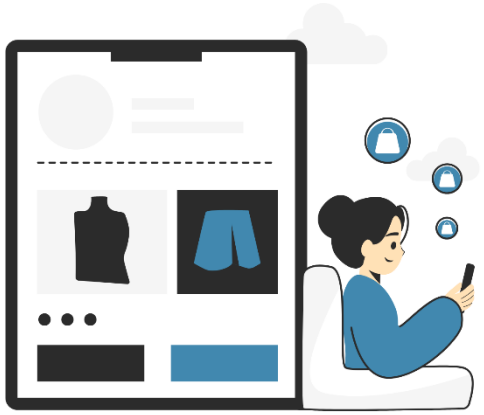
MACHINE LEARNING PROJECT

Brazilian E-Commerce

K-Means Customer Segmentation
Random Forest for Customer Retention Prediction

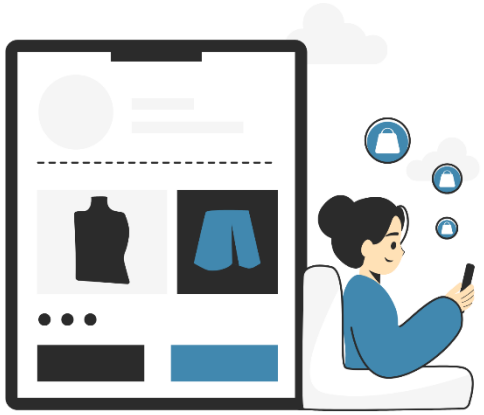
Patrícia Pereira
pereiraar.patricia@gmail.com



olist

BUSINESS UNDERSTANDING

Hypothetical Problem Statement



PROJECT

PROBLEM STATEMENT

The business faces a challenge in optimizing marketing efforts due to an inability to predict which customers are likely to make repeat purchases.

GOAL

Develop a clustering model that identifies customer segments for target market strategies, and a predictive model that identifies customers who are likely to return for purchases, based on their purchase history and behaviour before 2017/05/31.

SCOPE

Ecommerce public dataset of orders made at Olist Store with information of 100k orders from 2016 to 2018 made at multiple marketplaces in Brazil.

Dataset: https://www.kaggle.com/datasets/olistbr/brazilian-ecommerce

OLIST AND MARKETPLACE BENCHMARKING

Customers who **make repeat purchases spend 67% more** than first-time customers [1]

Retention Rates (after 6 months):
Walmart and Target - 14%
Temu - 28%
Amazon – 56% [2]

Less priority on **delivery speed** and more on **delivery reliability** [4]

For marketplace sellers, the **customer retention ratio is 20-40%** [5]

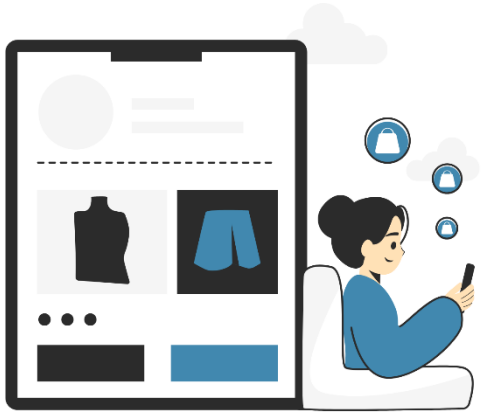
Approximately 90% of transactions involve **disintermediation** after the first purchase [3]

Importance of **shipping costs and flexible delivery options** [4]

