Predicting Advertisement Demand on Avito Platform

Artur Zagitov, Ildar Zalialiev, Artem Nazarov

June 25, 2024

Contents

1 Introduction

In the digital age, the online marketplace has become an essential platform for buying and selling goods. However, sellers often face challenges in predicting the demand for their products, which can lead to frustration and suboptimal pricing strategies. Avito, Russia's largest classified advertisements website, encounters these issues regularly. Sellers are frequently frustrated by the lack of demand for their listings or overwhelmed by unexpected high demand, indicating that the product description or price might not be optimal.

This report presents the first phase of the development of a machine learning system to predict demand for online advertisements. By using comprehensive data, provided by the platform itself, including ad descriptions, geographical context and other relevant information, we aim to provide sellers with insights to optimize their listings and better understand market interest.

2 Business and Data Understanding

2.1 Business Problem

The primary business problem is to accurately predict the demand for online advertisements on Avito. This involves understanding the details of product descriptions and other contextual factors that influence buyer interest. The goal is to provide sellers with insights to enhance their listings and manage expectations regarding demand.

Addressing this problem is crucial for several reasons:

- **Optimizing Listings:** By predicting demand more accurately, sellers can adjust their product descriptions, category and prices to better match market interest.
- **Reducing Frustration:** Sellers often face frustration when their products do not sell or sell too quickly, indicating potential underpricing. Accurate demand predictions can help mitigate these issues by providing realistic expectations.
- **Increasing Revenue:** For Avito, helping sellers optimize their listings can lead to higher transaction volumes, which will translate to increased revenue through commissions and advertising fees.
- **Enhancing User Experience:** Buyers benefit from more relevant and well-described listings, improving their overall shopping experience on the platform and increasing their retention.
- **Competitive Advantage:** In a competitive market, offering advanced tools and insights for sellers can set Avito apart from other classified advertisement websites.

2.2 Assessment of the Current Situation

Avito's platform hosts millions of advertisements across various categories and regions. Despite the platform's extensive reach, many sellers struggle with optimizing their ad listings due to a lack of understanding of what drives demand. This results in either overpricing or underpricing of products, leading to missed opportunities or dissatisfaction.

In data, provided by Avito, Avito computes deal probability based on a combination of how many people view the ad and how many people click the button. As illustrated in Figure 6, the deal probability for the majority of ads is very low, with approximately 66.4% of ads having a deal probability of less than 0.05. This indicates a significant challenge in generating interest and converting views into sales for a large portion of the listings on Avito. The histogram shows a long tail distribution with a few ads achieving high deal probabilities, suggesting that while some ads are highly effective, most struggle to attract buyers.

2.3 Terminology

2.3.1 Business Terminology

- **Ad Listing:** A product advertisement posted by a seller on Avito.

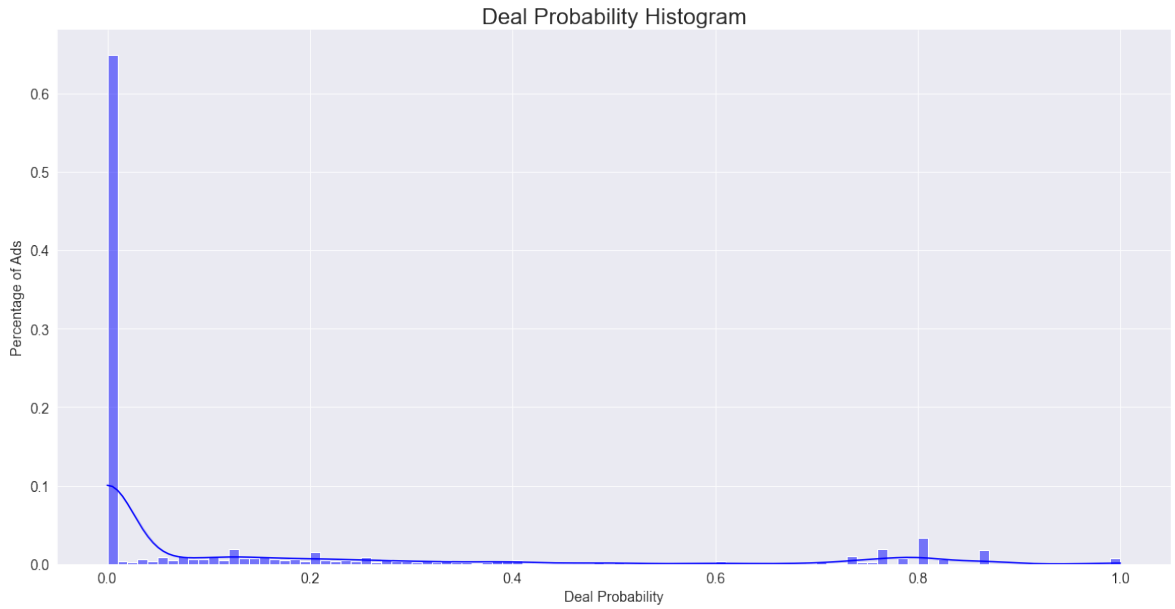

Figure 1: Deal Probability Histogram

- **Demand:** The interest shown by potential buyers, measured by views and interactions with the ad.
- **Deal Probability:** The likelihood that an advertisement will result in a transaction.
- **Optimization:** The process of adjusting ad features to increase demand.

