CUSTOMER SENTIMENT:
PREDICTING REVIEW SCORES WITH BIG DATA

Big Data Tools - 2024/2025

ABOUDI Arajem | GAILLARD Paul

TABLE OF CONTENTS

PROJECT OVERVIEW

DATA & FEATURE CREATION

MODELS AND ALGORITHM EVALUATION

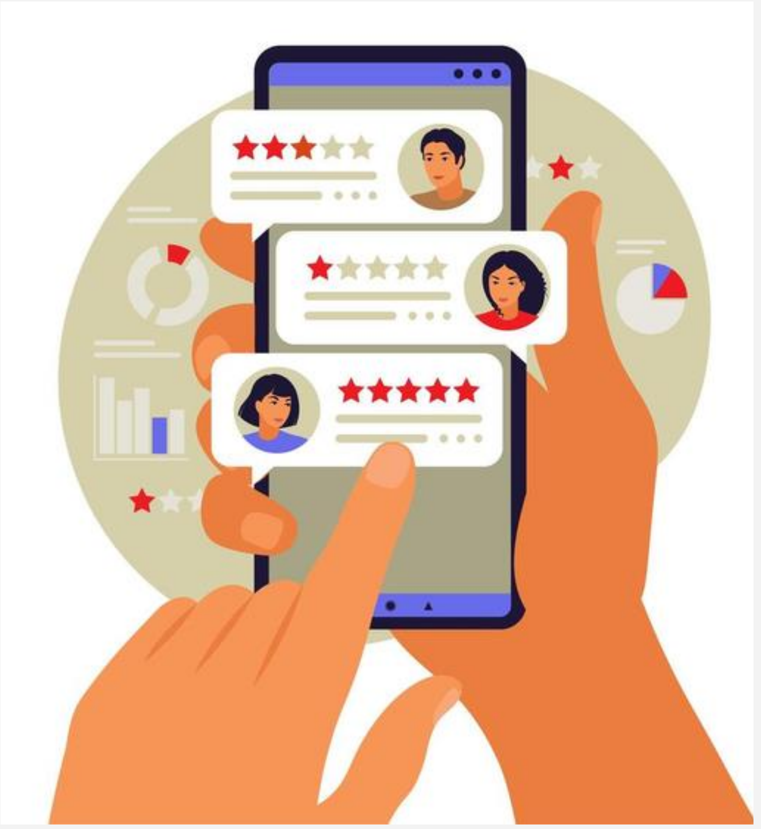
KEY FINDINGS & RECOMMENDATIONS

BLU - ECOMMERCE

BLU offers B2C and B2B customers across 100+ product categories.
It **excels in acquiring new customers** compared to Amazon and C-Discount

BUT

Despite strong customer acquisition, BLU is **losing existing customers to competitors**, impacting its long-term revenue and growth.

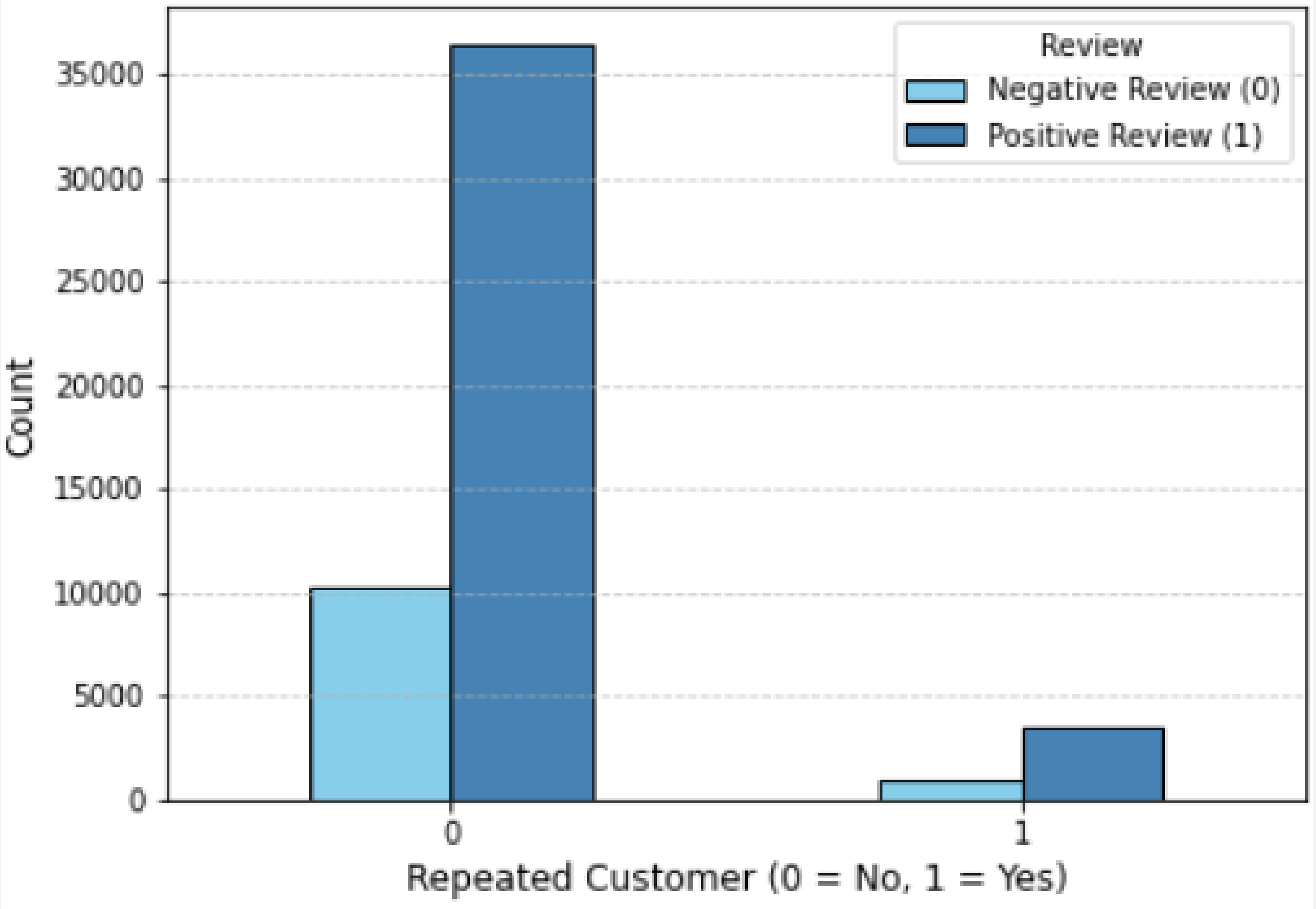


GOAL

Enhance customer satisfaction on orders:
POSITIVE (4-5)
vs.
NEGATIVE (1-3)