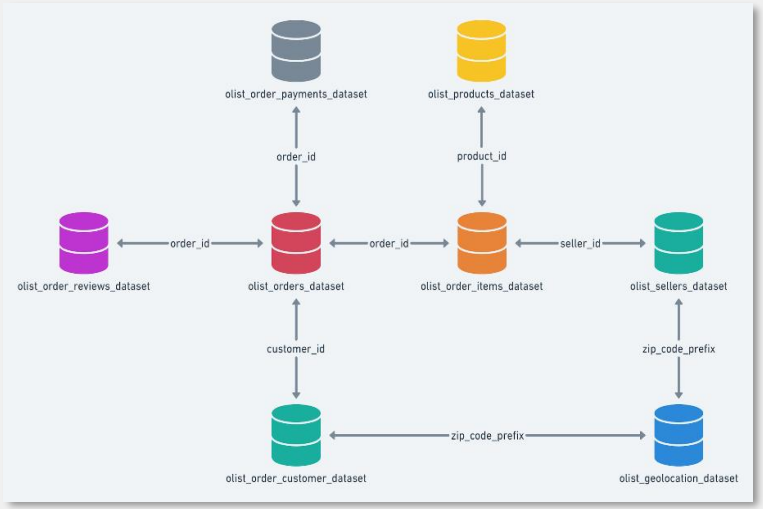
olist

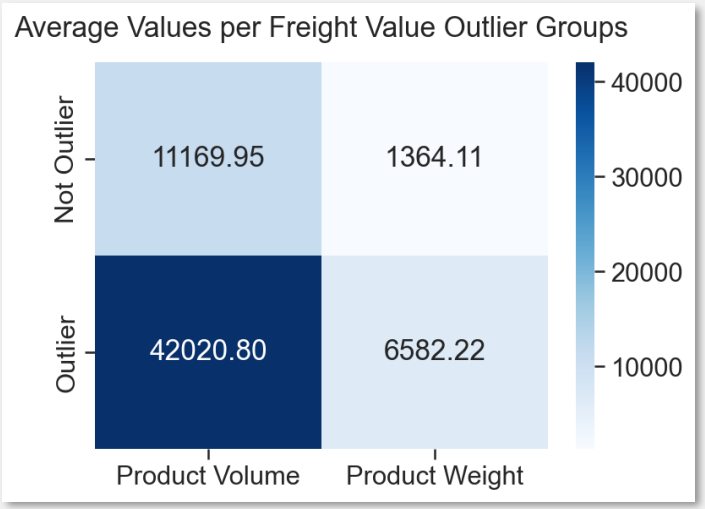
DATA UNDERSTANDING

Initial EDA



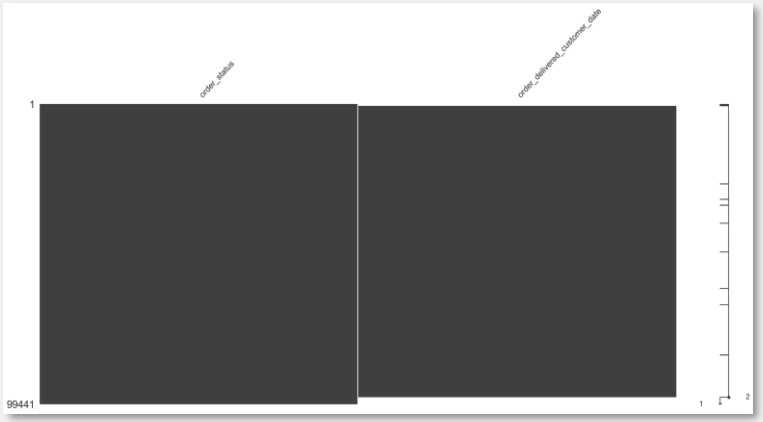
SCHEMA, OUTLIERS AND MISSING VALUES



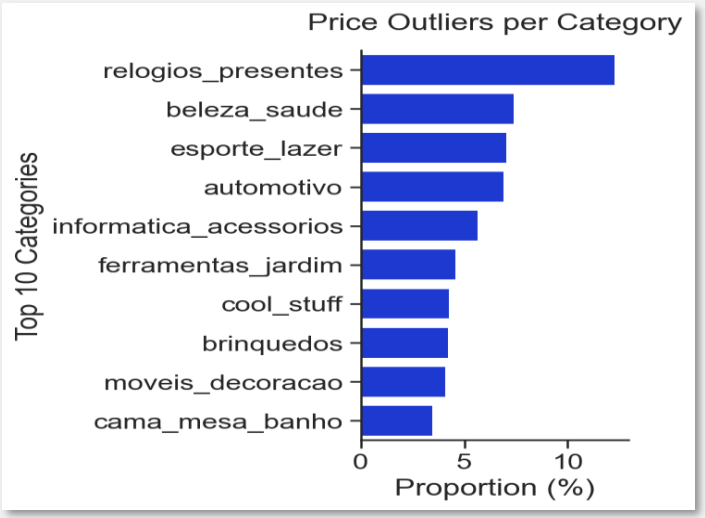


Freight Value Outliers:
above the third quartile plus 1.5 times the interquartile range (IQR).

The distribution of product's volume and weight for the Outliers group is **statistically greater** than the distributions for the Not Outliers group.



Missing dates in the orders dataset can be explained by the order status.



Price Outliers:
above the third quartile plus 1.5 times the interquartile range (IQR).

Price outliers are **present across multiple categories**, suggesting the need for further investigation with company.

INITIAL INSIGHTS

1.04
Average Orders

Each customer made 1.04 orders on average.

1.14
Average Products

Each order has 1.14 units of products, on average.

97%
Delivered

The Delivered Status represents around 97% of all orders.

74%
Credit Card

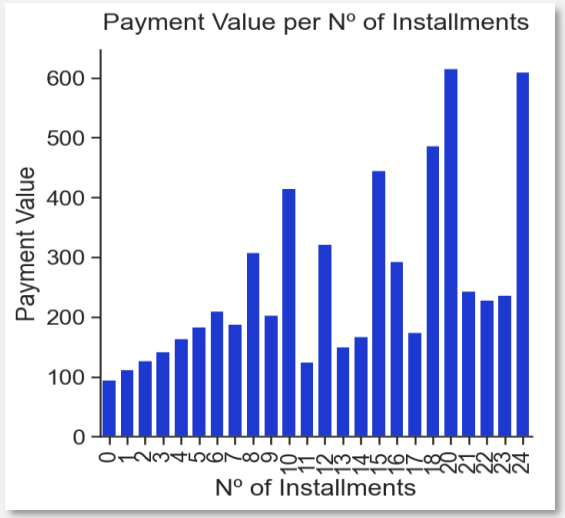
74% of all payments are made with Credit Card.

