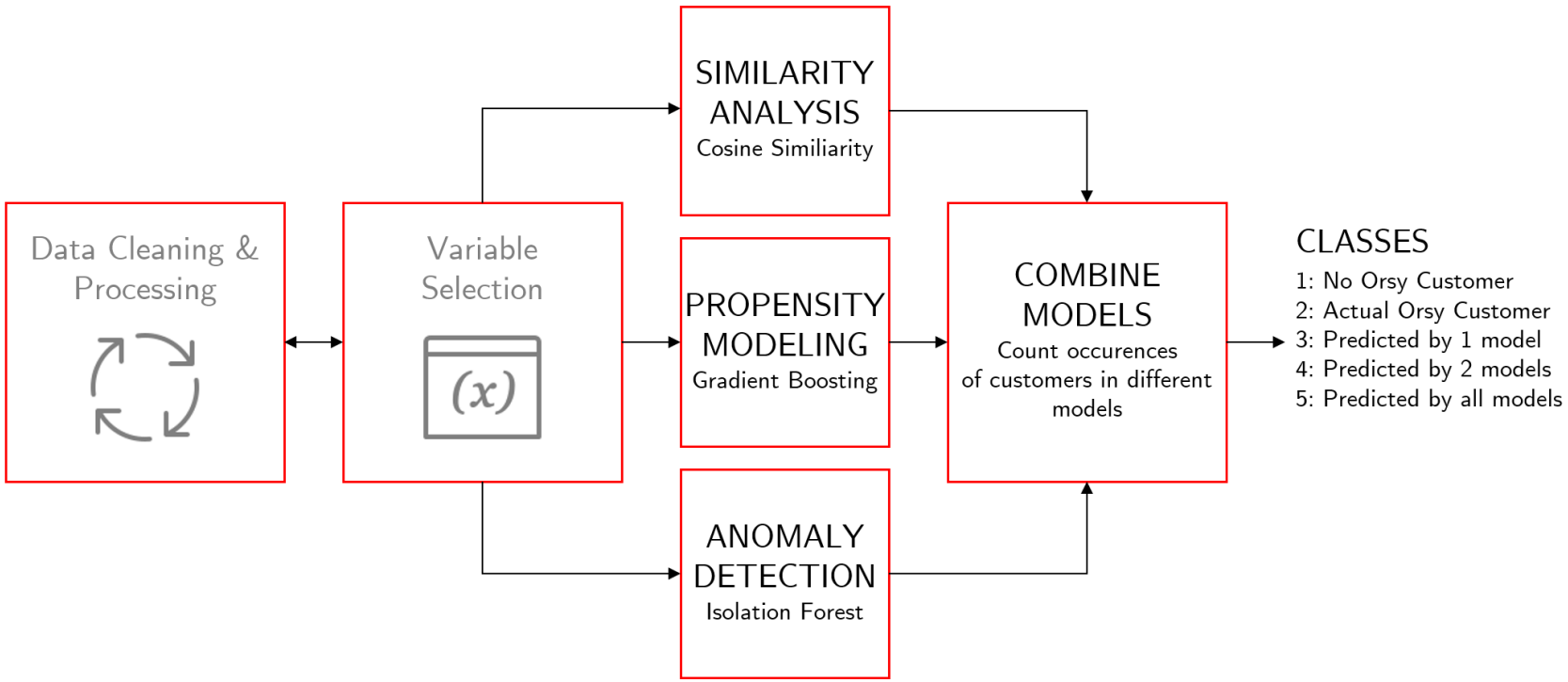
A cross-check with Lasso yields 47 identified features, with 19 matching features with the correlation method.

Approach

Limitations

1. No time dimension in the data
2. Multicollinearity
3. High dimensionality
4. Overfitting
5. Unequal distribution target label

Procedure


Data Cleaning & Processing
Variable Selection
(x)
SIMILARITY ANALYSIS
Cosine Similiarity
PROPENSITY MODELING
Gradient Boosting
ANOMALY DETECTION
Isolation Forest
COMBINE MODELS
Count occurences of customers in different models
CLASSES
1: No Orsy Customer
2: Actual Orsy Customer
3: Predicted by 1 model
4: Predicted by 2 models
5: Predicted by all models

Similarity Analysis - Cosine Similarity

Cosine Similarity is a mathematical method for measuring how similar two vectors are by calculating the cosine of the angle between them.

