ОПТ Транзакции 1

Arkhipov N.A.,
Alkhanashivli A.Z.,
Rukavishnikov N.A.,
Chechulin N.D.

Stages of project implementation

1 Data cleaning

Translation result
Data correction, correct merging into a pivot table, filling in gaps, removing outliers.

2 Showcase creation

Aggregation of data into showcases for subsequent application of models, analysis of the received showcases

3 MBA

Definition joint purchase rules

4 Response Model

Building models that predict the likelihood of a customer returning

5 Forming an offer

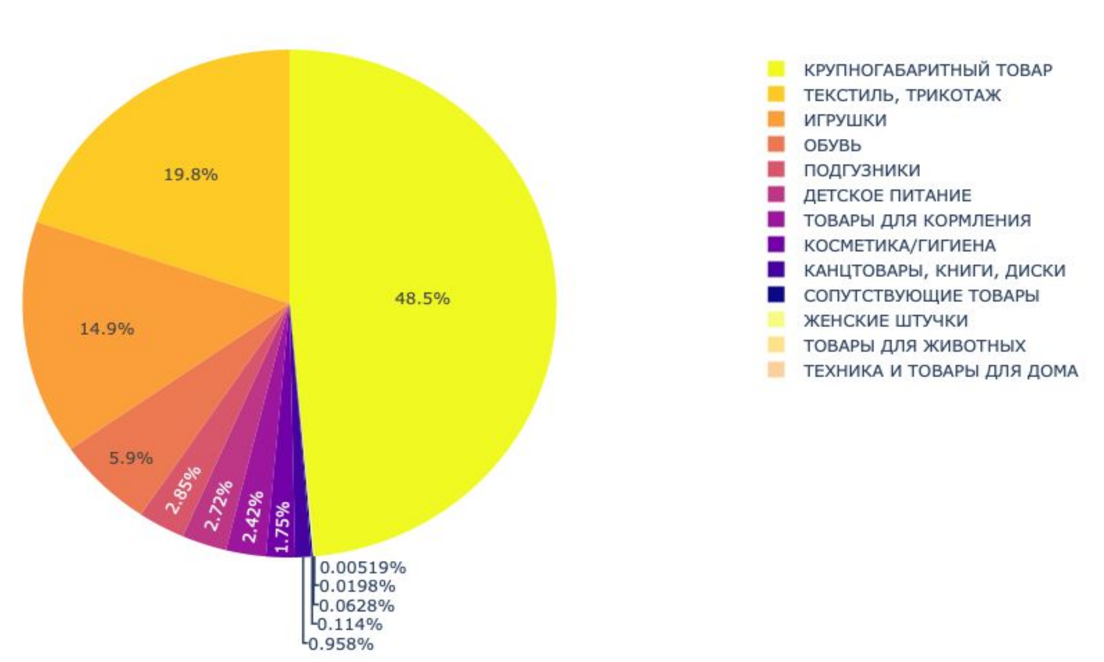
Formation of the best personal offers for each of the profiles