2.3.2 ML Terminology

- **Root Mean Squared Error (RMSE):** A measure of the differences between predicted and observed values. It is used to evaluate the accuracy of the machine learning model.
- **Feature:** An individual measurable property or characteristic of a phenomenon being observed.
- **Categorical Data:** Variables that contain label values rather than numeric values.
- **Text Data:** Unstructured data that contains words and sentences.
- **Regression:** A type of predictive modeling technique which estimates the relationships among variables.
- **Feature Engineering:** The process of using domain knowledge to create features that make machine learning algorithms work better.
- **Overfitting:** A modeling error which occurs when a function is too closely aligned to a limited set of data points.
- **Cross-Validation:** A technique for evaluating ML models by training several models on subsets of the available input data and evaluating them on the complementary subset of the data.

2.4 Scope of the ML Project

2.4.1 Background

Avito is the largest classified advertisements website in Russia, providing a platform for users to buy and sell a wide range of products and services. Founded in 2007, Avito has grown rapidly and now has millions of active listings and a large user base across the country. The platform handles vast volumes of data daily, including user interactions, ad postings, and transaction records, click streams,

etc. Avito operates in multiple regions, to a diverse audience with varying needs and preferences and allows to list advertisements for many different types of products and services.

Being Russia's largest e-commerce platform, Avito faces the challenge of maintaining high-quality listings and user satisfaction. The large volume of data processed by the platform necessitates robust machine learning and MLOps solutions to handle data ingestion, processing, model training, and deployment efficiently. Implementing a machine learning solution to predict demand for advertisements requires a scalable infrastructure that can manage large datasets and provide real-time insights to sellers.

2.4.2 Business Problem

To provide sellers with accurate predictions of demand for their advertisements based on ad descriptions, contextual information, and historical data.

2.4.3 Business Objectives

- Increase the efficiency of ad listings by providing sellers with predictive insights.
- Reduce the number of underpriced and overpriced ads by optimizing pricing strategies.
- Enhance user satisfaction.

2.4.4 ML Objectives

- Develop a machine learning model to predict deal probability based on ad features and historical data.
- Achieve an RMSE of 0.22 or lower to ensure the model's predictions are accurate and reliable.

2.5 Success Criteria

2.5.1 Business Success Criteria

The success of the ML application will be measured by its ability to increase the number of successfully completed transactions on Avito by providing more accurate demand predictions.

2.5.2 ML Success Criteria

The machine learning model must achieve an RMSE of 0.22 or lower on the validation dataset, indicating that the predictions are sufficiently accurate for practical use.

2.5.3 Economic Success Criteria

A key performance indicator (KPI) will be the increase in revenue generated from ad listings due to improved pricing strategies and higher demand.

2.6 Data Collection

2.6.1 Data Collection Report

The data used in this project is sourced from Avito's classified advertisements platform, which is available on Kaggle. The dataset initially consisted of 1.5 million rows, but was randomly undersampled to 100,000 rows. Dataset containts 17 columns, including features such as ad titles, descriptions, prices, and geographical information. The data is collected in CSV format and includes a mix of date, categorical, numerical, and text data types.

To assist with the data collection process, we use the Kaggle API. This allows us to programmatically download the dataset directly from Kaggle.

Once the data is downloaded, we manage its versioning using Data Version Control (DVC). DVC is an open-source tool that enables versioning of data and machine learning models. It allows us to track changes to the dataset over time, ensuring that we can reproduce experiments and maintain a history of the data.

2.6.2 Data Version Control Report

Data versions are managed using Data Version Control (DVC), which integrates with our Git version control system. DVC allows us to version control large datasets and machine learning models. Currently, there are 2 versions of the data: raw data with all 1.5 million rows, and undersampled data with 100,000 rows.

2.7 Data Quality Verification

2.7.1 Data Description

The dataset includes the following 17 columns and contains 100,000 rows. Table 1 presents the information about each feature.