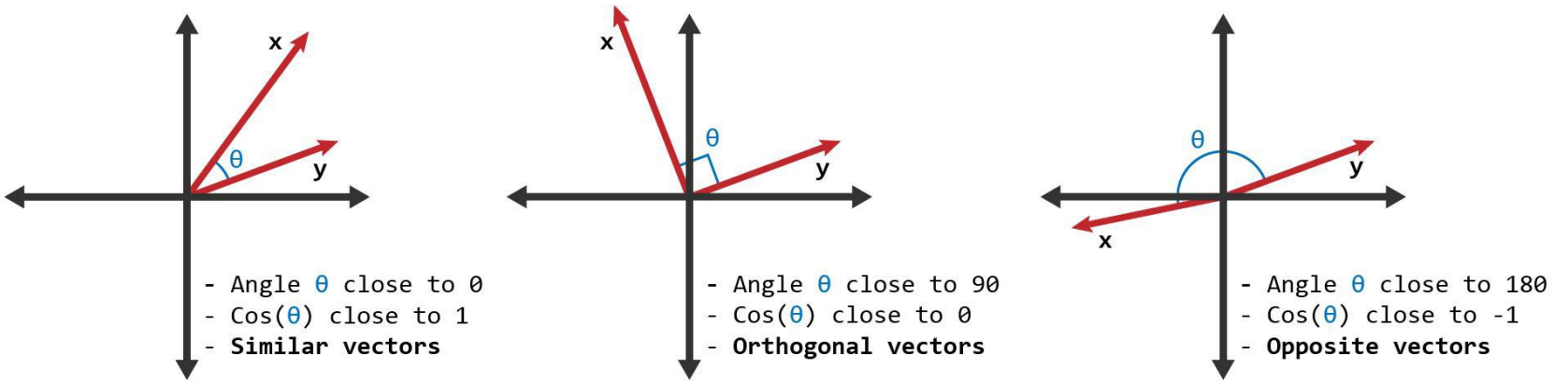


Figure 3: Cosine Similarity

Source: https://www.learndatasci.com/glossary/cosine-similarity/

Similarity Analysis - Cosine Similarity

Advantages:

+ Effective in sparse and high-dimensional data
+ Effective in unbalanced data sets

Disadvantages:

– Does not learn from the data
– Necessitates pairwise calculation

To tackle the drawback of only being able to compute pairwise cosine similarities, we calculated these similarities for each non-customer against all customers. Then, we counted how often each non-customer exceeded a predefined threshold, providing a concise measure of potential customer alignment.

Propensity Modeling - Gradient Boosting

There is a broad range of propensity models. We used a Logistic Regression, Random Forest, Gradient Boosting and Support Vector Machine and tested which models can predict potential ORSY-Shelf Customers with a high accuracy.

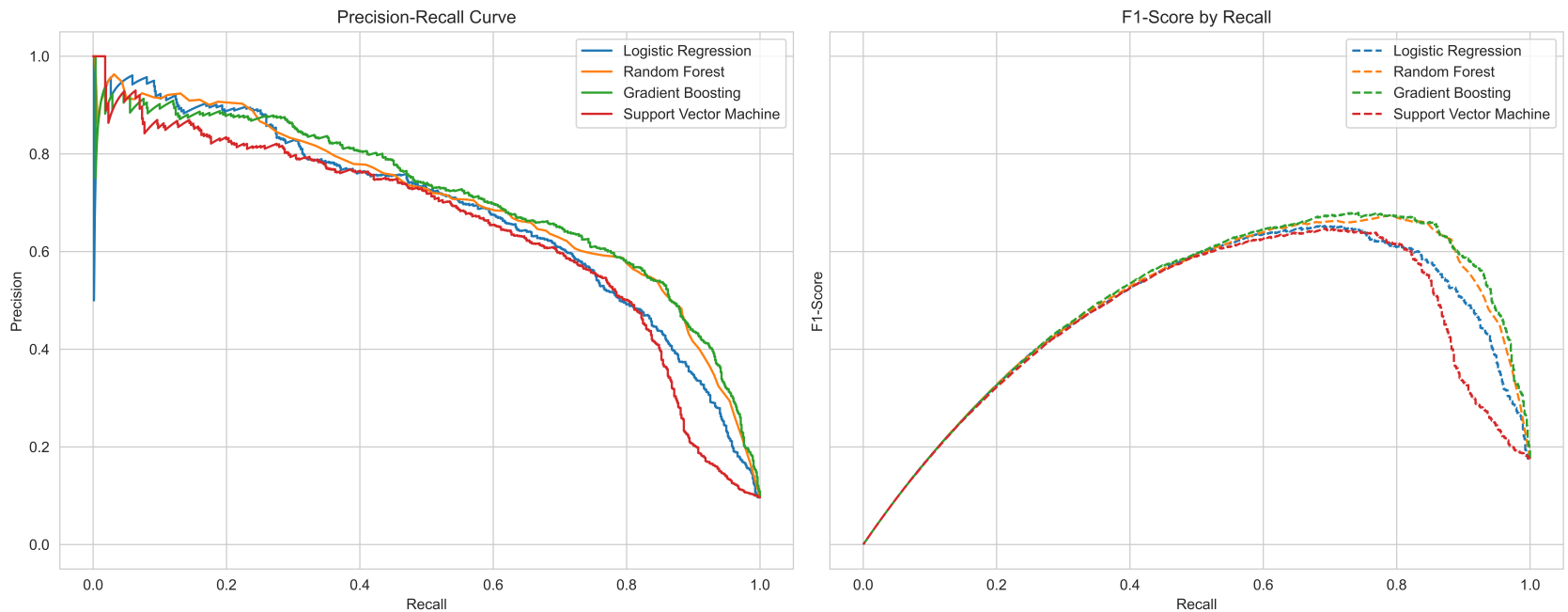


Figure 4: Model Evaluation

Propensity Modeling - Gradient Boosting

Advantages:

+ Good with multicollinearity.
+ Resilient and robust method to outliers.