6 Financial justification

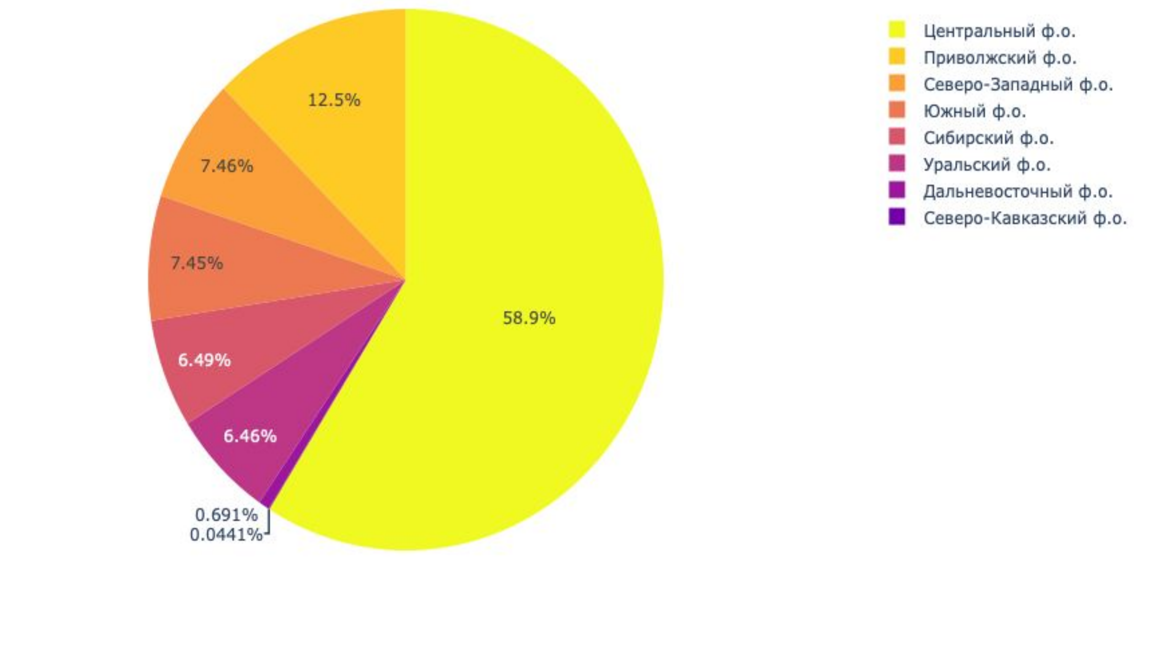
Financial justification

Data visualization

Margin by category

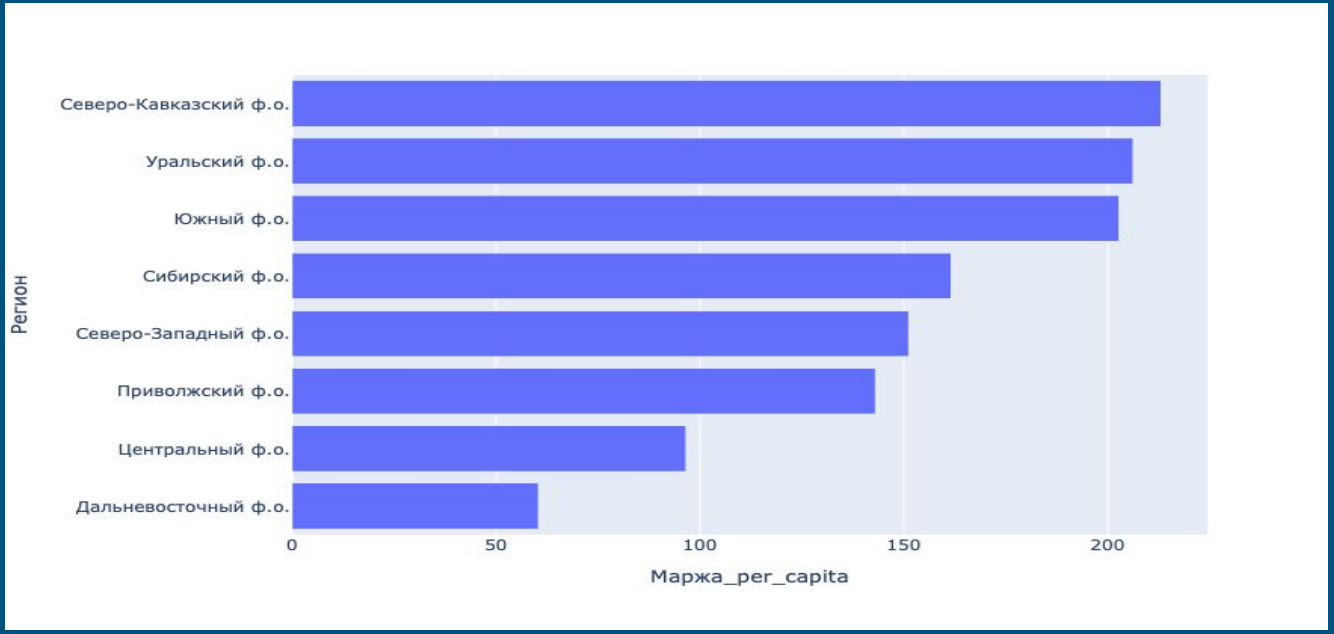


Margin by regions

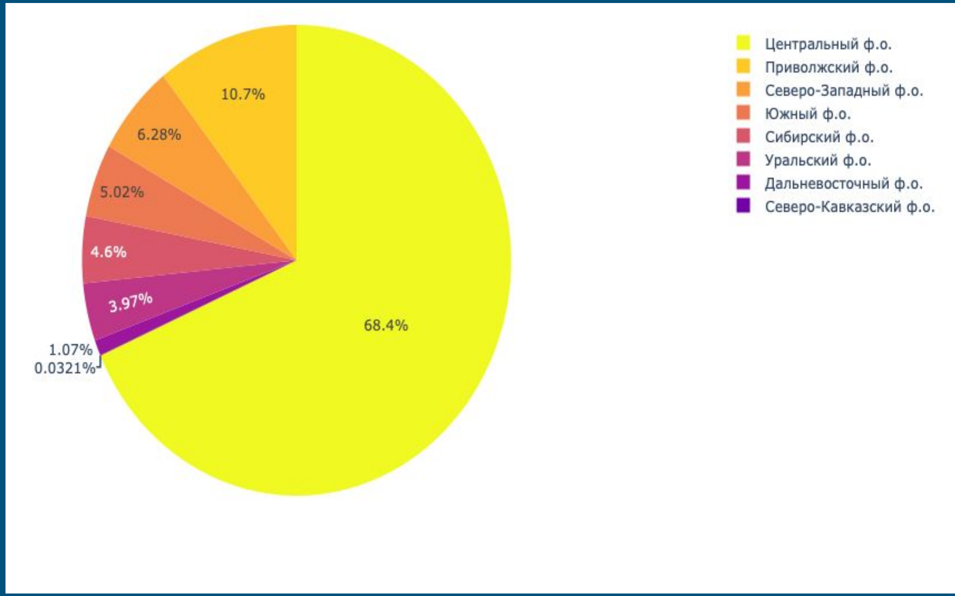


Data visualization

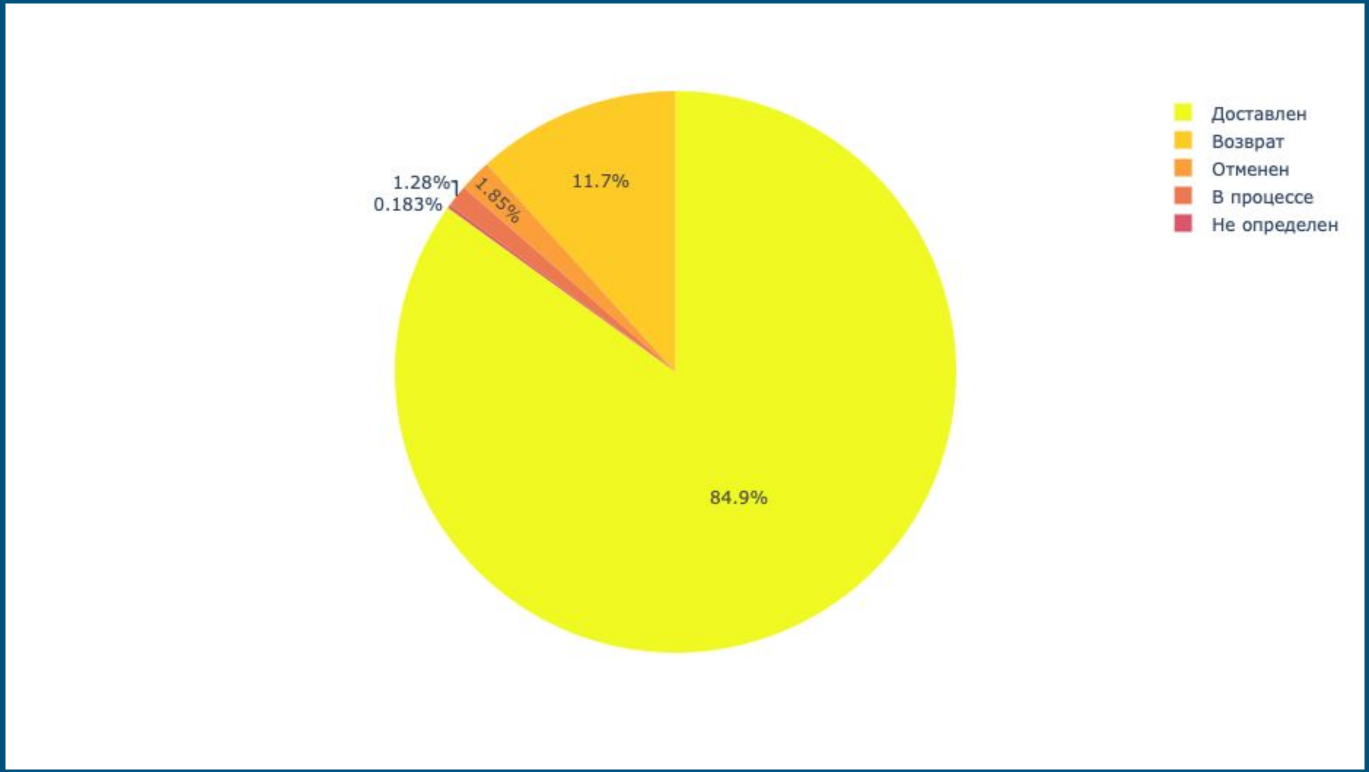