DATA ANALYSIS

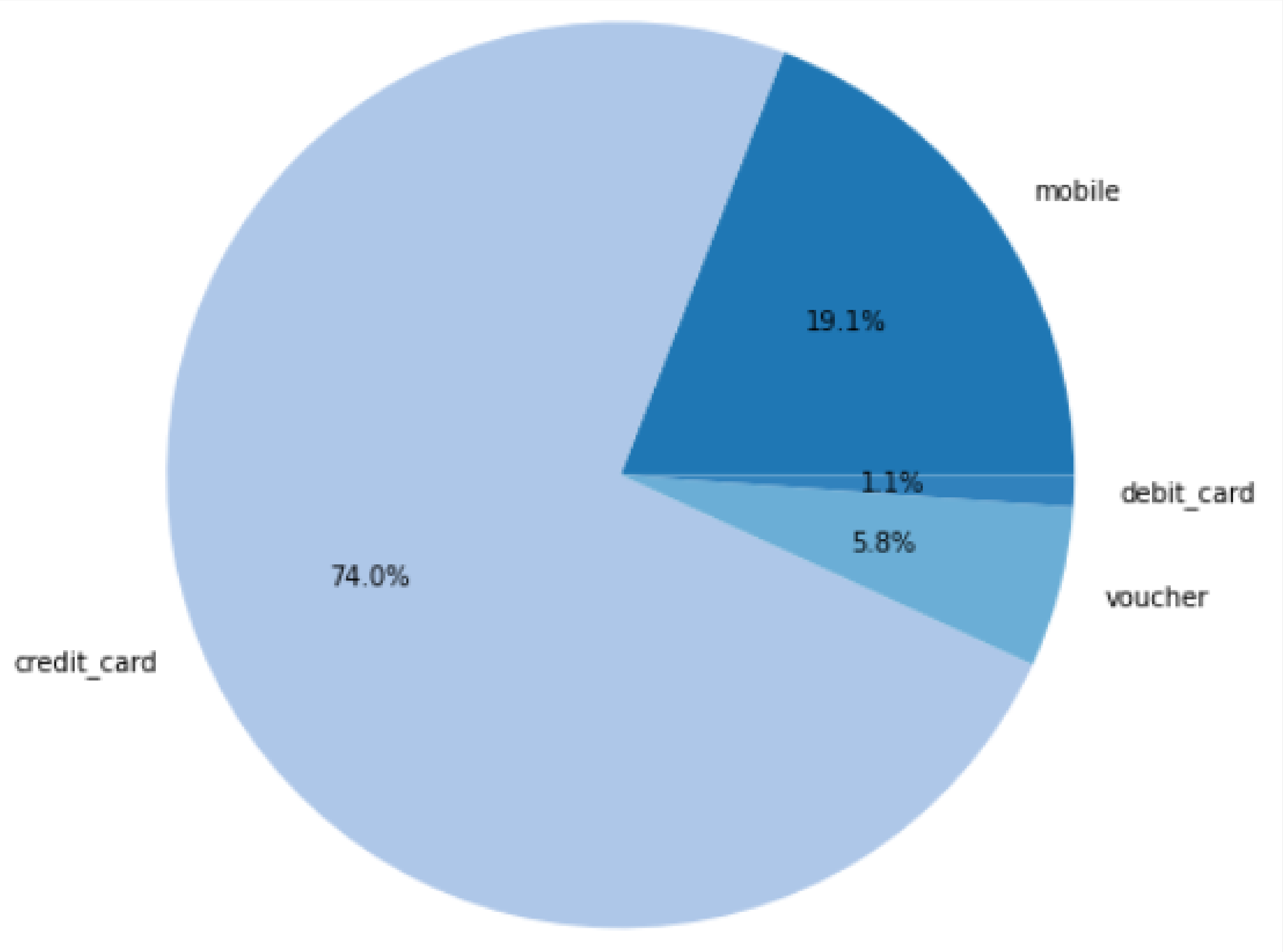
What are the customers doing?



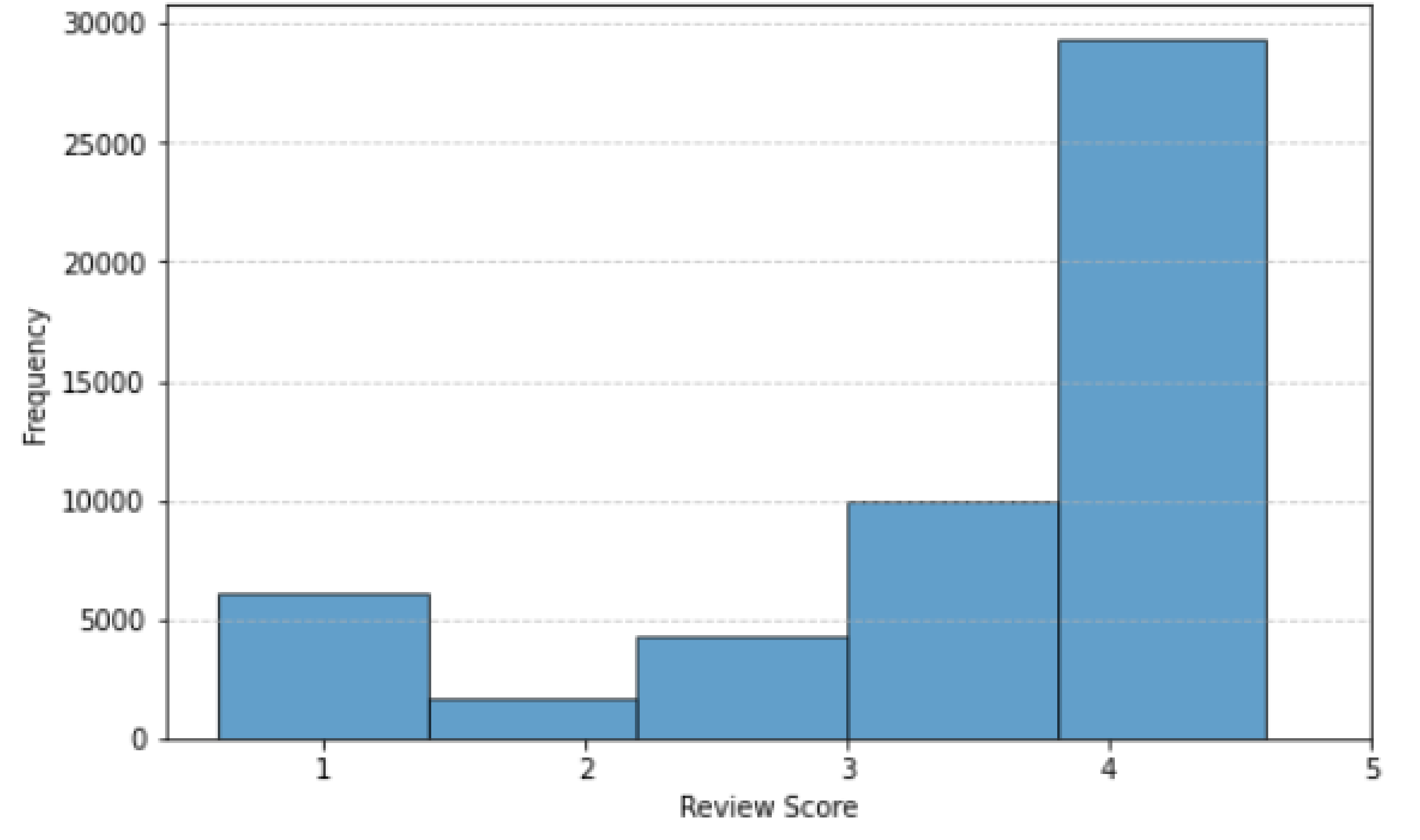
Important Insight:
CUSTOMER RETENTION is linked to positive experiences, **reinforcing** the importance of **loyalty-building strategies**.