Disadvantages:

– May overfit within small datasets.
– Computational expensive with large datasets.

Propensity Modeling is prone to overfitting within small datasets. Hence, the model is unsuitable as a single benchmark model and needs additional support to yield robust results.

Anomaly Detection - Isolation Forest

The Isolation Forest is an unsupervised machine-learning algorithm for anomaly detection.

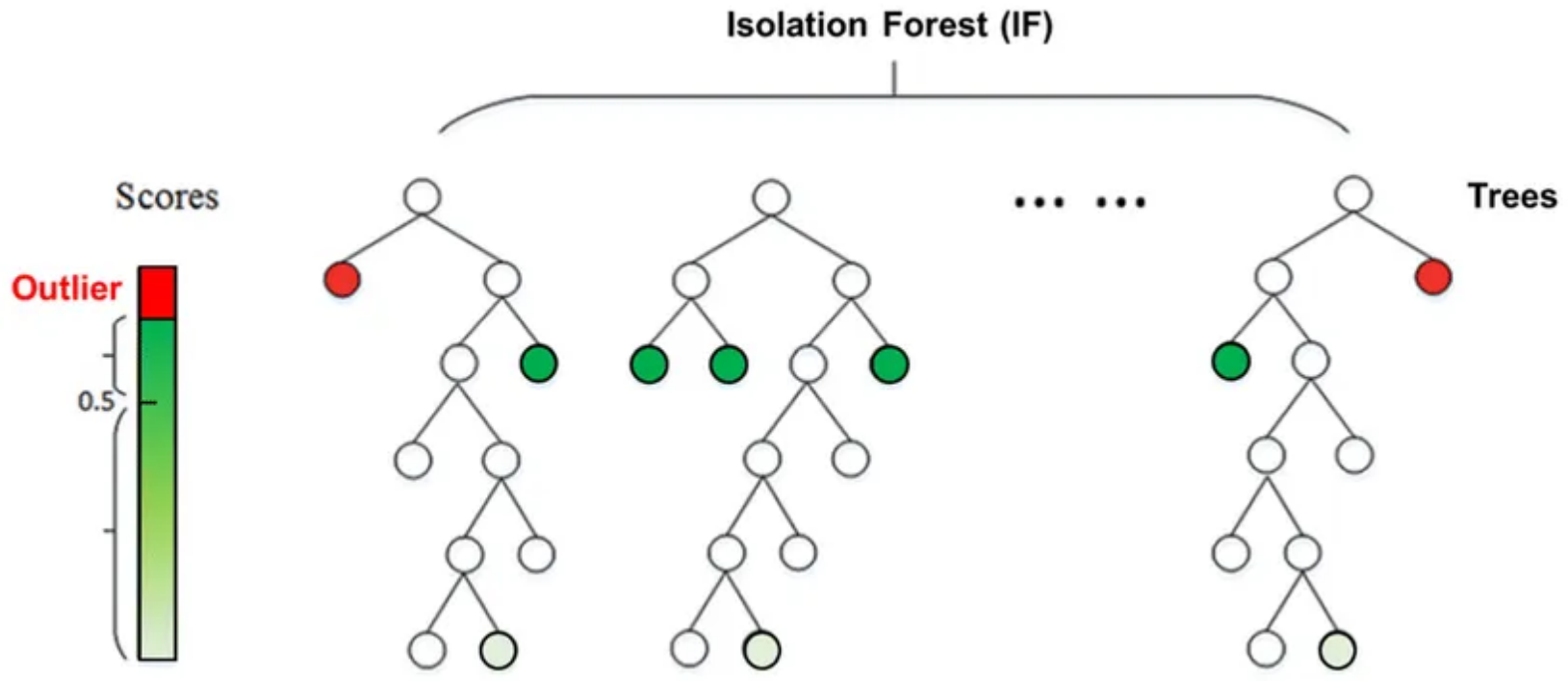


Figure 5: Isolation Forest

Source: www.researchgate.net/figure/Overview-of-the-isolation-forest-method-Light-green-circles-represent-common-normal_fig3_341629782

Anomaly Detection - Isolation Forest

Advantages:

+ Good at detecting anomalies without prior knowledge.
+ Effective in high-dimensional datasets.

Disadvantages:

– Randomness in algorithm.
– Performance suffers from too many irrelevant features.

To mitigate the disadvantages the method is inizialized over 100 random states and different contamination levels. Only customers who were identified at least 30% of the time in all four datasets are considered outliers, and thus, identified as potential customers.