1176
Black Friday Day

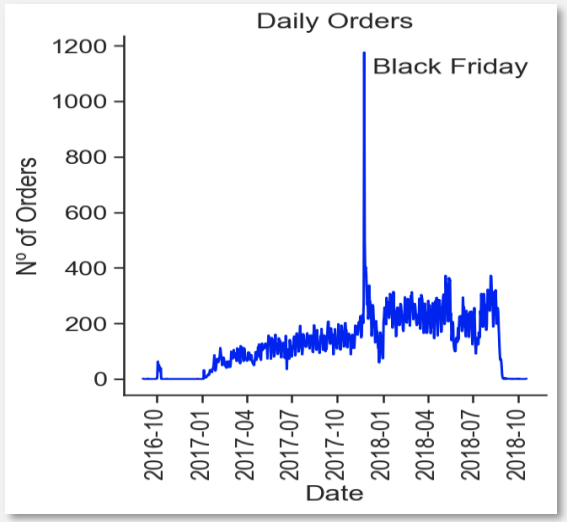
There were 1176 orders on Black Friday day (BF).

305
Black Friday Week

The average orders per day on BF week (without that day) was 305.



The high rate of family debt in Brazil (77.60%) underscores the relevance of this chart, which demonstrates their **tendency to utilize installment plans.** [6]



Black Friday is an excellent tool for a massive **injection of revenue at one time.**

However, it did not prove to be a good tool for increasing customer retention.

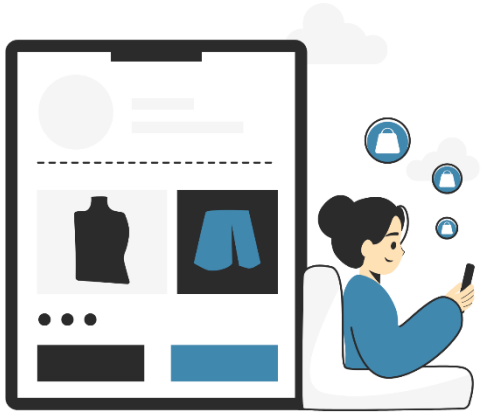
olist

olist

DATA PREPARATION

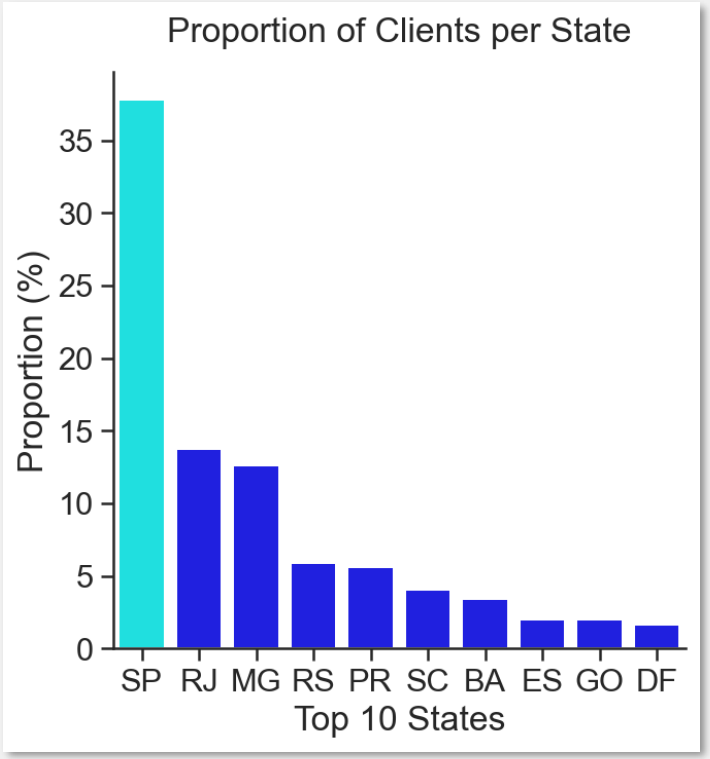
Feature Engineering and EDA for Retention Prediction

Delivered orders until the split date (2017/05/31)