DATA ANALYSIS

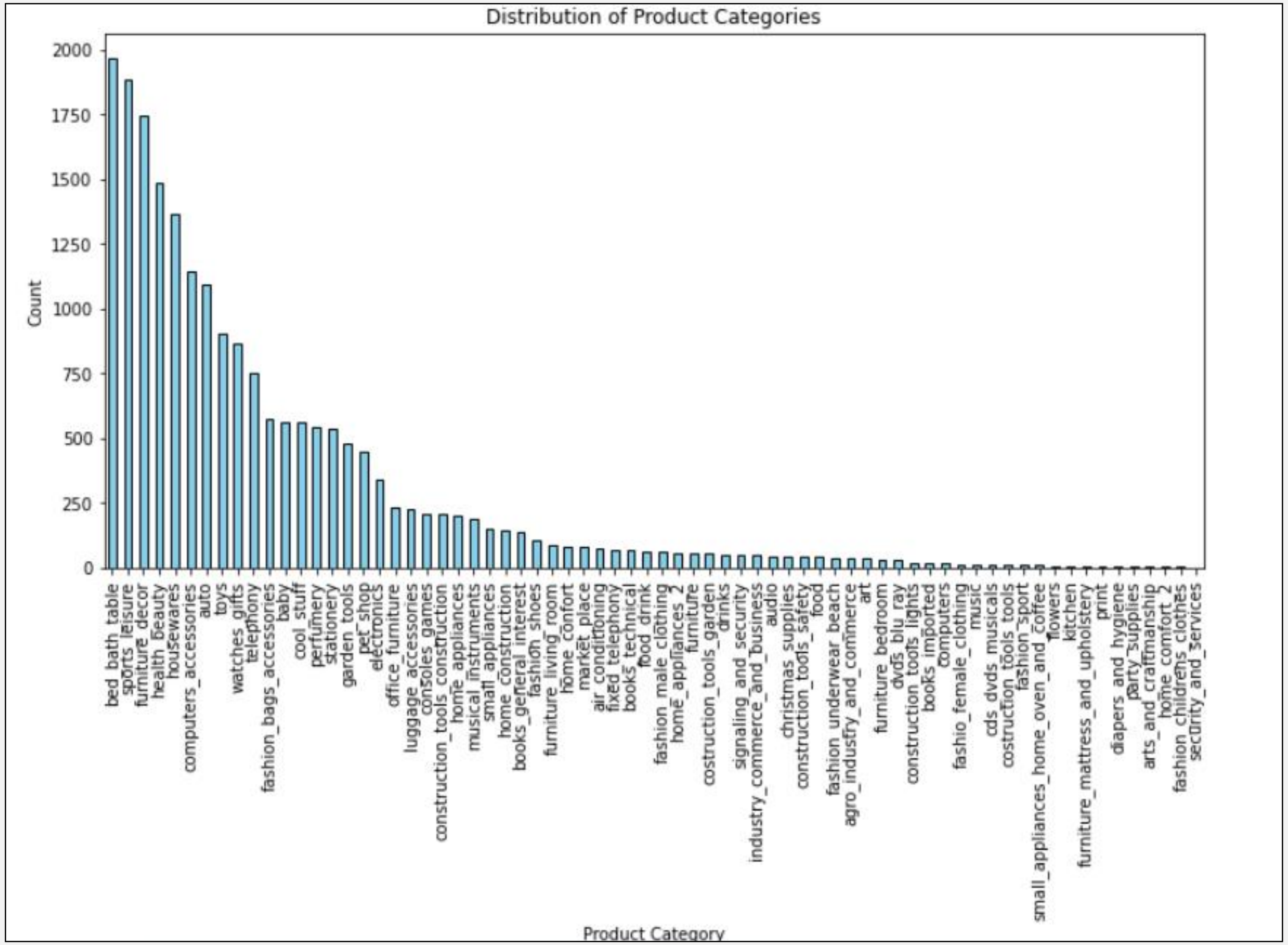
Distribution of payment type



Distribution of Review Score



DATA ANALYSIS



Distribution of the top 5 categories

Product Category Name	Count
bed_bath_table	1964
sport_leisure	1884
furniture_decor	1746
healthy_beauty	1484
housewares	1366