Marginality by regions



Number of checks by region



Data cleaning and transformation



Exploratory Data Analysis

I

Remove irrelevant columns

II

Convert columns to the correct data types

III

Remove incorrect values, clean data and validate calculations

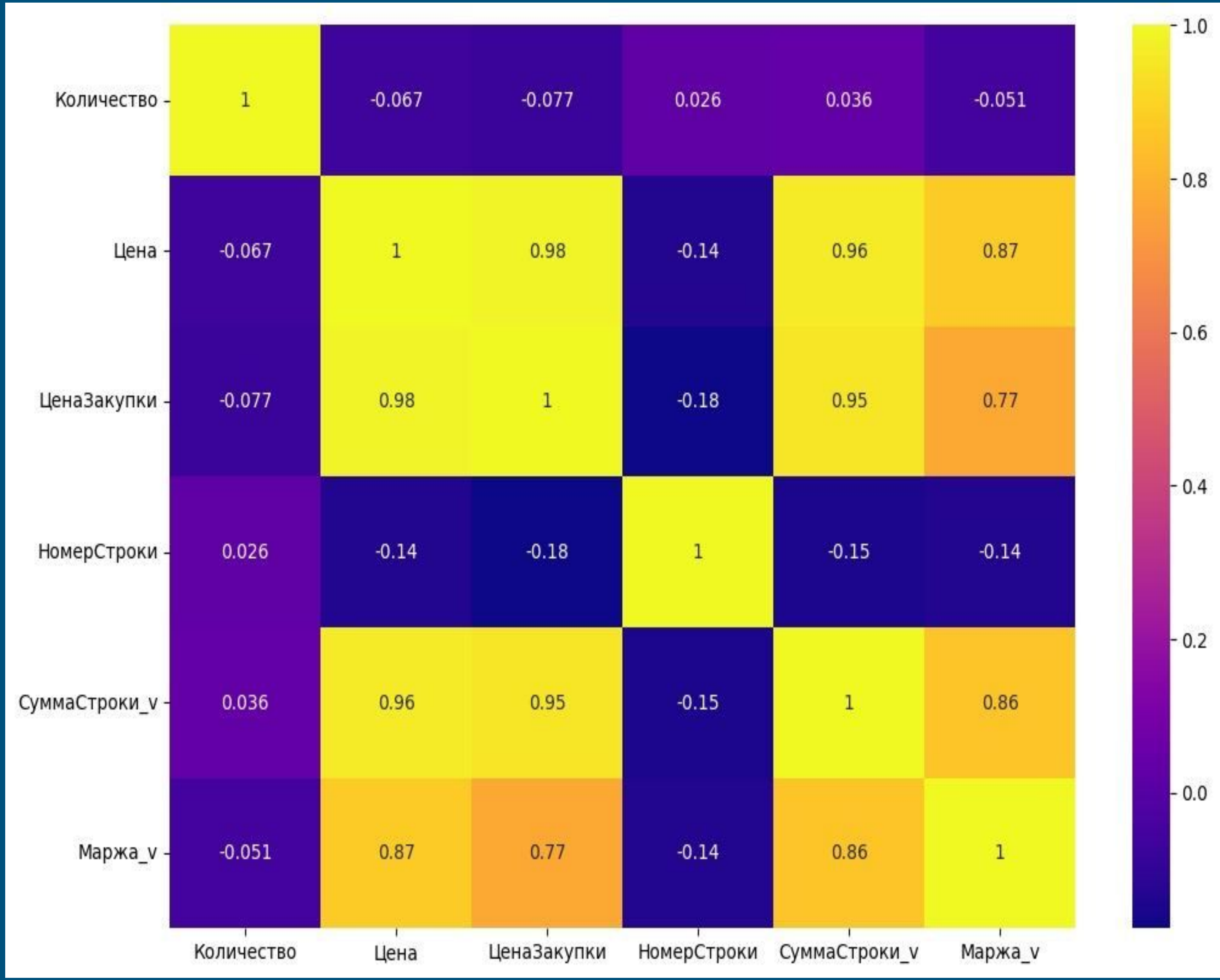
IV

Divide geo-data into federal districts

V

Omit rejected positions and prepare data for pivot tables and analytical base tables

Exploratory Data Analysis



Correlation between features

Realization stages

1. **Data cleaning**

   Data correction, correct merging into a pivot table, filling in gaps, removing outliers.