Results

What do these identified Customers have in common?

On average, a typical potential ORSY-Shelf Customer ..

 .. is a long-time customer.

 .. buys frequently.

 .. is at least considered a medium sized customer.

 .. has a much higher contribution margin.

 .. has a high order count.

 .. buys different products in many product fields.

 .. visits stores and the webshop frequently.

A typical potential ORSY-Shelf Customer buys frequently

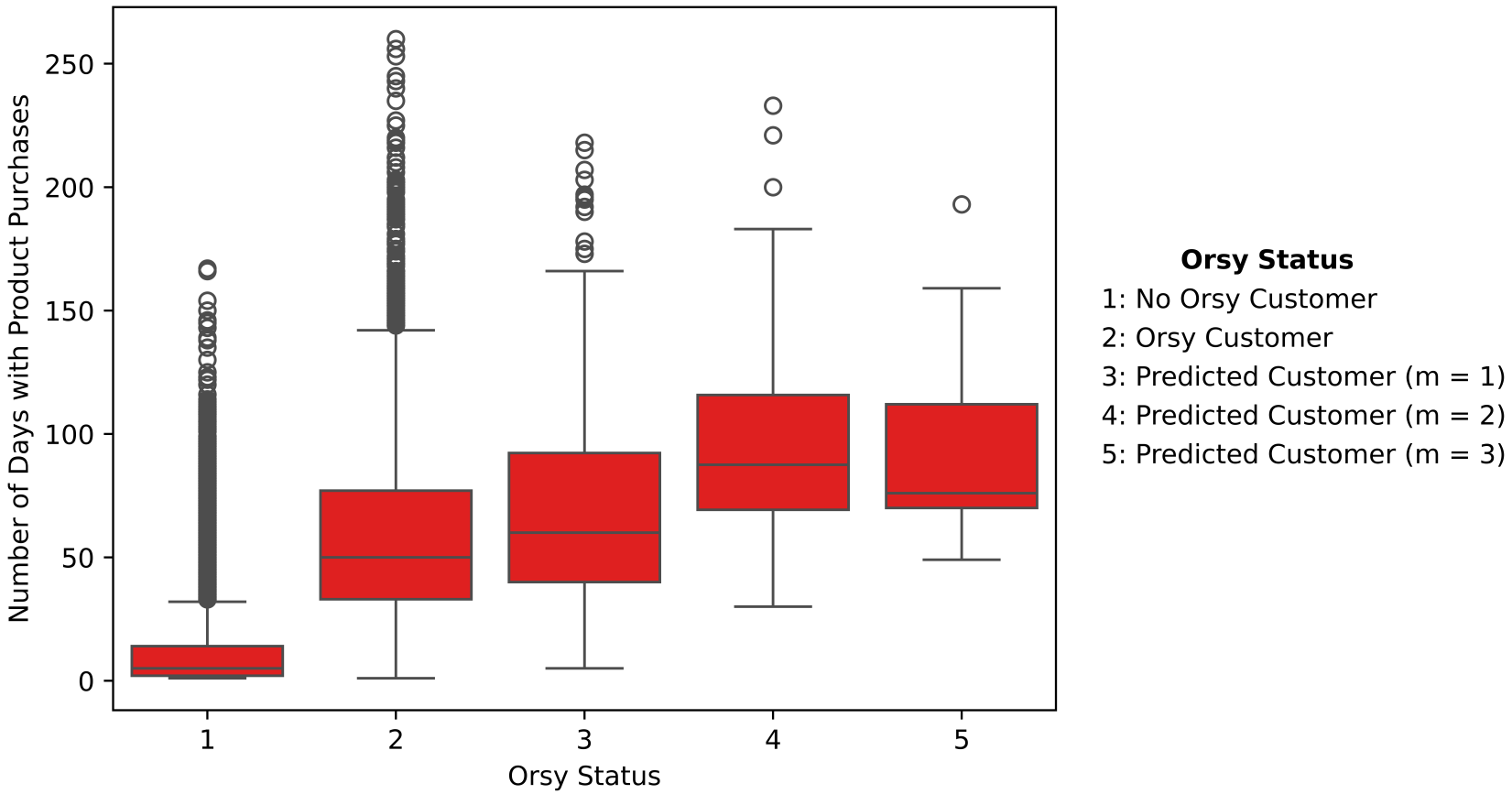


Figure 6: Boxplot of the Number of Days with Product Purchases by ORSY Status