DATA PRE-PROCESSING

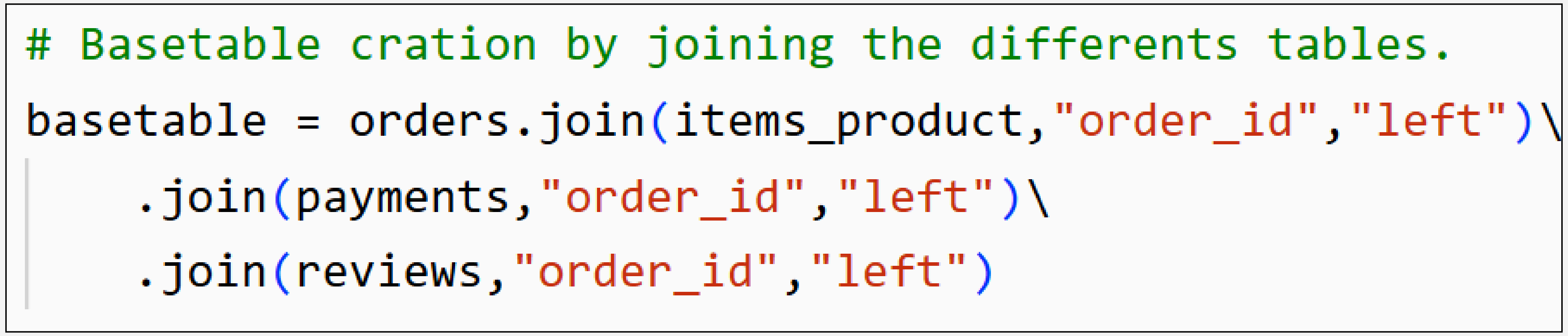
Tables

PRODUCTS
ORDERS
ITEMS
PAYMENTS
REVIEWS

Handling Data

MISSING VALUES
NULL VALUES
DATA CONVERSION
DUPLICATES TREATMENT
FEATURE SCALING

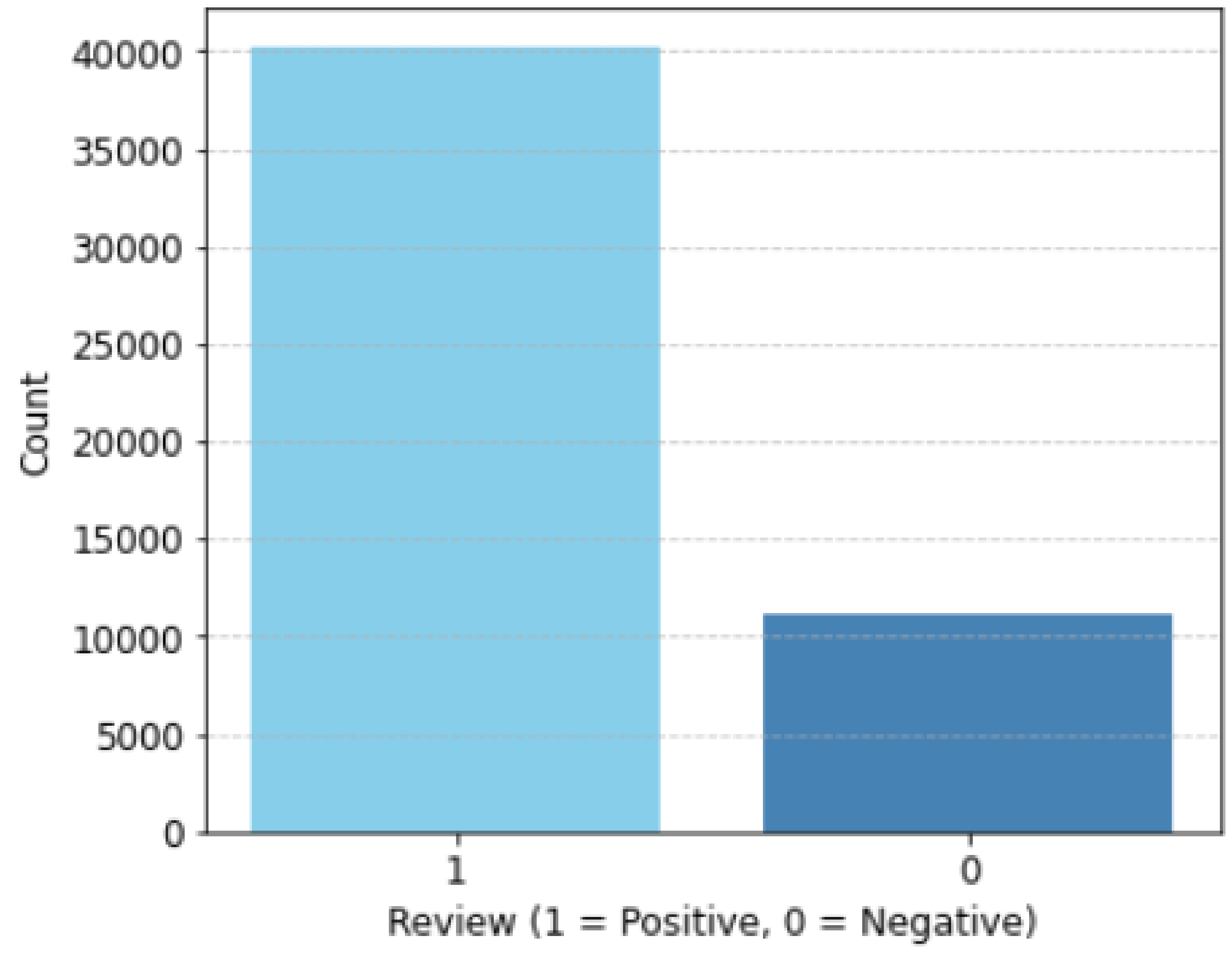
AGGREGATION on ‘ORDER_ID’



```
# Basetable cration by joining the differents tables.
basetable = orders.join(items_product,"order_id","left")\
    .join(payments,"order_id","left")\
    .join(reviews,"order_id","left")
```

FEATURE CREATION

Distribution of Reviews



What IMPACTS reviews?