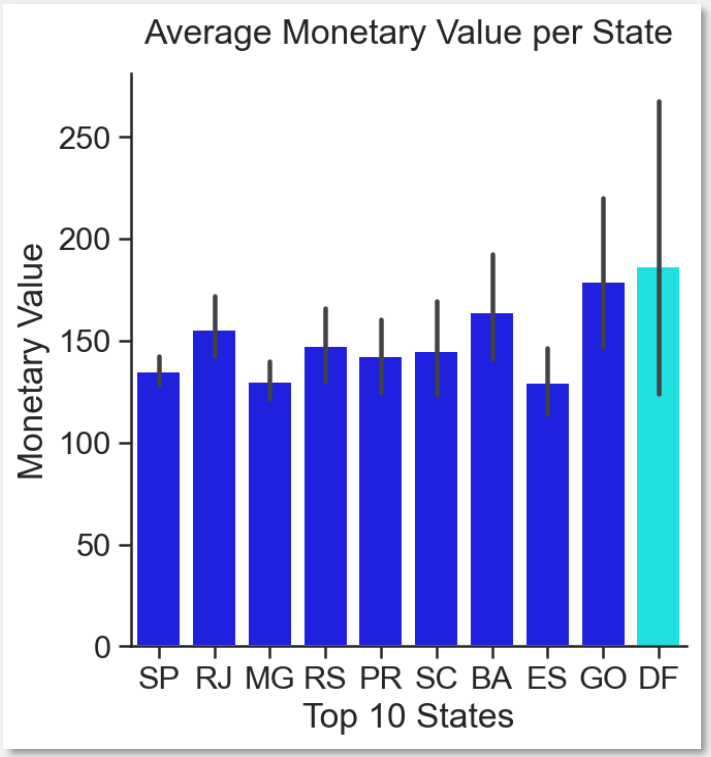
INSIGHTS

Most populous state in Brazil: São Paulo [7]



Approximately 38% of total orders are from São Paulo (SP)

Highest GDP per capita in Brazil: Distrito Federal [8]



The highest average monetary value comes from Distrito Federal (DF)

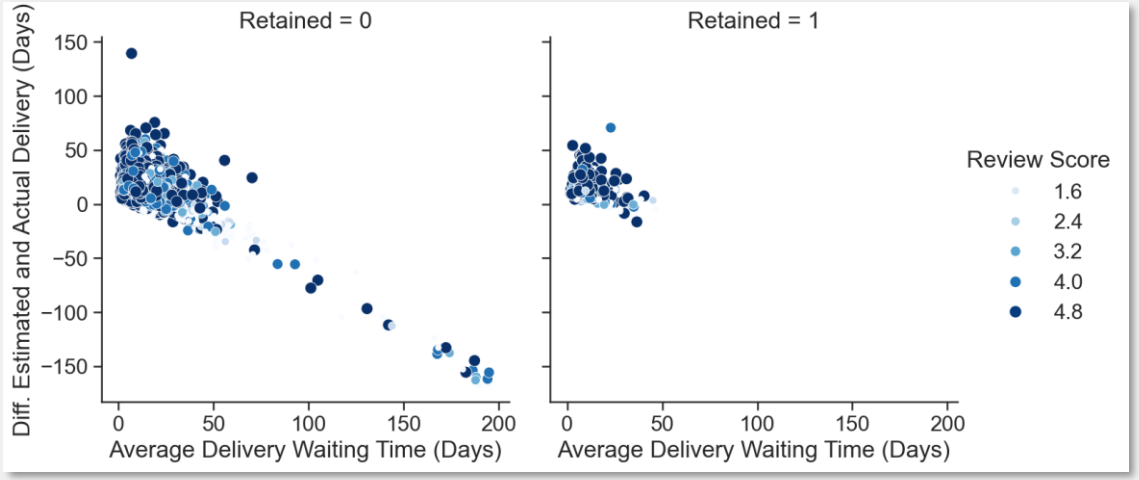
INSIGHTS PER RETAINED CUSTOMER

DELIVERY TIME

Retained customers have shorter delivery waiting days and a smaller discrepancy between estimated and real delivery days.

REVIEW SCORE

The average **review score of retained customers is statistically higher** than the scores of non-retained customers.

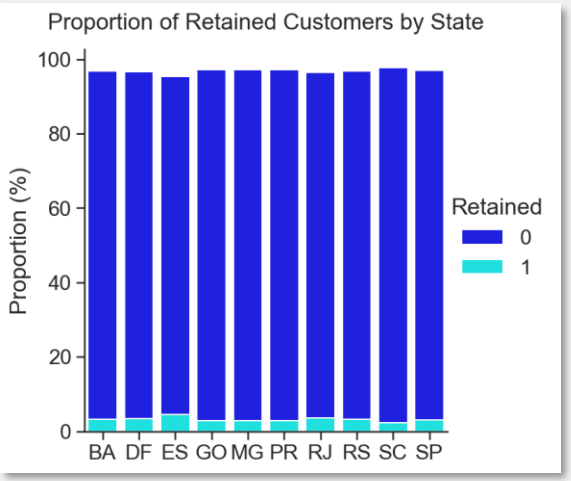


FREIGHT VALUE

Statistical analysis shows **no significant difference** in the average freight value for retained and non-retained customers.

CUSTOMER STATE

There is **no statistically significant association** between customer state and customer retention.