Column Name	**Description**	**Data Type**	**Feature Type**
item_id	Ad id.	object	Categorical
user_id	User id.	object	Categorical
region	Ad region.	object	Categorical
city	Ad city.	object	Categorical
parent_category_name	Top level ad category as classified by Avito's ad model.	object	Categorical
category_name	Fine grain ad category as classified by Avito's ad model.	object	Categorical
param_1	Optional parameter from Avito's ad model.	object	Categorical
param_2	Optional parameter from Avito's ad model.	object	Categorical
param_3	Optional parameter from Avito's ad model.	object	Categorical
title	Ad title.	object	Text
description	Ad description.	object	Text
price	Ad price.	float64	Numerical
item_seq_number	Ad sequential number for user.	int64	Numerical
activation_date	Date ad was placed.	date	Datetime
user_type	User type.	object	Categorical
image	Id code of image.	object	Categorical
image_top_1	Avito's classification code for the image.	int	Categorical
deal_probability	The target variable. This is the likelihood that an ad actually sold something. It can be any float from zero to one.	float64	Numerical

Table 1: Dataset Description

2.8 Data Exploration

In the data exploration stage, we aim to understand the underlying structure of the dataset, identify initial trends and patterns, and formulate hypotheses that can guide further analysis and model development. Here we present the results of our exploration, highlighting the most insightful findings.

2.8.1 Exploratory Data Analysis

Parent Category Name and Category Name

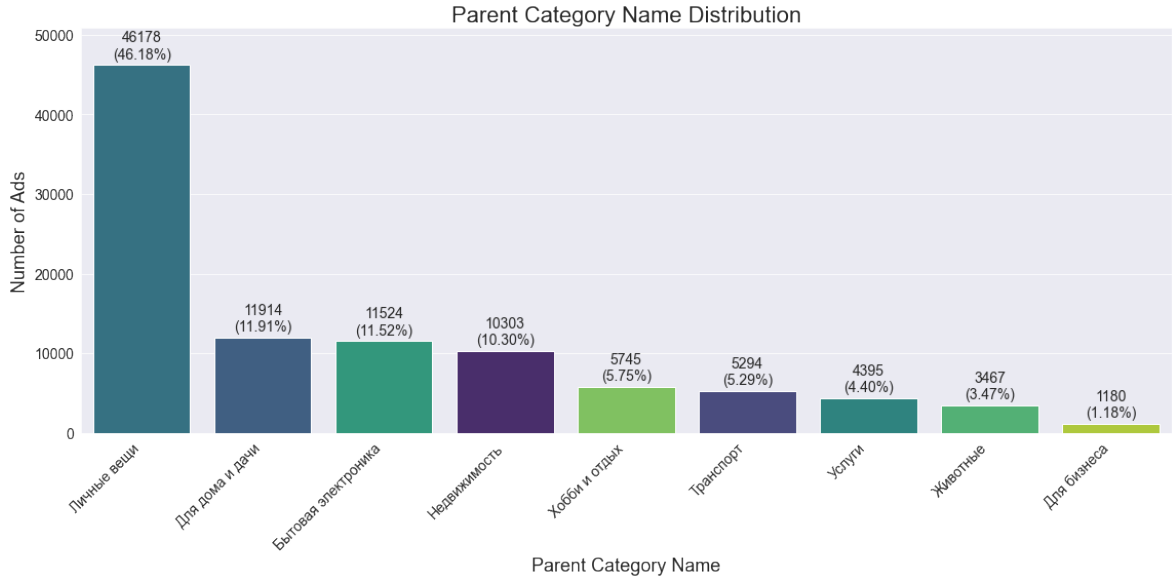

Figure 2: Parent Category Name Distribution

The distribution of parent category names shows that the most common parent category is "Personal belongings," followed by "For the home and garden" and "Consumer electronics." The distribution of ads across parent categories is not uniform, with some categories having significantly more ads than others, especially "Personal belongings", which contains almost half of the ads in the dataset. This indicates a heavy skew towards personal items in the dataset, which may impact the demand predictions.

There are also 47 unique categories in the dataset. The most common category is "Clothing, shoes, and accessories," followed by "Children's clothing and shoes". The distribution of ads across categories is not uniform. The first two categories, related to clothing, have significantly more ads than others, while 36 categories have less than 1% presence in the dataset.

User Type

The majority of ads are posted by private users, followed by companies and shops. Private users dominate the dataset, which might suggest different behaviors and demand patterns compared to commercial users.