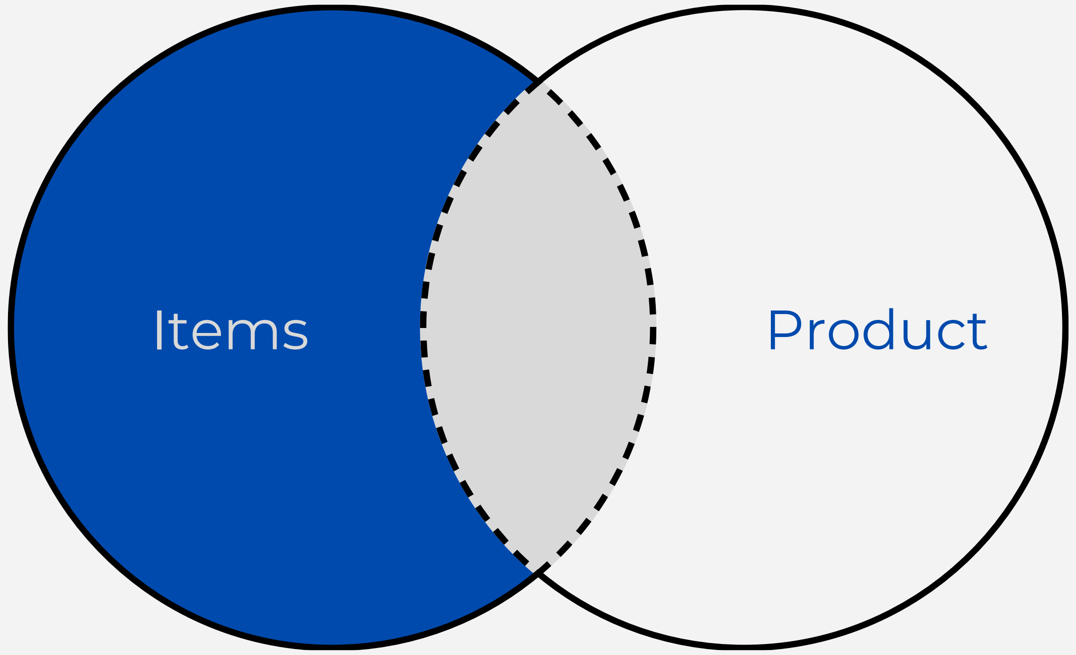
1. Delivery & Logistics
2. Price & Value for Money
3. Product
4.Incentives & Discounts

↓

is_repeated_customer
delivery_duration
approval_duration
delivery_delay
review_responsiveness
size_to_weight_ratio
product_volume