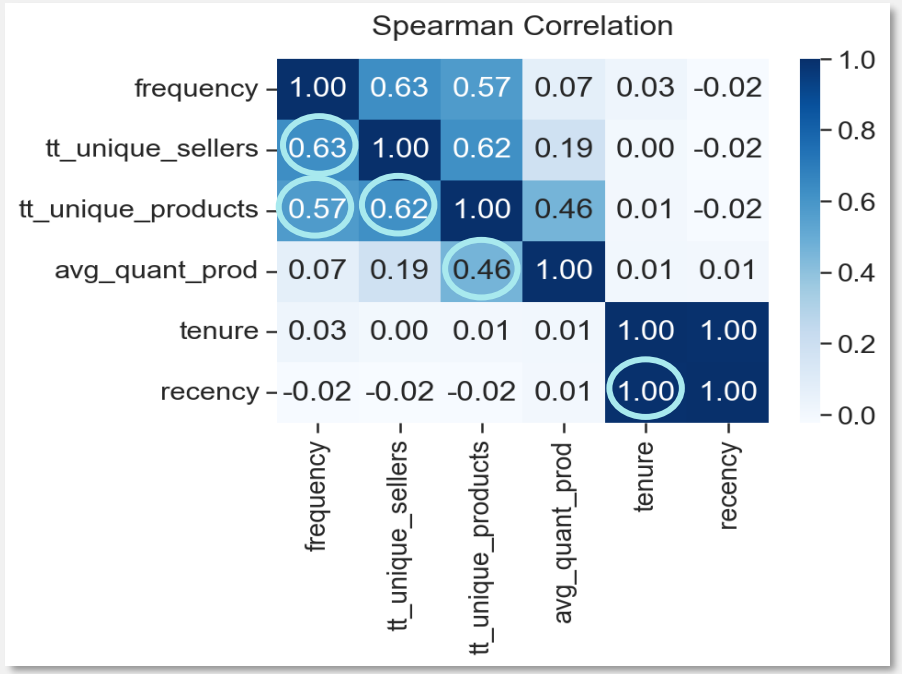


olist

FEATURES TABLE

Correlated Features

Highly correlated features were eliminated (confirmed by the Pearson/ Spearman correlation hypothesis tests).



Features Table

- Recency (Days since last purchase)
- Frequency (Number of purchases)
- Monetary Value (Total amount spent)
- Delivery Waiting Time (Average days)
- Average Quantity of Products per Order
- Average Number of Installments per Order
- Preferred Payment Type
- Average Customer Review Score
- Customer State (São Paulo, Distrito Federal)
- Clusters (after running K-Means Model)

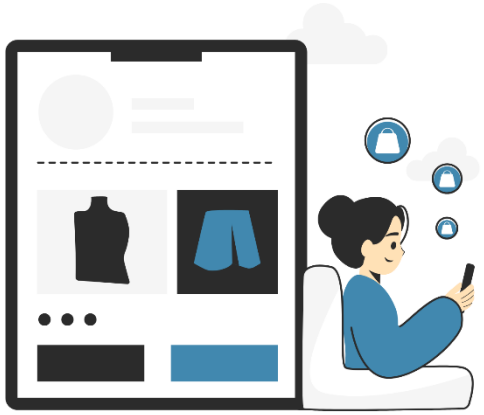
3%
Retained Customers

With only 3% of total customers making repeat purchases after 2017/05/31, there is a significantly imbalanced target class for predictive models.

olist

MODELING
K-MEANS

Customer Segmentation



K-MEANS

FEATURE SELECTION

- Various feature combinations across multiple iterations.
- Features related to customer characteristics, such as location and spending habits.

DATA PREPROCESSING

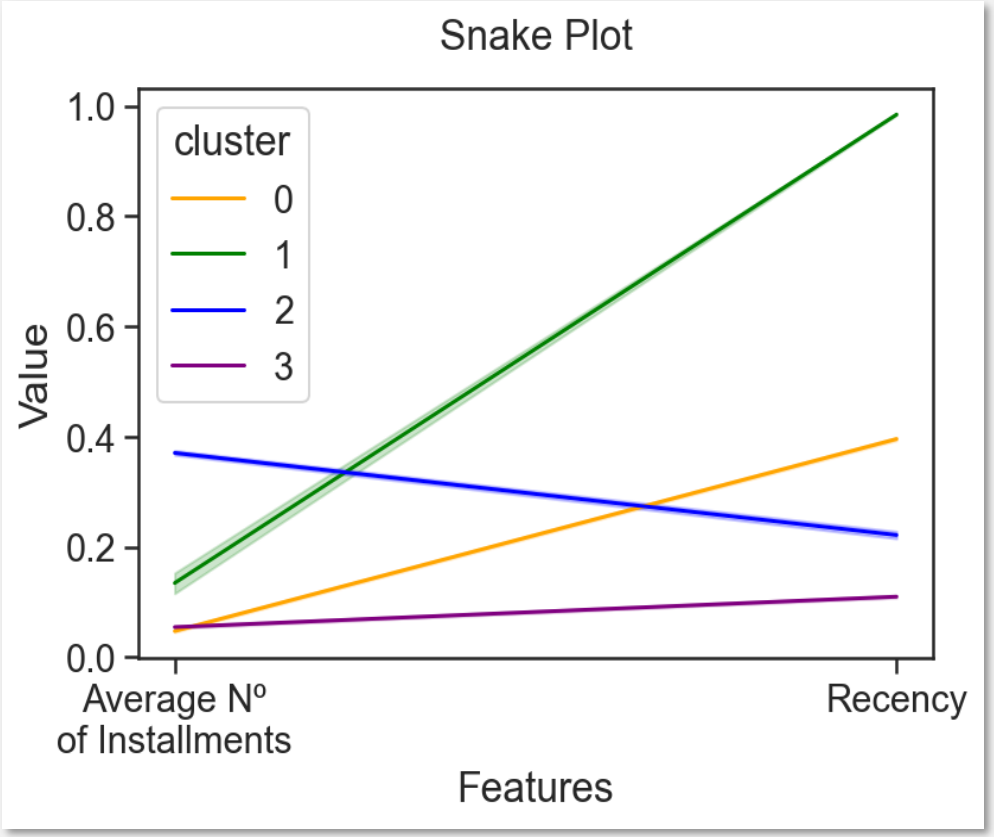
- One-hot encoding.
- Scaling numerical features to the 0-1 range.