A typical potential ORSY-Shelf Customer has a much higher contribution margin

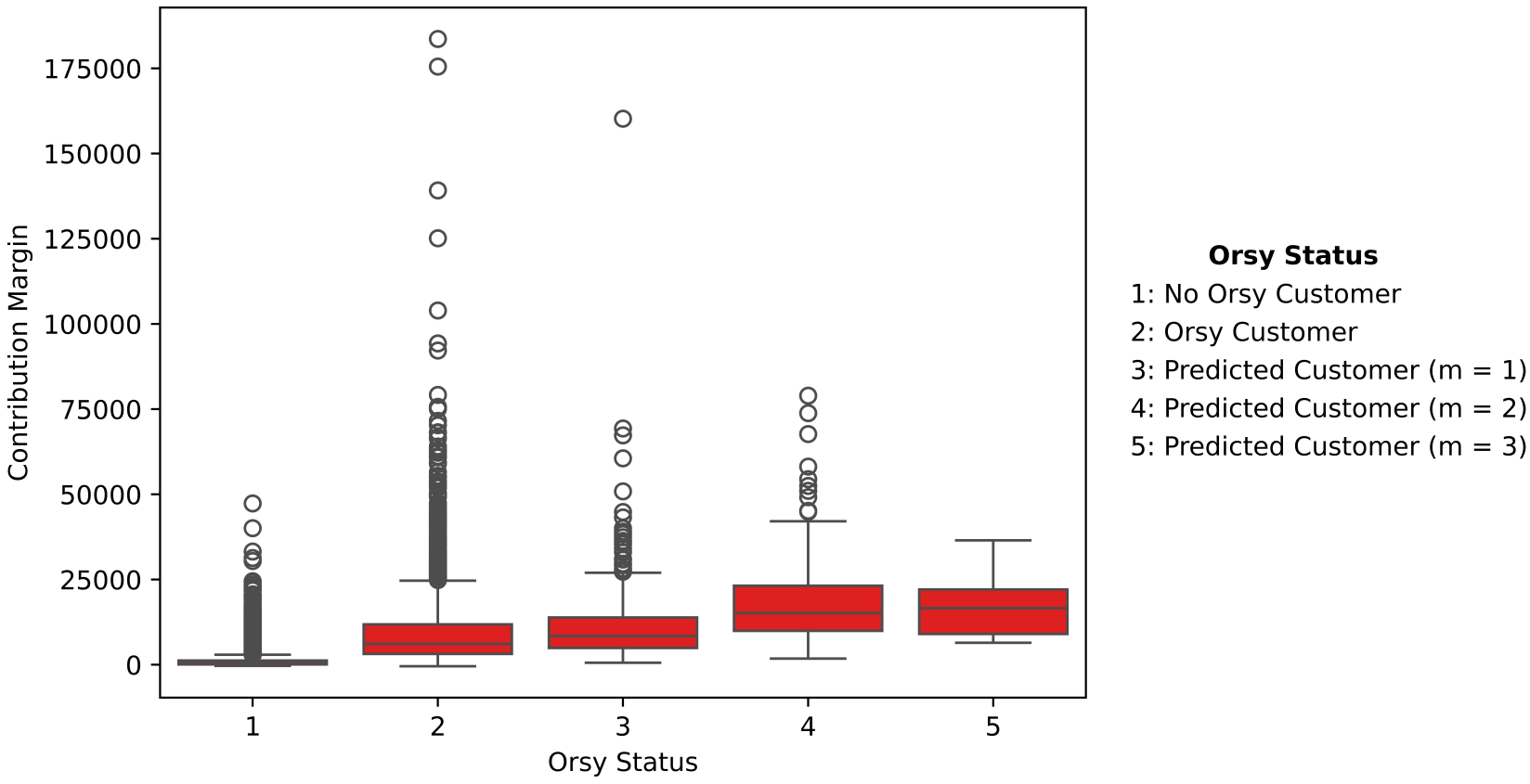


Figure 7: Boxplot of the contribution margin by ORSY Status

A typical potential ORSY-Shelf Customer has a high order count

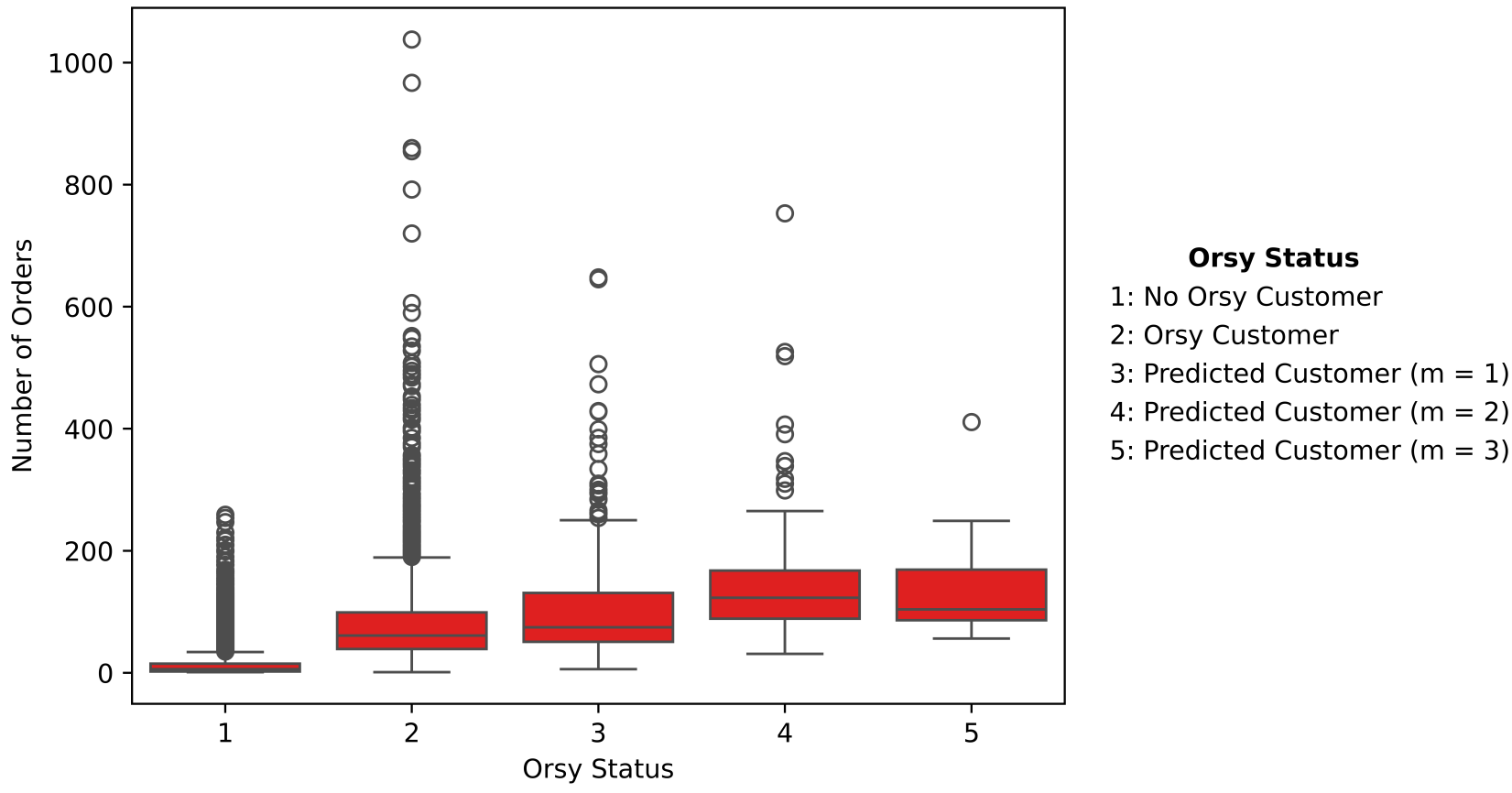


Figure 8: Boxplot of the order count by ORSY Status

A typical potential ORSY-Shelf Customer buys products in many product fields

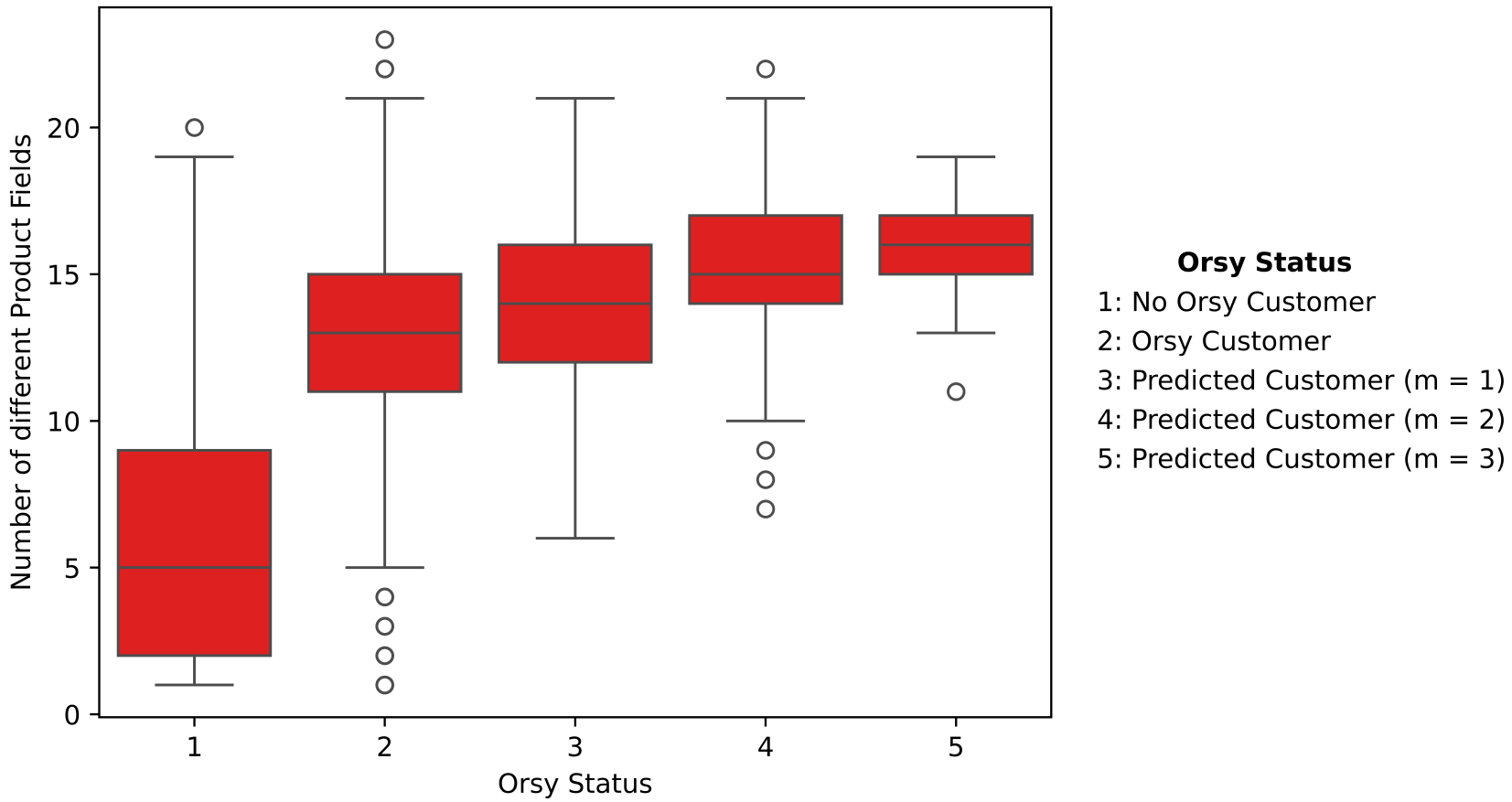


Figure 9: Boxplot of the Number of different product fields by ORSY Status

Comparision of potential ORSY-Shelf Customers at different Customer Contact Points (CCP)

The predicted ORSY-Shelf Customers who mostly use the CCP ..

- .. **Branch Office** bring in **over 1Mio.€ ORSY relevant sales**.
- .. **Sales Representatives** have a **higher share of ORSY relevant sales** than the respective Non ORSY-Shelf Customers.
- .. **Telesales / Internal Staff** generate not many sales but the **highest share of ORSY relevant sales**.

Comparision of potential ORSY-Shelf Customers at different Customer Contact Points