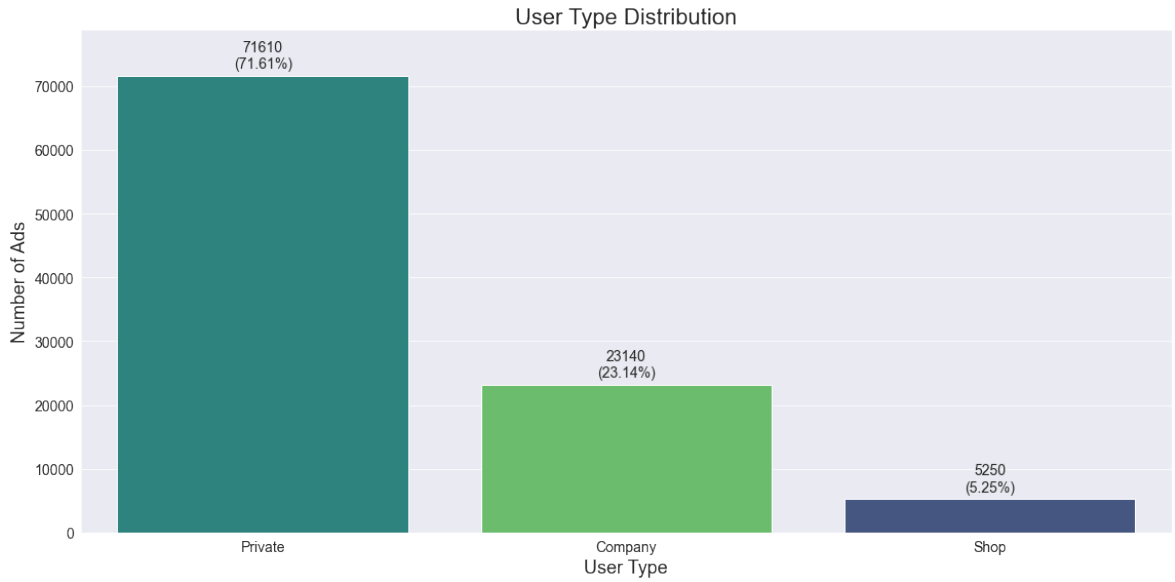

Figure 3: User Type Distribution

Region

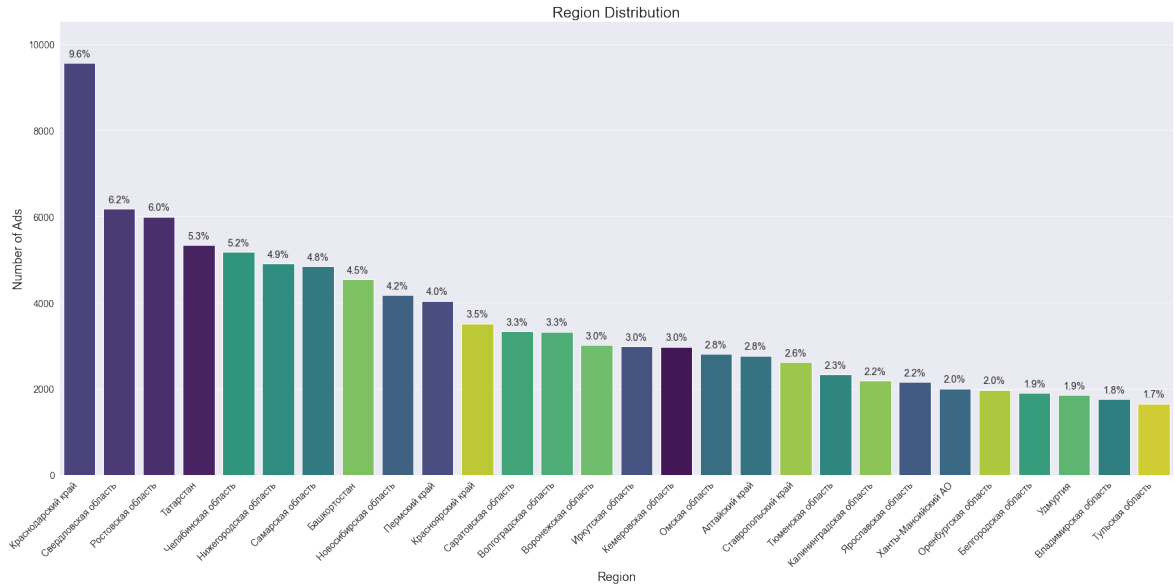

Figure 4: Region Distribution

The most common region is "Krasnodar region", followed by "Sverdlovsk region" and "Rostov region". The distribution of ads across regions is not uniform, with some regions having significantly more ads than others, especially "Krasnodar region", however, the difference is not very significant.

Price

Figure 5: Log_{10} of Price Histogram

The price distribution is right-skewed with a peak around 100 to 10,000 rubles. There are significant outliers with very high prices, which may need to be handled separately in the modeling phase to prevent skewing the predictions.