2. **Showcase**

   Aggregation of data into showcases for subsequent application of models, analysis of the received showcases

3. **MBA**

   Defining co-purchase rules

4. **Response model**

   Building models that predict the likelihood of a customer returning

5. **Formation of an offer**

   Formation of the best personal offers for each of the profiles

6. **Financial justification**

   Calculation of the financial and economic justification of the project

Market Basket Analysis

Discover which products are frequently purchased together by customers

Association Rule Mining

Find frequent itemsets and generate rules indicating item co-purchases

Rule Evaluation

Evaluate and measure the strength and significance of these rules

Recommendation Systems

Analyze purchase patterns, recommend related items and boost sales

Inventory Management

MBA optimizes inventory by identifying frequently co-purchased items

Apriori	FP-Growth	Association rules	FPMax	HMine
Generates association rules by scanning the database multiple times and incrementally discovering itemsets with increasing length.	Constructs a compact data structure called an FP-tree to represent the transactions and then performs recursive mining to find frequent itemsets.	Technique used that utilizes frequent itemsets obtained from algorithms like Apriori or FP-Growth to generate rules of the form "if-then."	An extension of the FP-Growth algorithm that focuses on finding maximal frequent itemsets, which are itemsets that do not have any proper supersets that are also frequent.	Optimized algorithm that improves upon Apriori and FP-Growth, by utilizing the H-struct data structure and a more efficient search space traversal method.
Easy to understand, widely used, and works well for small to medium-sized datasets.	Faster than Apriori as it avoids costly database scans. It can handle large datasets efficiently and has a reduced memory footprint.	Association Rules provide valuable insights into item associations and can guide business decision-making.	Efficiently identifies maximal frequent itemsets, reducing redundancy in the generated rules and improving the interpretability of the results.	The H-struct is a hybrid data structure that combines the benefits of both horizontal and vertical data layouts, making it more efficient for frequent itemset mining.
It can be computationally expensive for large datasets due to its multiple database scans, and it may generate a large number of candidate itemsets, leading to slower execution.	The initial construction of the FP-tree can be memory-intensive for very large datasets. Additionally, it may generate a large number of candidate itemsets for association rule generation.	Association Rules may generate a large number of rules, including many that are irrelevant or less actionable. Selecting meaningful rules and interpreting their significance can be challenging.	FPMax may still generate a large number of candidate itemsets for large datasets, which can impact the execution time and require additional filtering or post-processing steps.	Has minimal and predictable space overhead, operates quickly in memory-based settings, and can scale to large databases through partitioning. Additionally, it dynamically constructs (conditional) FP-trees during the mining process for dense datasets.
The Apriori algorithm uses the "Apriori principle" which states that if an itemset is infrequent, then all its supersets must also be infrequent.			Maximal frequent itemsets can be useful when a concise set of rules is desired, or when the focus is on the most important or unique associations.	

Provide personalized recommendations to users, improving user experience and driving additional profits

SVD

Decomposes the matrix into latent factors representing user preferences and item characteristics. It predicts user ratings for unknown items based on the similarity between user and item latent factors, enabling personalized recommendations.

SVD++

SVD++ incorporates implicit feedback signals from user interactions. It improves recommendation accuracy and relevance by capturing latent factors representing user preferences and item characteristics more accurately than traditional SVD. This approach is especially useful when explicit ratings are unavailable but user interactions, such as purchases, can be leveraged to infer user preferences.

NMF

Decomposes a non-negative matrix into two non-negative matrices representing user preferences and item characteristics. It uncovers latent patterns for accurate and relevant recommendations, especially with non-negative data such as counts or frequencies.

Name	RMSE	MAE	TIME
SVD++	0.004	0.001	0:00:24
SVD++ with cache	0.004	0.001	0:00:23
SVD	0.011	0.003	0:00:13
NMF	0.062	0.057	0:00:24

Comparison of recommender systems algorithms on the whole dataset

Realization stages

1 Data cleaning

Data correction, correct merging into a pivot table, filling in gaps, removing outliers.