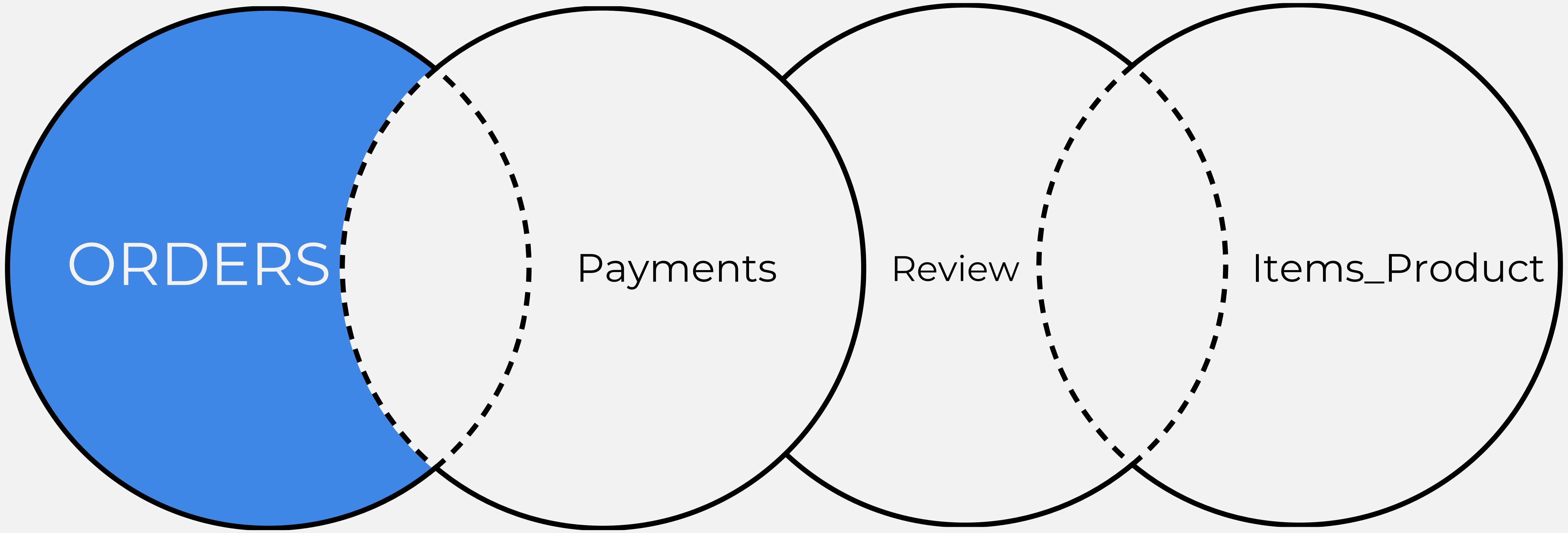
DATA PRE-PROCESSING

JOINING

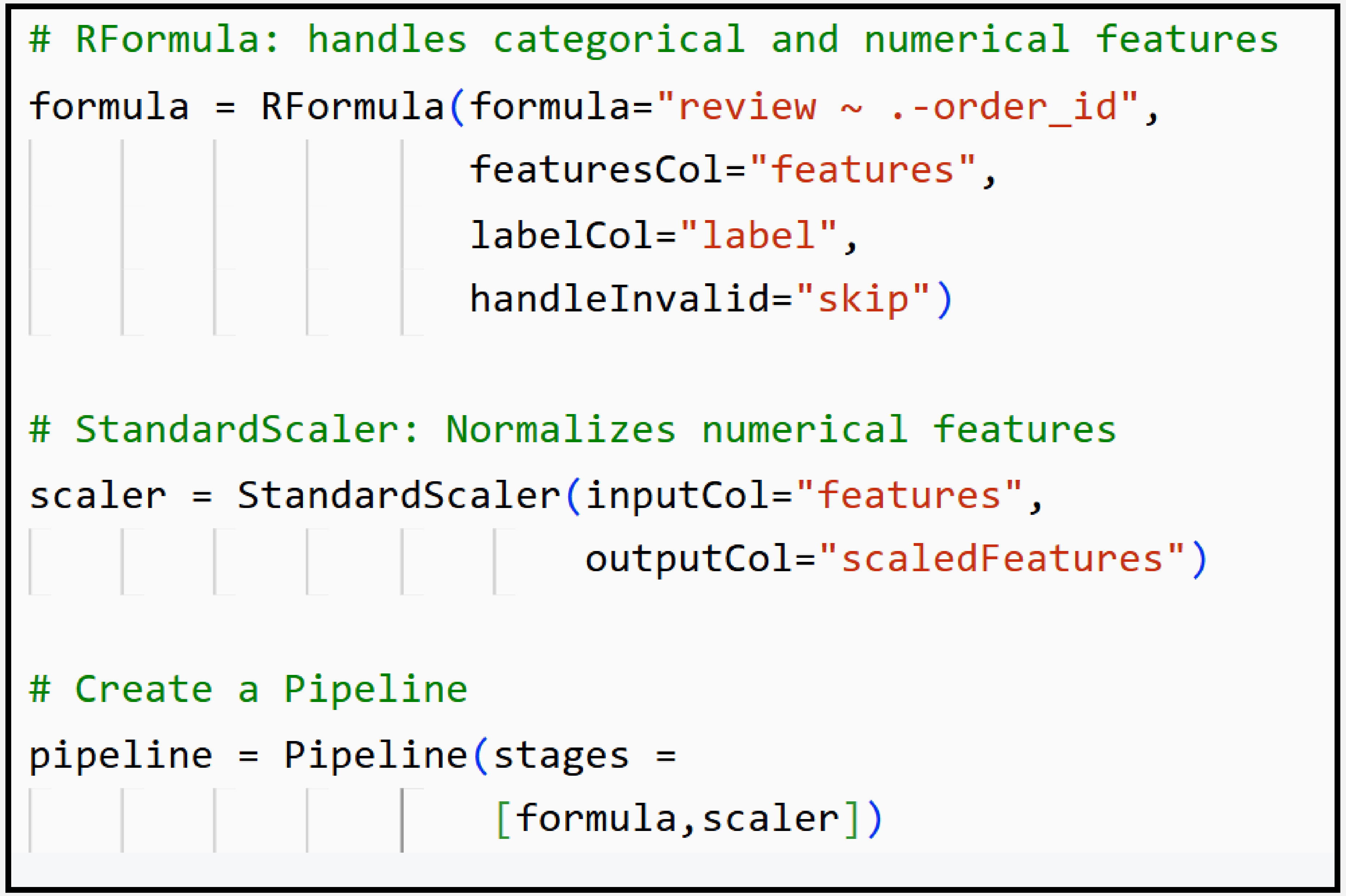


FEATURE CREATIONS

JOINING



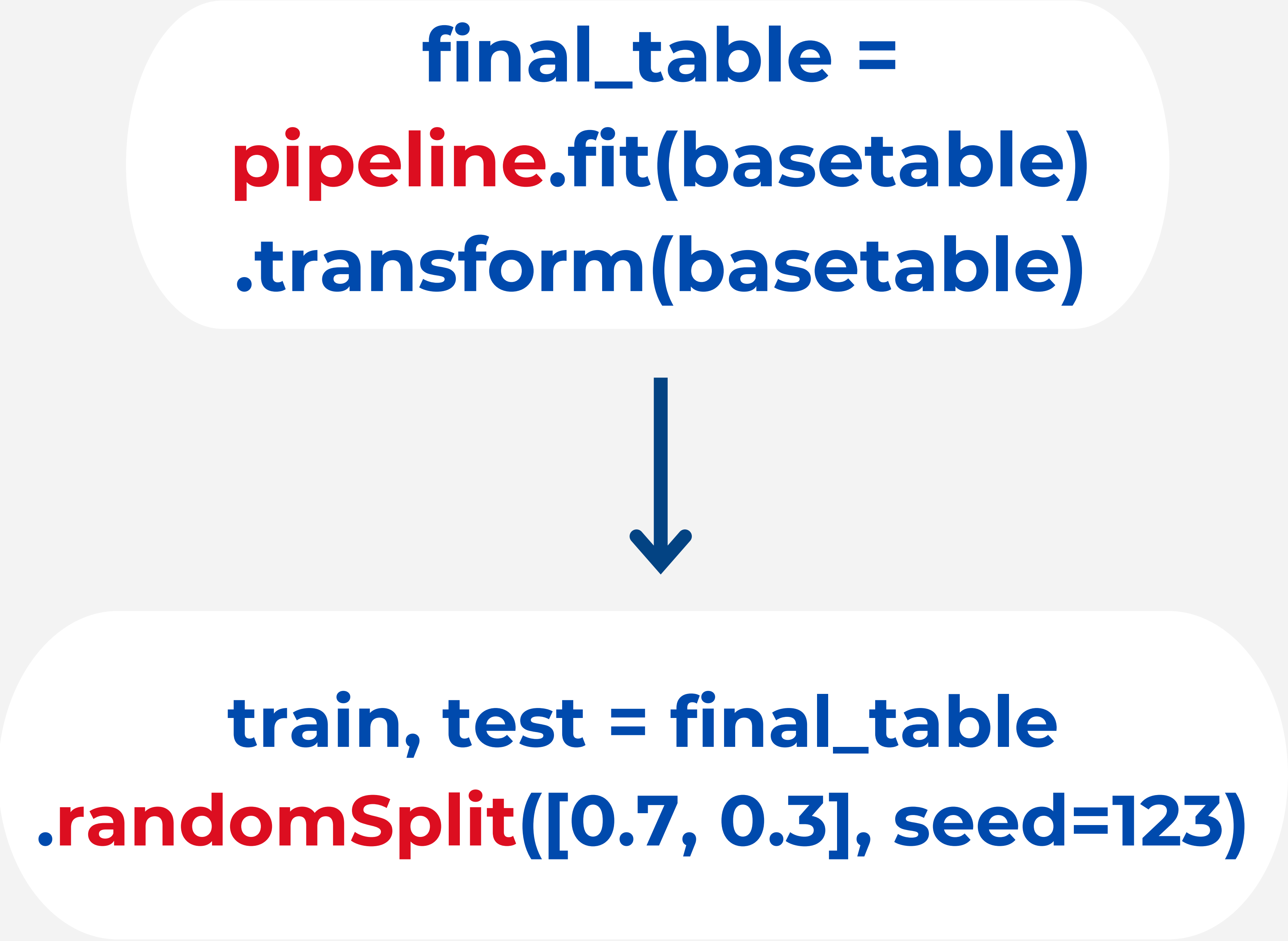
ESTIMATION



```
# RFormula: handles categorical and numerical features
formula = RFormula(formula="review ~ .-order_id",
                   featuresCol="features",
                   labelCol="label",
                   handleInvalid="skip")

# StandardScaler: Normalizes numerical features
scaler = StandardScaler(inputCol="features",
                        outputCol="scaledFeatures")

# Create a Pipeline
pipeline = Pipeline(stages =
                    [formula,scaler])
```