Deal Probability

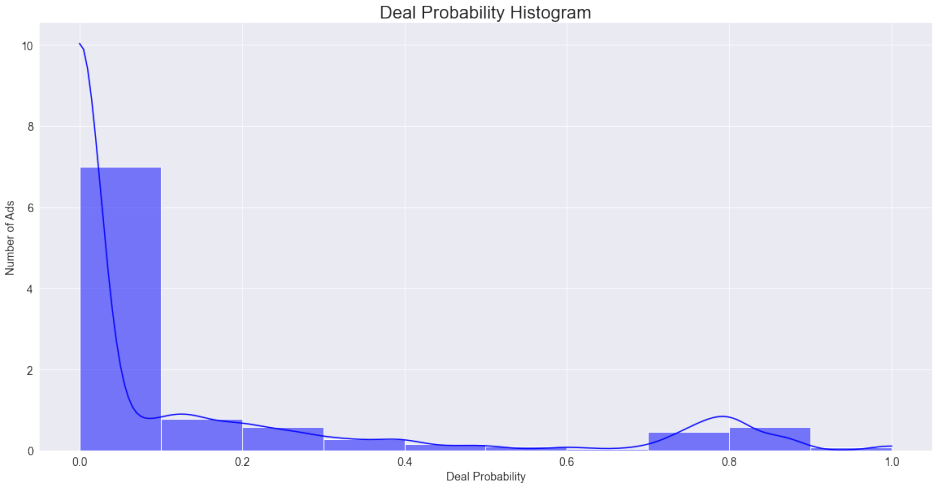

Figure 6: Deal Probability Histogram

The deal probability distribution is heavily right-skewed, with approximately 66.4% of ads having a deal probability of less than 0.05. This indicates that most ads struggle to convert views into transactions. There is also a small peak around 0.8, indicating a smaller number of ads with higher deal probabilities.

If we consider deal probability as a binary feature, where 1 means that the ad has a deal probability greater than 0.5, and 0 otherwise, the difference becomes clearer.