2 Showcase

Aggregation of data into showcases for subsequent application of models, analysis of the received showcases

3 MBA

Defining co-purchase rules

4 Response model

Building models that predict the likelihood of a customer returning

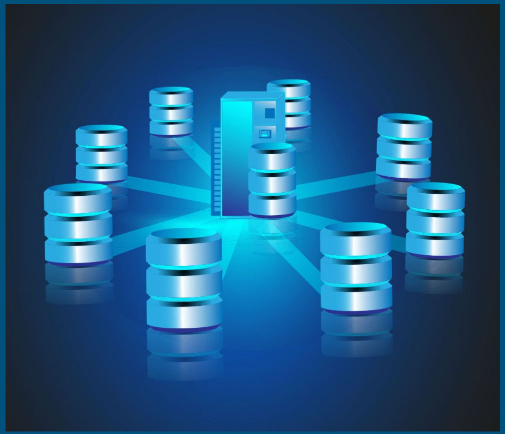
5 Formation of an offer

Formation of the best personal offers for each of the profiles

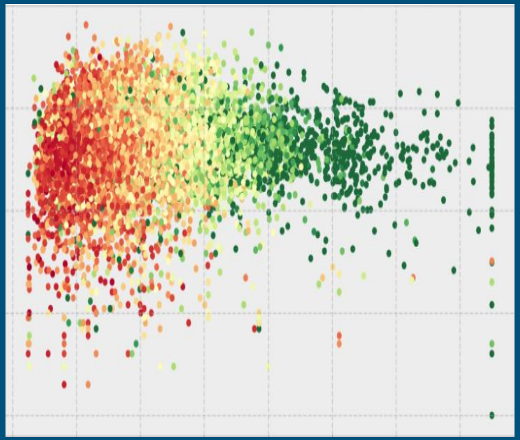
6 Financial justification

Calculation of the financial and economic justification of the project

Building a Response Model



Data preparation



Building Models
Classifications

Selection of hyperparameters

Choosing the best model

Classification models

RandomForest

Cat boost

XGboost

Gradient Boosting

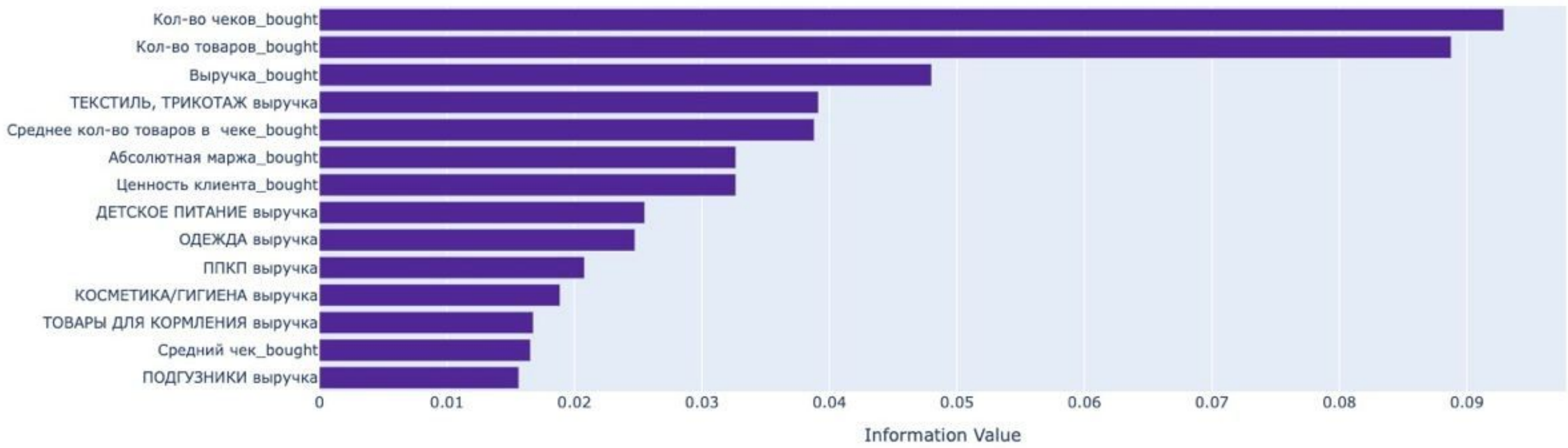
Выбранные метрики качества:

- ROC-AUC
- F1
- Gini

Model comparison

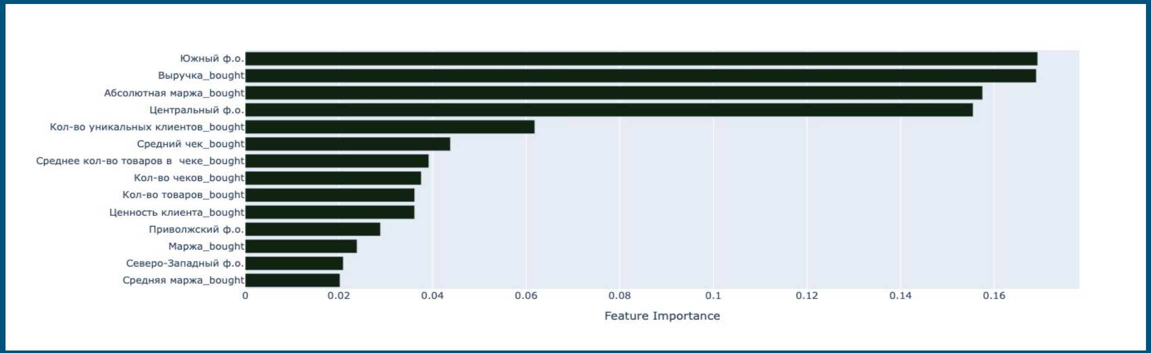
Метрики качества	**RandomForest**	**CatBoost**	**XGboost**	**Gradient boosting**
ROC AUC Train	0.96	0.67	0.96	0.96
ROC AUC Test	0.71	0.63	0.71	0.71
F1 Train	0.96	0.71	0.96	0.96
F1 Test	0.74	0.67	0.74	0.74
Gini Train	0.92	0.35	0.91	0.91
Gini Test	0.42	0.26	0.41	0.41