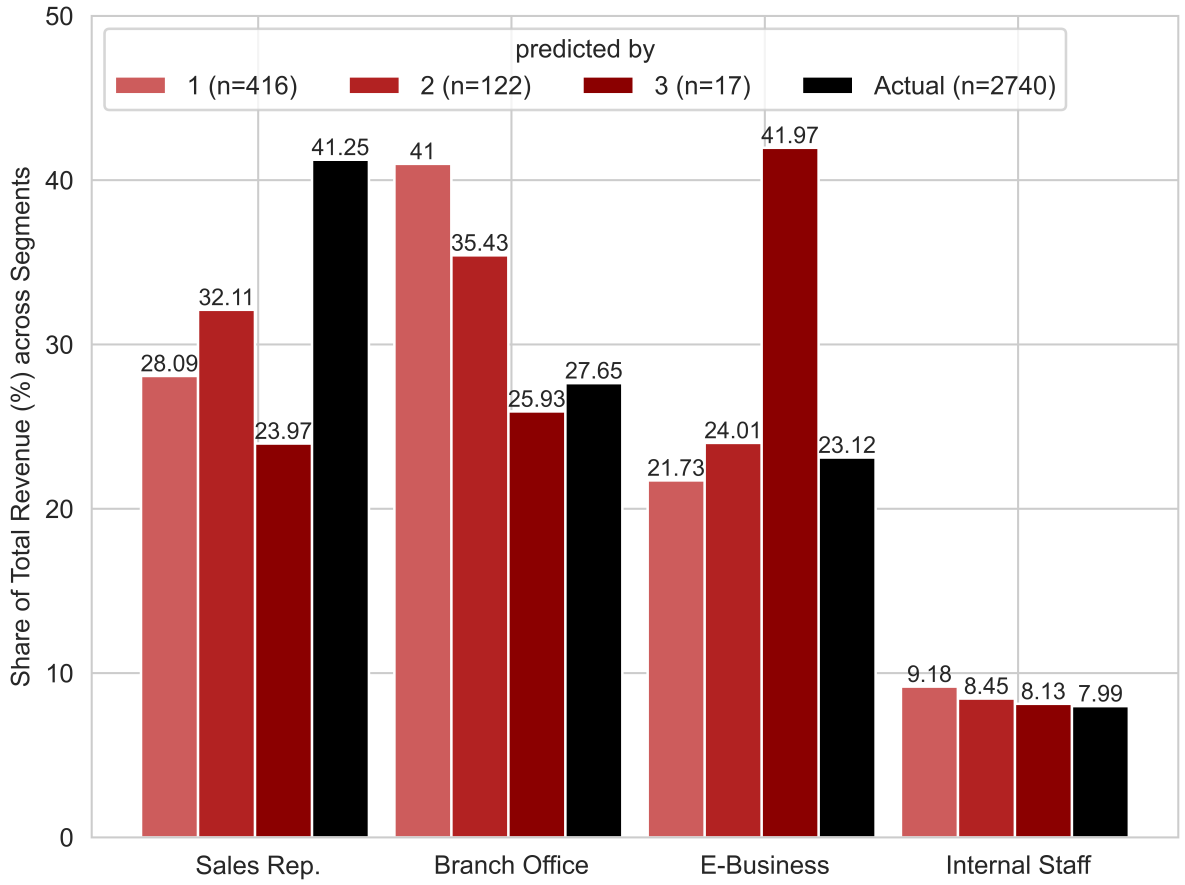


Figure 10: Revenue Share Among Predicted Customers Across Contact Points

Comparision of potential ORSY-Shelf Customers at different Customer Contact Points

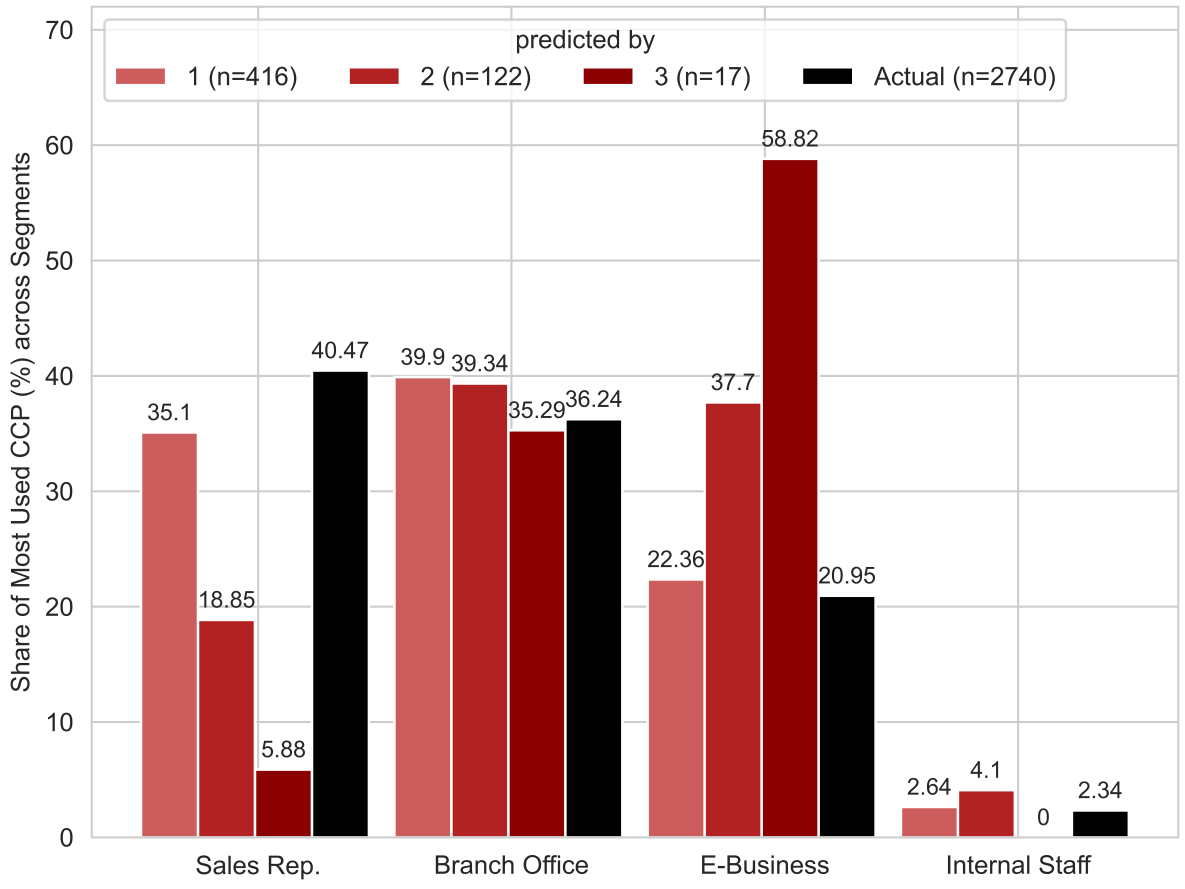


Figure 11: Most Used CCP Among predicted Customers

Recommendations

Recommend an ORSY-Shelf to all customers who ...

bought products on

>50

DAYS

have a

>3000€

CONTRIBUTION MARGIN

Thereby focus on the CCPs ...

E-BUSINESS

- High share of ORSY relevant sales
- Frequently visitors of the E-Shop

BRANCH OFFICE

- Lot of product purchases in Branch Offices
- Most-used CCP