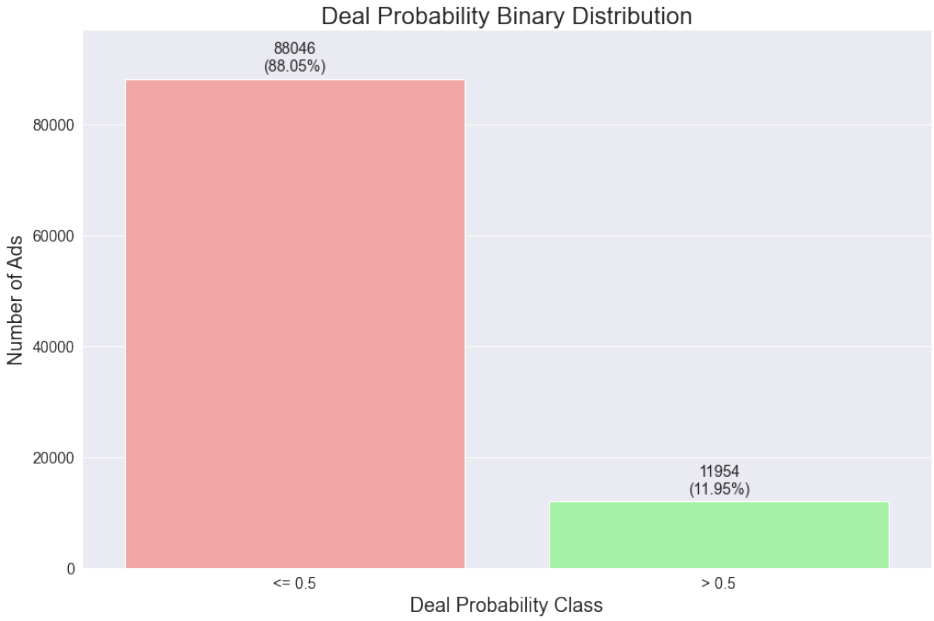

Figure 7: Deal Class Distribution

Title length

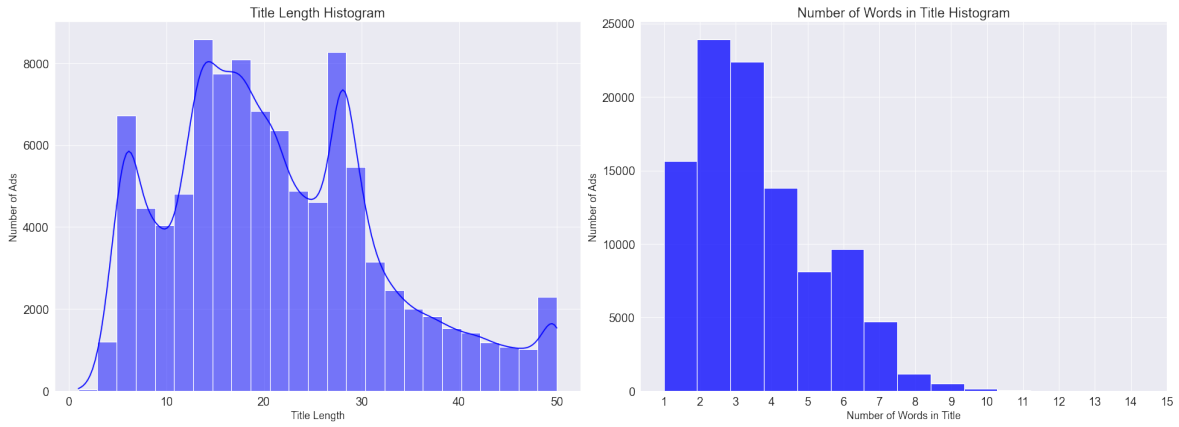

Figure 8: Title length Distribution

The title length histogram shows that the titles have a length between 0 and 50 characters, and 50 seems to be the character limit for titles. Majority of titles have a length between 0 and 30 characters and after that, the number of ads decreases. The number of words in the title histogram shows that the titles have between 1 and 15 words, with most titles having 1 to 10 words, and the majority having 1 to 5 words.

Price vs. Deal Probability

Figure 9: Price vs Deal Probability

- There is a high density of ads with lower prices (left side of the plot).
- Many of these low-priced ads have varying deal probabilities
- A few ads with low prices have a deal probability close to 1
- There is a line of ads with around 0.78 deal probability that have prices between 0 and 3,000,000, which could indicate a specific category of ads or a specific users group.
- As the price increases, the density of ads decreases, and the deal probability tends to be lower.
- There are a few outliers where some high-priced ads have achieved moderate deal probabilities.

User Type and Parent Category vs. Deal Probability

- The box plot in Figure 10 shows the distribution of deal probabilities for each parent category and user type.
- Deal probabilities for services are the highest, followed by transport and animals.
- Shop users are present only in 4 parent categories: 'Real estate', 'Transport', 'Animals', and 'Services'.
- Shops have the lowest median deal probability, while private users have the highest median deal probability.

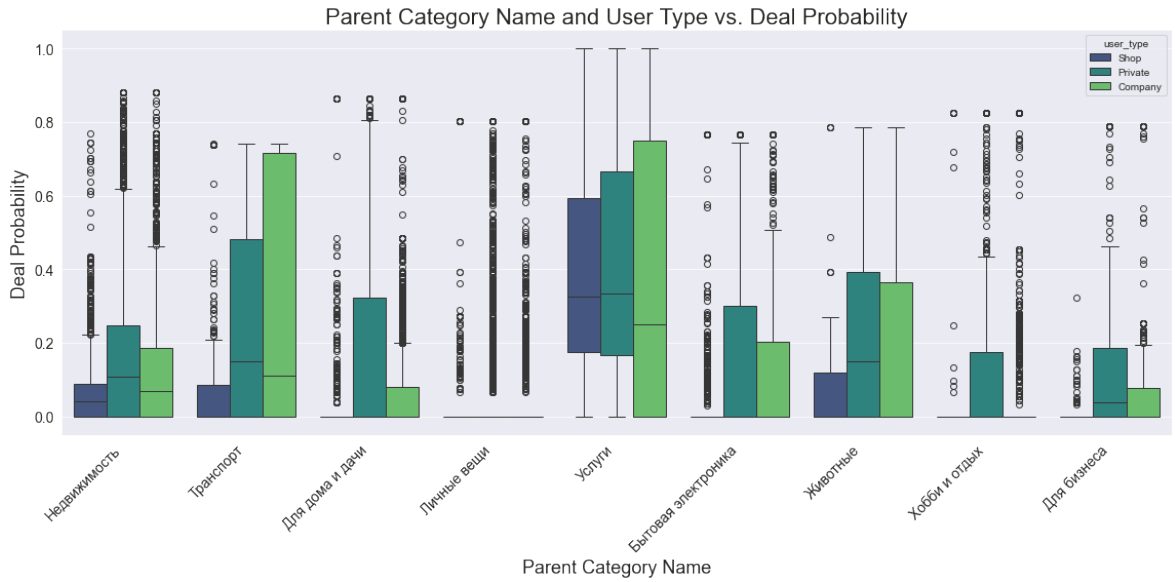

Figure 10: User Type and Parent Category vs. Deal Probability

2.8.2 Correlation Analysis

- The correlation matrix in Figure 11 shows the correlation coefficients between numerical features in the dataset.
- There is a correlation between the length of the title and the length of the description, as well as between the number of words in the title and the number of words in the description.
- Surprisingly, there is a correlation between image_top_1 and deal_probability, which could indicate that the image classification code may have some influence on the deal probability and may have some order in it, but it was supposed to be a categorical nominal feature. We may try to use it as a numerical feature in the model instead of encoding it.
- There is also a large negative correlation between image_top_1 and params_length, which could indicate that parameters are somehow related to the image classification code.
- Other then that, there are no significant correlations between deal probability and other numerical features.