Sample of significant features

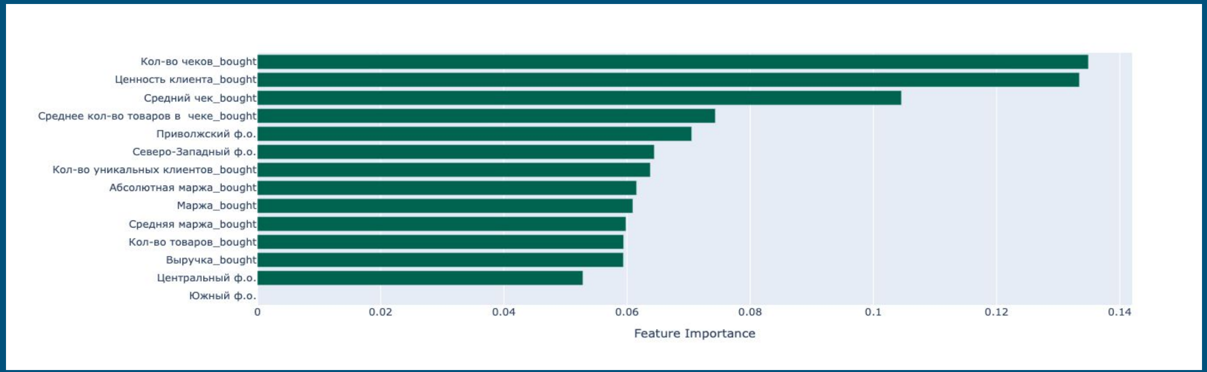


Information values for all models

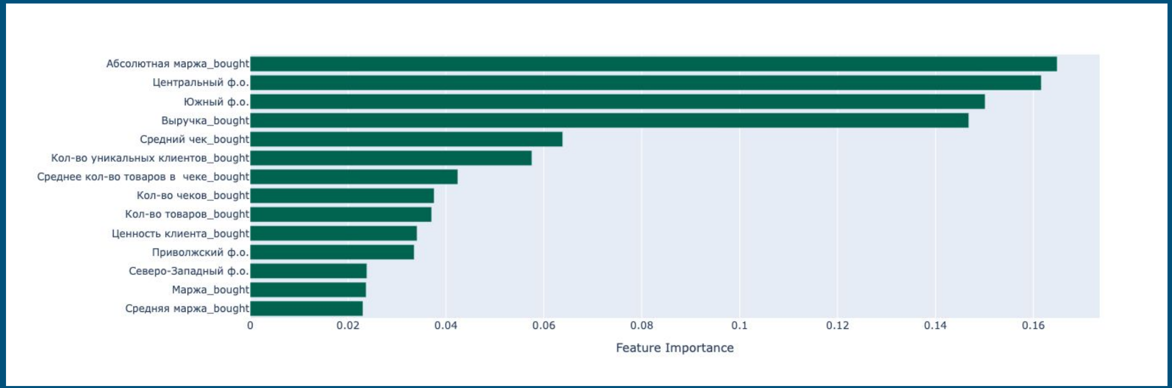
Random Forest



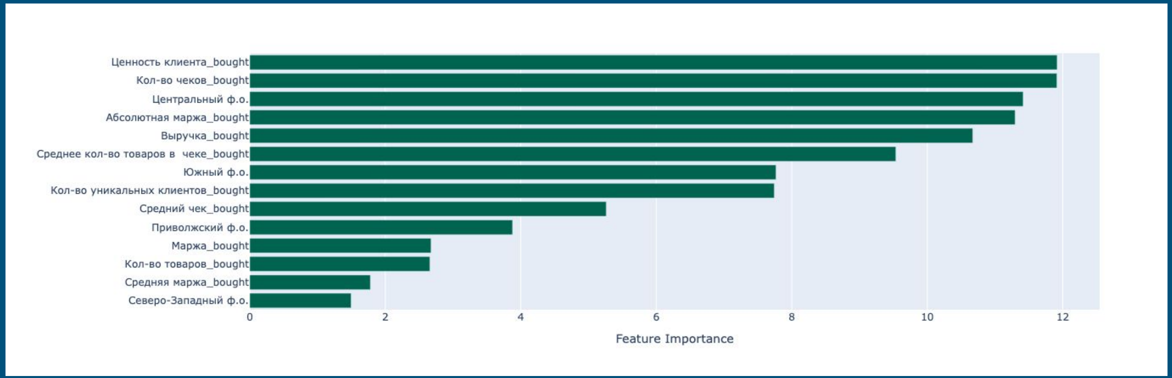
XGBoost



CatBoost



Gradient Boosting



Realization stages

1 Data cleaning

Data correction, correct merging into a pivot table, filling in gaps, removing outliers.

2 Showcase

Aggregation of data into showcases for subsequent application of models, analysis of the received showcases

3 MBA

Defining co-purchase rules

4 Response model

Building models that predict the likelihood of a customer returning

5 Formation of an offer

Formation of the best personal offers for each of the profiles

6 Financial justification

Calculation of the financial and economic justification of the project

Definition of rules for targeted campaigns and selection of customers

Make a purchase of goods of a certain line from a certain amount and get a 10% cashback

I goal

Increase in the average check
Increased shopping frequency
Cart expansion
Cart Margin Increase

II goal

Unstable category
more expensive brand
More marginal brand
Minimum basket

III goal

Loyalty points increase
Fixed discount
Percentage discount
Second item as a gift

Next best offer

Segment	Aim	Customer processing method	Expected profit (thousand rubles)
I segment baby food	Increasing the average check	Distribution of special offers among clients	4000
II segment baby clothes	Increasing the average check, expanding the audience	Adding coupons for next purchases when paying online with delivery	3085

Next best offer

Segment	Aim	Customer processing method	Expected profit (thousand rubles)
III segment Diapers	Increasing the number of goods in the receipt, unloading warehouses	When buying three packs of diapers, the fourth one is a gift	4175
IV segment Clothes	Increasing the average check	When buying goods from the category "Clothing" in the amount of 3,500 rubles or more, charge 10% of the amount in the form of bonuses	6090
V segment Toys	Expanding the variability of goods in the receipt	When buying several different products from the group "Toys" in the amount of 1500 rubles or more, a 7% discount	2580

Realization stages

1 Data cleaning

Data correction, correct merging into a pivot table, filling in gaps, removing outliers.

2 Showcase

Aggregation of data into showcases for subsequent application of models, analysis of the received showcases

3 MBA

Defining co-purchase rules

4 Response model

Building models that predict the likelihood of a customer returning

5 Formation of an offer

Formation of the best personal offers for each of the profiles

6 Financial justification

Calculation of the financial and economic justification of the project

Result

Aim: **Automating the process of selecting customer lists for communication using mathematical modelling**

Done:

1. EDA. Data preparation for further analysis
2. Pivot table analysis is done
3. Analytical base table construction
4. Conducting customer segmentation and results analysis
5. Customer response models construction and analysis
6. Conducting Market Basket Analysis
7. Choosing the optimal offer for each of the customer segments