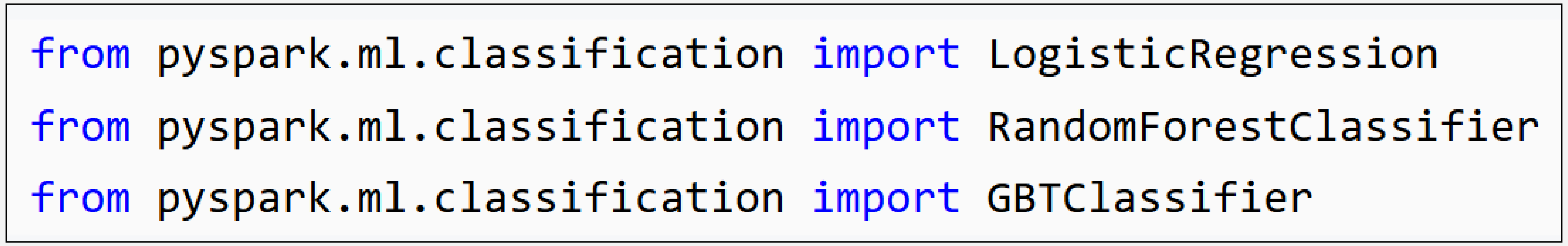
final_table = pipeline.fit(basetable) .transform(basetable)

train, test = final_table .randomSplit([0.7, 0.3], seed=123)

MODELLING PHASE

IT IS A **CLASSIFICATION PROBLEM**

MODEL	Usage
LOGISTIC REGRESSION	Binary Classification that predicts probabilities
RANDOM FOREST	Ensemble model of decision trees to reducing overfitting
GRADIENT BOOSTING	Corrects errors made in decision trees



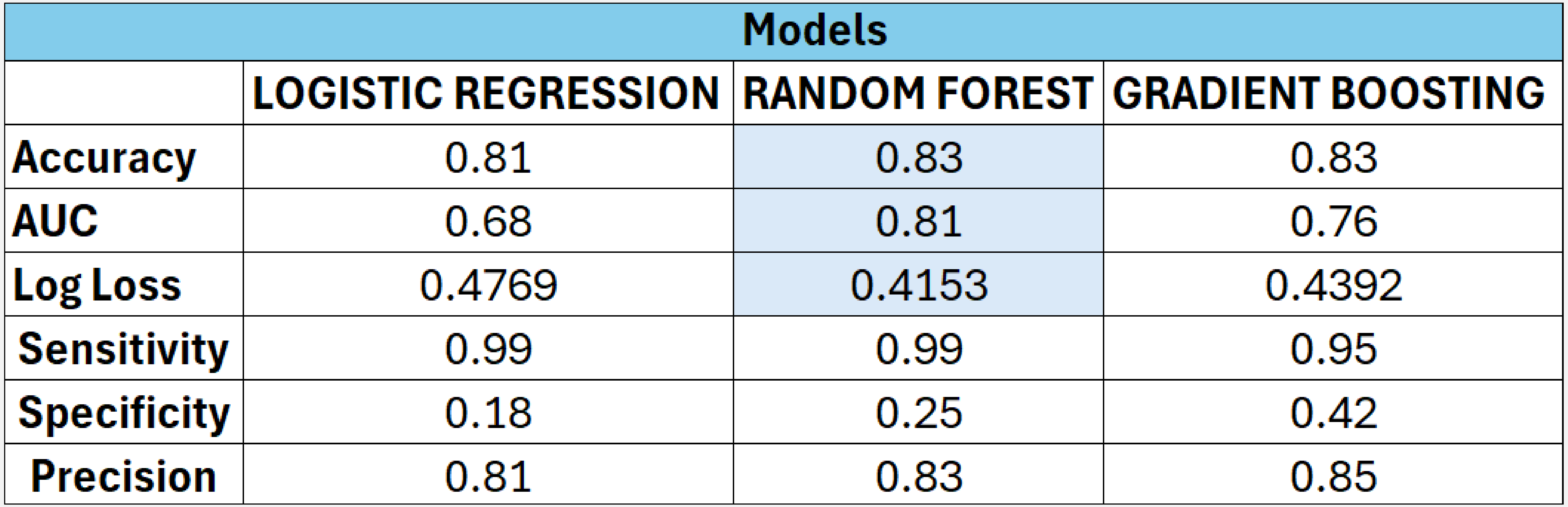
```python
from pyspark.ml.classification import LogisticRegression
from pyspark.ml.classification import RandomForestClassifier
from pyspark.ml.classification import GBTClassifier
```

HOW DO WE CHOOSE THE MOST EFFECTIVE ONE?