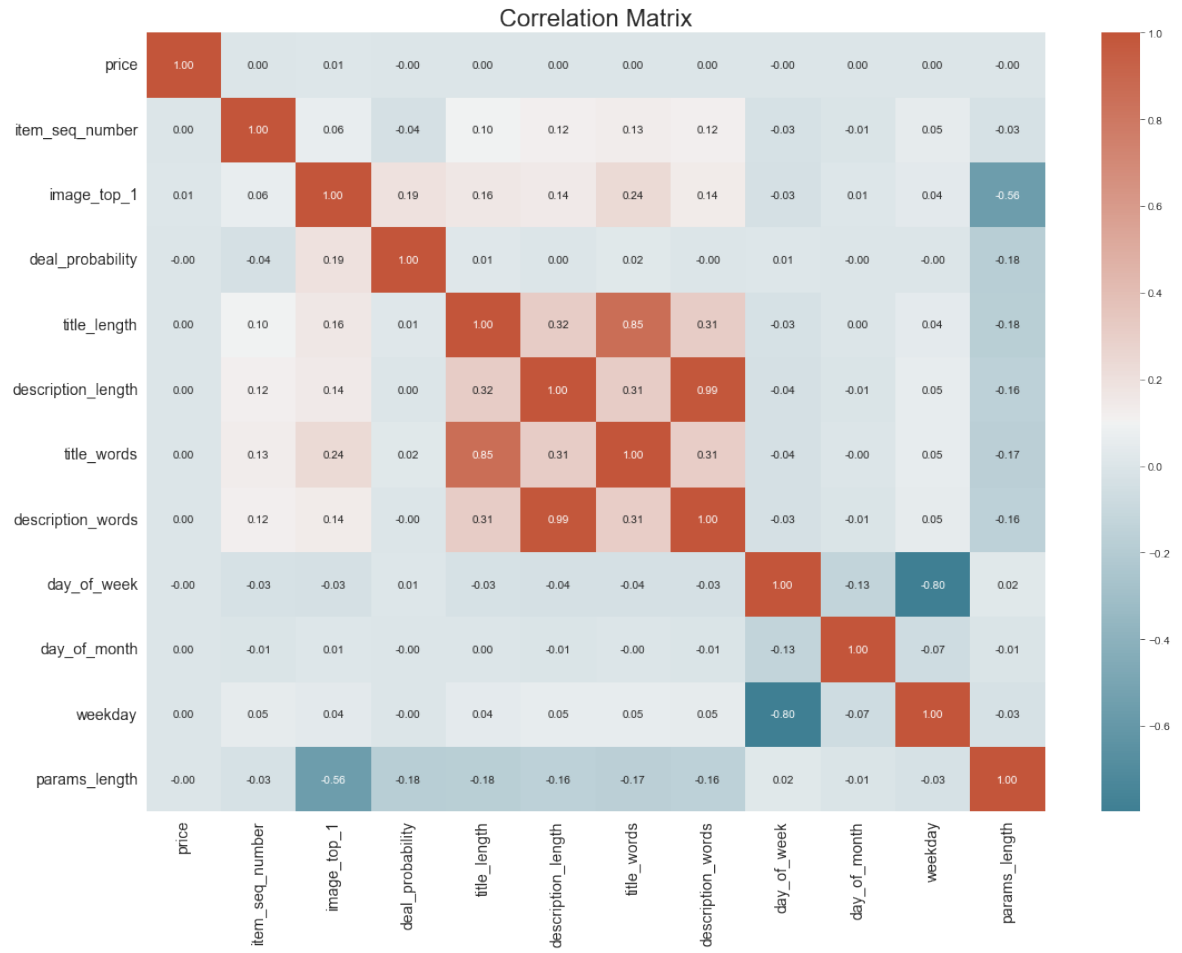

Figure 11: Correlation heatmap of numerical features

2.8.3 Initial Hypotheses and Impact

Based on our initial exploration, we hypothesize that:

- **Price Sensitivity:** Ads with moderate prices (neither too low nor too high) are more likely to have higher deal probabilities.
- **User Type Influence:** Ads posted by private users may have different demand patterns compared to those posted by companies or shops.
- **Category Impact:** Certain categories, such as "Personal belongings" and "For the home and garden," may have higher demand due to their prevalence in the dataset.