CLUSTER ANALYSIS

- Elbow method and Silhouette score.

CHALLENGES

- High cardinality in categorical variables (K-Modes clustering for future modeling).
- Monetary value distribution: Highly skewed.



Cluster	0	1	2	3
Proportion	37.46%	2.41%	17.72%	42.41%

olist

CLUSTER ANALYSIS (CUSTOMER SEGMENTATION)

Reactivation-Ready Spenders

Cluster 0

Characteristics:

- Large segment
- Less recent purchases

Marketing Strategy:

- Analyse why they did not buy recently
- Re-Engagement campaigns

Long-Lapsed Spenders

Cluster 1

Characteristics:

- Smallest segment
- Long inactivity

Marketing Strategy:

- Evaluate if the effort to re-engage them is cost-effective

Premium Installment Spenders

Cluster 2

Characteristics:

- Confortable with spreading out payments
- Recent purchases

Marketing Strategy:

- Offer flexible payment options
- Premium products

New Value Spenders

Cluster 3

Characteristics:

- Largest segment
- Recent purchases

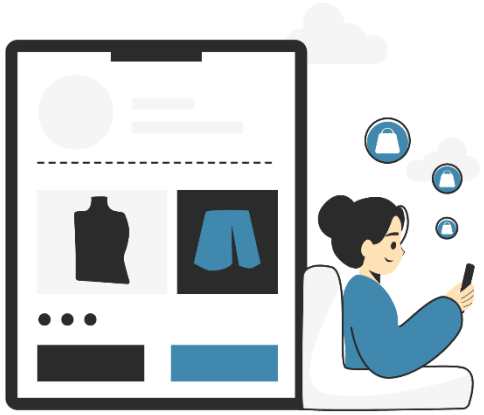
Marketing Strategy:

- Onboarding and Welcome campaigns

olist

MODELING RANDOM FOREST

Retention Prediction



RANDOM FOREST METHODOLOGY

FEATURE SELECTION

- Various feature combinations across multiple iterations.
- Features that could impact customer's likelihood to make a repeat purchase.

DATA PREPROCESSING

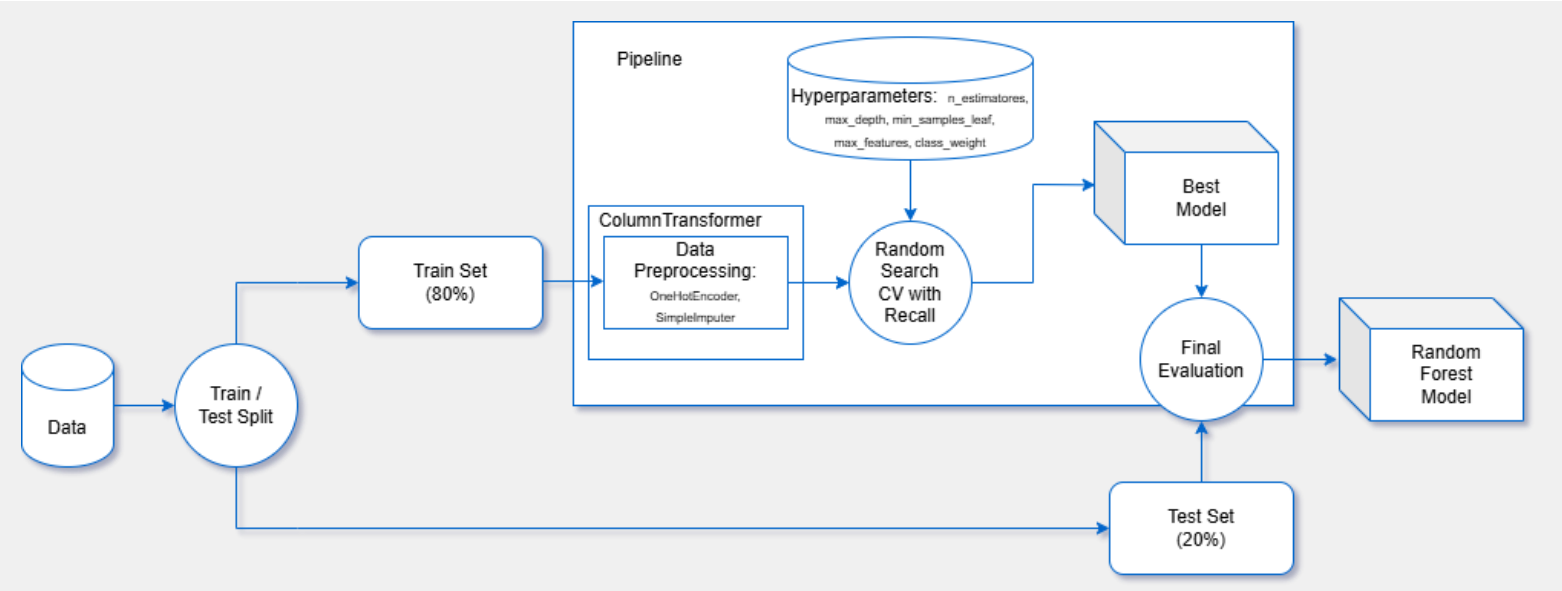
- One-hot encoding.
- Numerical imputation for average score feature.
- Pipeline to avoid data leakage.