AUC

Accuracy

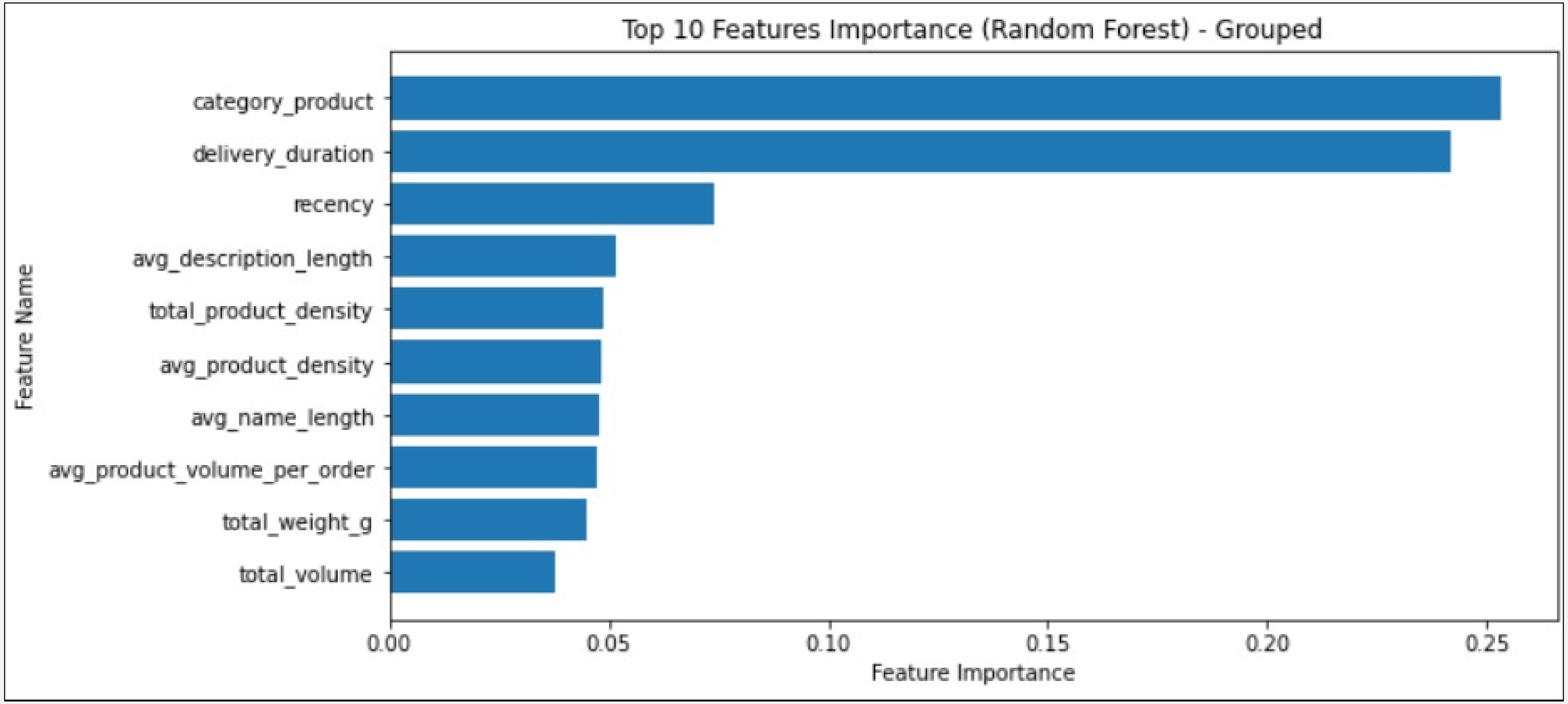
LogLoss



Models			
	LOGISTIC REGRESSION	RANDOM FOREST	GRADIENT BOOSTING
Accuracy	0.81	0.83	0.83
AUC	0.68	0.81	0.76
Log Loss	0.4769	0.4153	0.4392
Sensitivity	0.99	0.99	0.95
Specificity	0.18	0.25	0.42
Precision	0.81	0.83	0.85

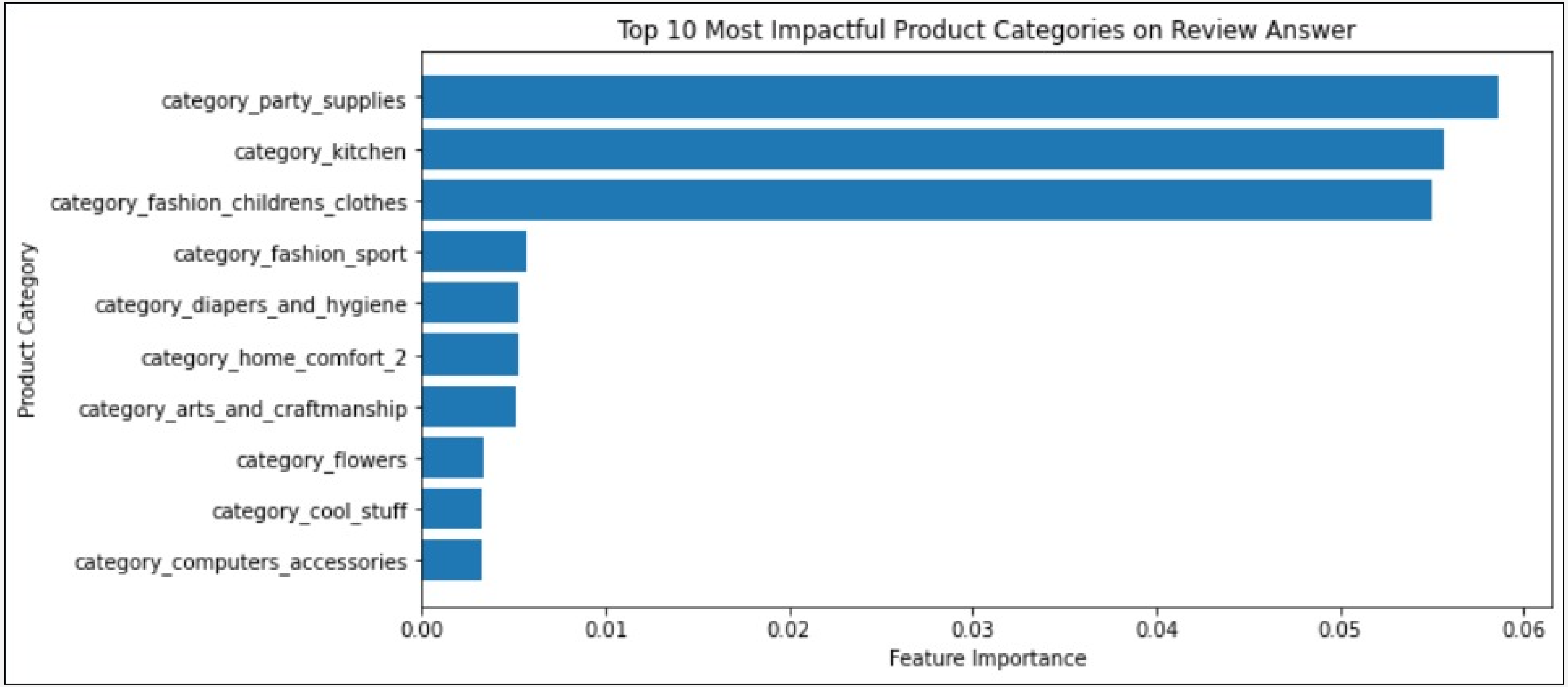
FEATURE SELECTION

RANDOM FOREST



FEATURE SELECTION

More into Detail...



BUSINESS IMPLICATIONS

Our Suggestions

OPTIMIZE DELIVERY & LOGISTICS

Provide accurate delivery estimates & fast shipping

WHITE-GLOVE SERVICE & ASSEMBLY SUPPORT

Offering assembly and setup services for purchased products.

- Furniture & Home Decorations
- Electronics & Appliances
- Fitness Equipment

THANK YOU !