2.8.4 Data Requirements

To ensure the reliability and accuracy of our predictions, we define the following data requirements:

- **Completeness:** All mandatory fields, such as 'title', 'description', 'price', and 'deal_probability', should not have missing values. Optional parameters ('param_1', 'param_2', 'param_3') should be imputed as 'missing' if not provided.
- **Validity:** Numerical features like 'price' and 'deal_probability' should have values within plausible ranges (e.g., 'price' ¿ 0, 'deal_probability' between 0 and 1). Categorical features should have predefined valid categories.
- **Consistency:** Data should be consistent in terms of types and formats. For example, 'activation_date' should be a valid date, and 'user_type' should be one of the predefined categories (e.g., Private, Company, Shop).

- **Accuracy:** The data should accurately reflect real-world conditions. Any anomalies or outliers should be investigated and handled appropriately.

2.8.5 Data Quality Verification Report

We conducted a thorough quality check of the dataset, focusing on completeness, correctness, and the presence of missing values.

Completeness:

- The dataset is complete with respect to mandatory fields like 'title', 'description', and 'price', which are essential for demand prediction.
- Optional fields like 'param_1', 'param_2', and 'param_3' have a high number of missing values, which we addressed by imputing them as 'missing'.

Correctness:

- The 'deal_probability' values are within the expected range of 0 to 1.
- There are no duplicate entries in the dataset.
- The 'activation_date' values are valid dates within the specified range.

Missing Values:

- Missing values were primarily found in 'description', 'price', 'param_1', 'param_2', 'param_3', and 'image_top_1'.
- We addressed missing values by imputing with appropriate placeholder values for categorical features and price feature with the median price within each category to maintain the distribution of prices within similar ads.

2.9 Project Feasibility

2.9.1 Inventory of Resources

The project has access to the following resources:

- **Personnel:** A team of consiting of 3 students with the following roles: data scientist, machine learning engineer and data engineer.
- **Data:** Access to historical advertisement data from Avito, including ad descriptions, prices, and other information.
- **Computing Resources:** PC with 16 CPU cores, 32 GB of RAM and NVIDIA GPU RTX 3060.
- **Software:** Machine learning libraries such as scikit-learn, TensorFlow, and PyTorch; data processing tools like pandas; data visualization tools like Matplotlib and Seaborn; ETL Framework Apache Airflow.

2.9.2 Requirements, Assumptions, and Constraints

Requirements:

- The project must be completed within six months.
- The model should achieve an RMSE of 0.22 or lower.
- Data privacy and security measures must be adhered to, in compliance with legal regulations.

Assumptions:

- The data provided is representative of the overall advertisement behavior on Avito.

- The features selected for modeling are sufficient to capture the factors influencing deal probability.

Constraints:

- Limited historical data may restrict the model's ability to generalize.
- High-dimensional data may require significant computational resources for processing and model training.
- Large amount of unique categories in categorical features may cause curse of dimensionality after one-hot encoding.

2.9.3 Risks and Contingencies

Risks:

- **Data Quality Issues:** Poor data quality could affect model performance.
- **Model Overfitting:** The model may overfit the training data and perform poorly on unseen data.
- **Computational Limitations:** High computational requirements may delay model training and deployment.

Contingencies:

- Conduct regular data quality checks and cleaning processes.
- Implement cross-validation techniques to mitigate overfitting.
- Optimize computational resources and GPU-based algorithms.

2.9.4 Costs and Benefits

Costs:

- Investment in computational resources.
- Time and effort spent by the data science and engineering team.
- Potential costs associated with data storage and management.

Benefits:

- Improved accuracy in demand predictions, leading to optimized ad listings and pricing strategies.
- Increased user satisfaction and retention on the Avito platform.
- Higher revenue generation through improved transaction volumes.

2.9.5 Feasibility Report

Based on the initial exploration and data quality checks, the project is feasible. The data is sufficiently detailed and comprehensive to support the development of a robust machine learning model. The potential benefits in terms of increased revenue and user satisfaction justify the investment in this project.

We conducted a preliminary proof-of-concept (POC) using a linear regression model and simple feature engineering and transformations. The initial results show an RMSE of 0.2408, which is close to the target RMSE of 0.22. With further feature engineering and model tuning, it is likely that we can achieve the desired performance.

Overall, the project is viable and has the potential to significantly enhance Avito's service offering, providing a competitive edge in the online marketplace.