MODEL TRAINING

- Train/ Test split
- RandomSearchCV (K-fold = 10) trained on the training dataset with Recall score and then evaluated on the independent test dataset.

CHALLENGES

- Imputation of missing values. Given the highly skewed distribution of review scores, the median was chosen as a more robust measure.
- Imbalanced target class.



RANDOM FOREST METRICS

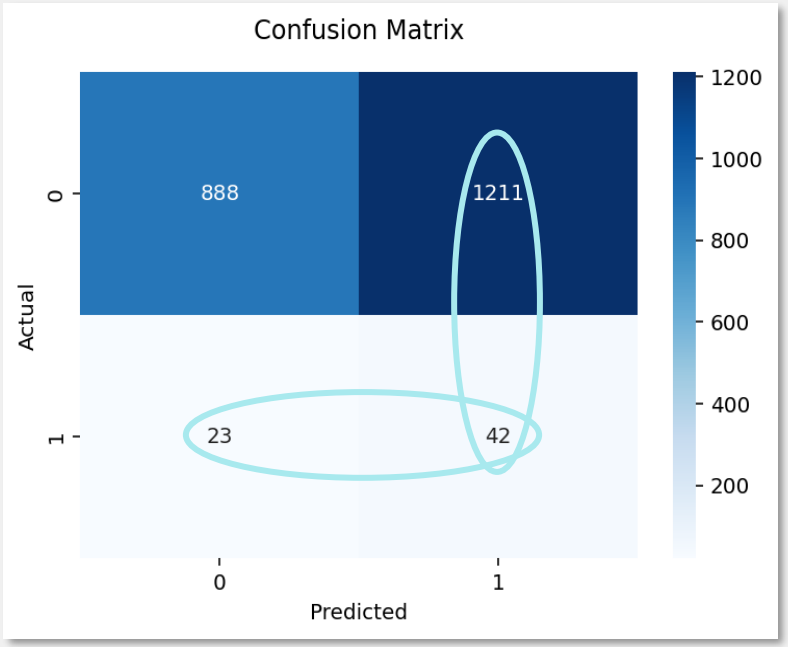
RECALL SCORE

Given imbalanced retention data, accuracy is misleading due to majority bias. The goal is to correctly **identify all retained customers** (positive class) for targeted marketing strategies and minimize false negatives. Failing to do so, means missing **opportunities for increased revenue** through these strategies. Therefore, recall is key.

0.69
Train Recall Score

0.65
Test Recall Score

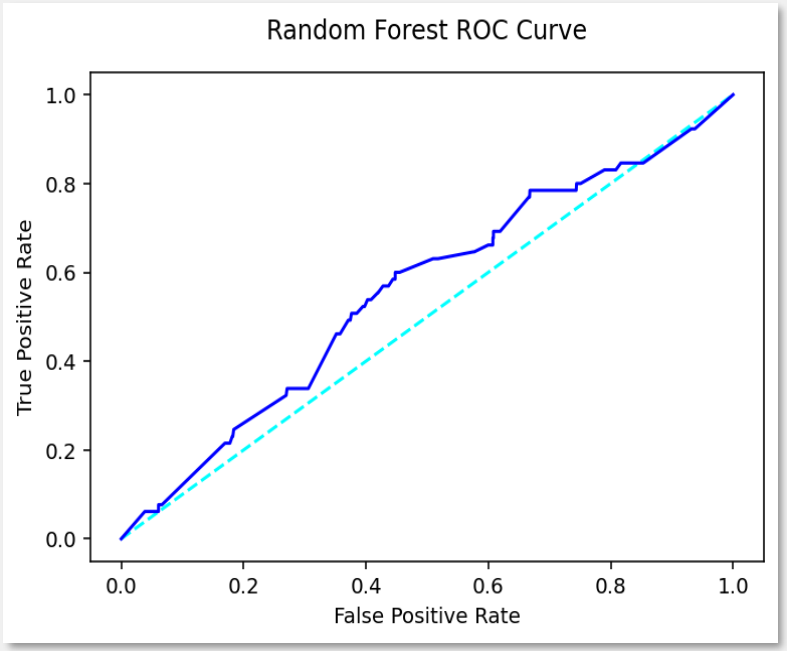
Precision Score **0.03**



The model shows low precision. Precision ensures efficient resource allocation, helps to accurately measure campaign effectiveness and understand true loyalty.

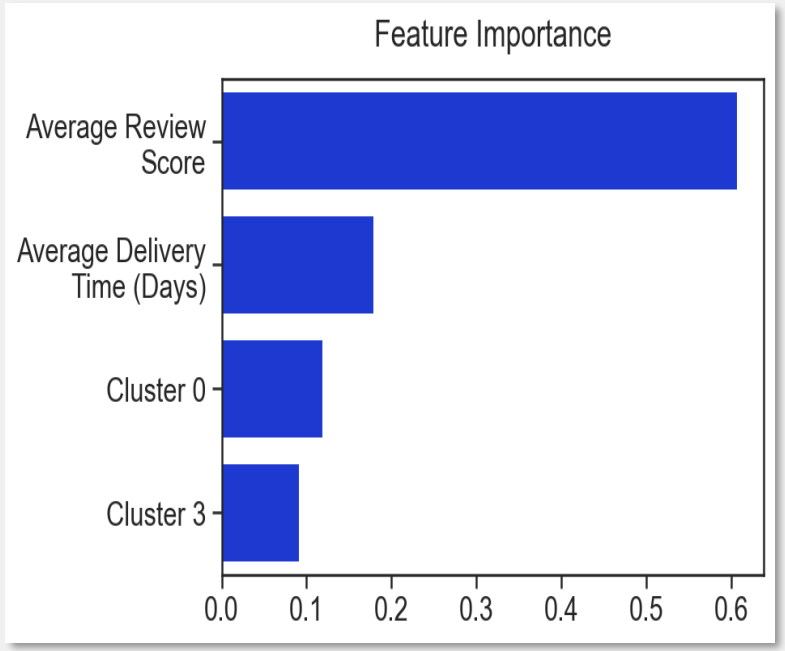
RANDOM FOREST METRICS

ROC / AUC Area 0.56



While the TPR increased, it led to a significant rise in the FPR. With a low AUC, the model is only slightly better than random guessing.

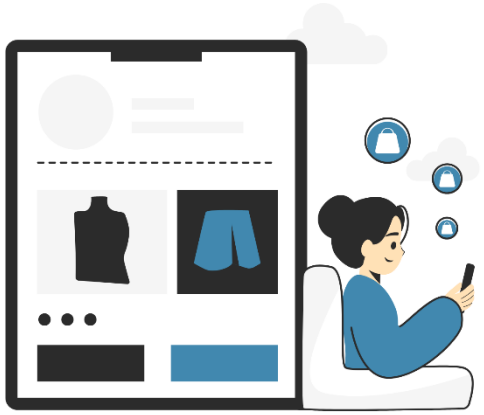
Nº of Relevant Features 4



Product and service quality (Review Score), logistical issues (Delivery Waiting Time) and Customer Segmentation (Cluster 0 and 3) are the most relevant features.

olist

CONCLUSIONS AND RECOMENDATIONS



CONCLUSIONS

CUSTOMER SEGMENTATION

The K-Means algorithm identified **four distinct customer segments.** One key differentiator was activity level, separating clients with recent purchases from those who have become inactive. Additionally, the algorithm distinguished groups based on payment preferences, highlighting customers who tend to pay upfront versus those who prefer to spread their payments.

IMBALANCED DATA

With a mere 3% of customers making repeat purchases after May 31, 2017, the target class is severely imbalanced, which challenges the performance of any predictive model. **While exploring alternative algorithms** beyond Random Forest might offer some improvement**, this class imbalance would remain a significant obstacle.**

RECALL AND PRECISION

While random forest model achieves a recall of 0.65, this comes at a cost of low precision (0.03). This **trade-off** results in a large number of false positives (1142 customers).

olist

BUSINESS RECOMENDATIONS

INCREASE CUSTOMER RETENTION RATE

MARKETING STRATEGIES

The **customer segments provide valuable insights** into different profiles, enabling targeted marketing strategies tailored to their specific behaviours and preferences.

FEATURE IMPORTANCE

Analysing customer comments to highlight the specific aspects driving negative **reviews** is crucial. Furthermore, optimizing the logistics process to ensure **timely delivery** is essential.

COST-BENEFIT ANALYSIS

It is crucial to conduct a cost-benefit analysis to determine if the **revenue gained** from correctly identifying retained customers outweighs the **expenses associated with incorrectly targeting the 1142 customers** with retention efforts.

REFERENCES

[1] https://www.business.com/articles/returning-customers-spend-67-more-than-new-customers-keep-your-customers-coming-back-with-a-recurring-revenue-sales-model/

[2] https://www.earnestanalytics.com/insights/temus-retention-grows-over-time-leads-walmart-trails-amazon

[3] https://platformpapers.com/trust-and-platform-disintermediation/

[4] https://www.mckinsey.com/industries/logistics/our-insights/what-do-us-consumers-want-from-e-commerce-deliveries

[5] https://salesduo.com/blog/customer-retention-on-amazon/

[6] https://portal-bucket.azureedge.net/wp-content/2025/05/Relatorio_Peic_abr25.pdf

[7] https://cidades.ibge.gov.br/brasil/sp/sao-paulo/panorama

[8] https://cidades.ibge.gov.br/brasil/df/panorama

olist

THANK YOU