Roni Muradov, Rotem Doron & Yoav Sela

Sberbank Russian Housing Market

predective analysis for the sberbank Russian housing market

Presentation - 2024

Introduction

Project overview

Our project aimed to develop an accurate predictive model for housing prices in Russia, using a comprehensive dataset provided by Sberbank. This dataset includes diverse features related to property details, geographical factors, and socio-economic indicators

Significance of the Project:

This research is crucial for real estate stakeholders and economic analysts as it provides insights into the factors driving housing prices in one of the world's most extensive and varied markets

Objectives

Technical Objective:

Implement and compare several advanced machine learning techniques to identify the most effective model for predicting housing prices.

Analytical Objective

Analyze the impact of different features on housing prices and enhance model performance through strategic feature engineering and selection.

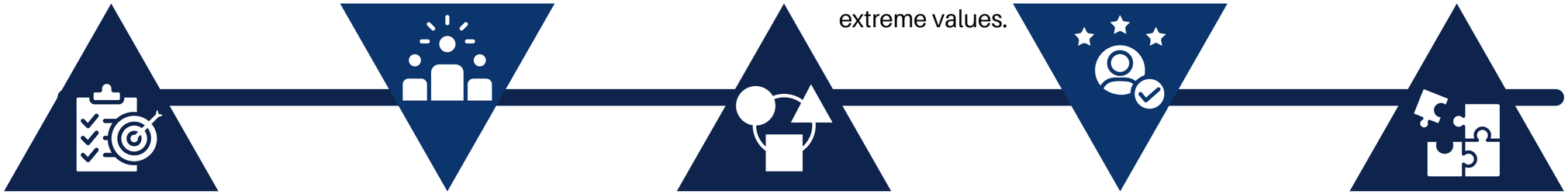
Data Overview and Initial Analysis

Initial Data Exploration

We began with an exploratory data analysis (EDA) to understand the distributions, presence of outliers, and potential correlations among the variables. This step was crucial for identifying anomalies and trends that could influence our predictive models

Feature Engineering:

- Utilized polynomial feature generation.
- Generated interaction features.
- Decomposed 'timestamp' into discrete 'year' and 'month' features for more granular analysis.
- Applied OneHotEncoder to transform categorical data.
- normalized several skewed features using logarithmic transformations to reduce the effect of extreme values.



Dataset Characteristics:

comprises a rich array of features totaling over 290 variables. These include quantitative attributes like apartment area, number of rooms, and price, along with qualitative attributes such as ecological condition and proximity to key facilities

Handling Missing Values

Significant challenges with missing data, which we addressed by employing multiple imputation techniques.
To tackle this, we categorized features into integers and floats, conducting imputation for integer features using the median and for float features using the mean.

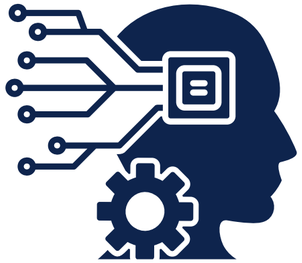
Data Preprocessing:

=Selected important features based on their correlation with the target variable 'price_doc', prioritizing impactful predictors.
- Conducted missing value imputation to maintain dataset integrity.
- Implemented preprocessing steps such as outlier removal to refine the dataset.

Insights from Initial Analysis:

All these measures are directed at constructing a robust model capable of capturing underlying patterns, which enhances predictive accuracy and optimizes the model for performance during the training phase.

Methodology and Model Development



Model Selection and Tuning:

Employed advanced models including RandomForest, XGBoost, and LightGBM. Tuned hyperparameters using cross-validation to enhance performance and ensure robust generalization

Feature Engineering:

Refined our approach by creating polynomial and interaction features, emphasizing selection based on feature importance to boost predictive accuracy

Ensemble Techniques:

LightGBM itself is a boosting algorithm, which inherently uses an ensemble method by building sequential trees that correct the previous trees' errors.

Challenges Addressed

Model Complexity and Overfitting: Addressed challenges related to high dimensionality and potential overfitting by applying feature selection based on model importance scores.

Evaluation and Comparison

Comparison of LGB and XGBoost Results:

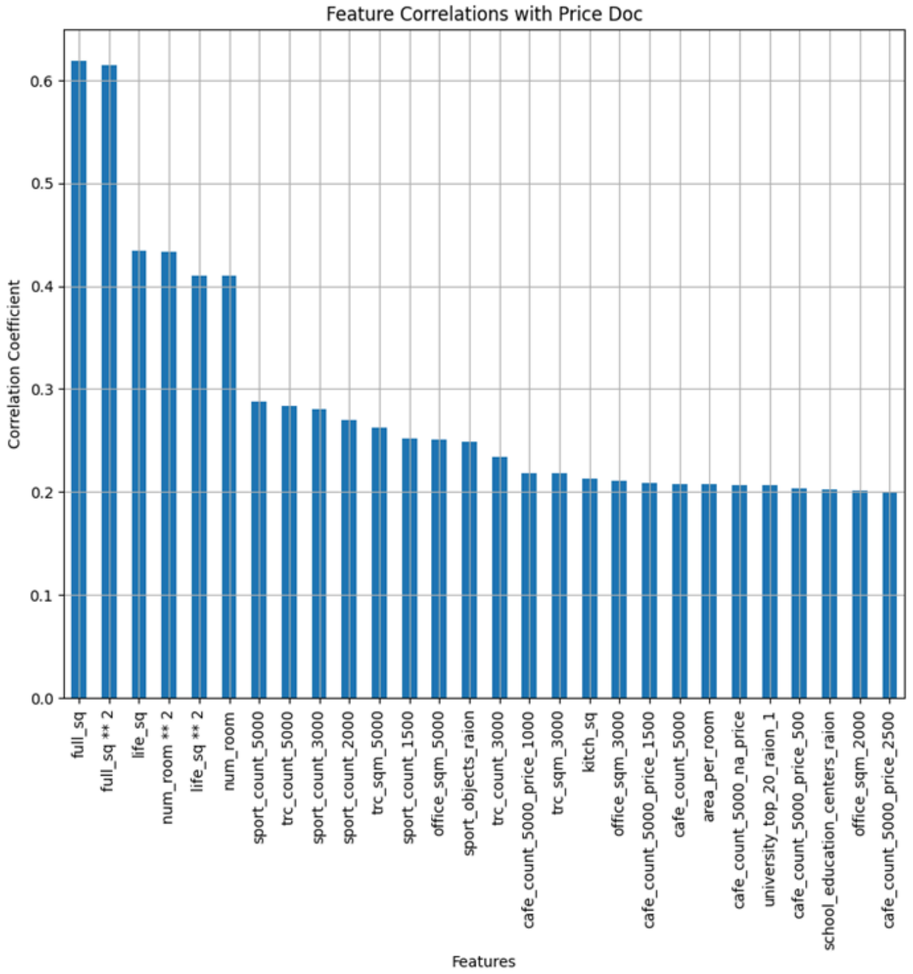
LightGBM (LGB) Results:
•Mean Squared Error: 0.1995959724730162
•Mean Absolute Error: 0.270406748908720 3
•R-squared: 0.4131338715713392
•Root Mean Squared Error: 0.4467616783558 1154

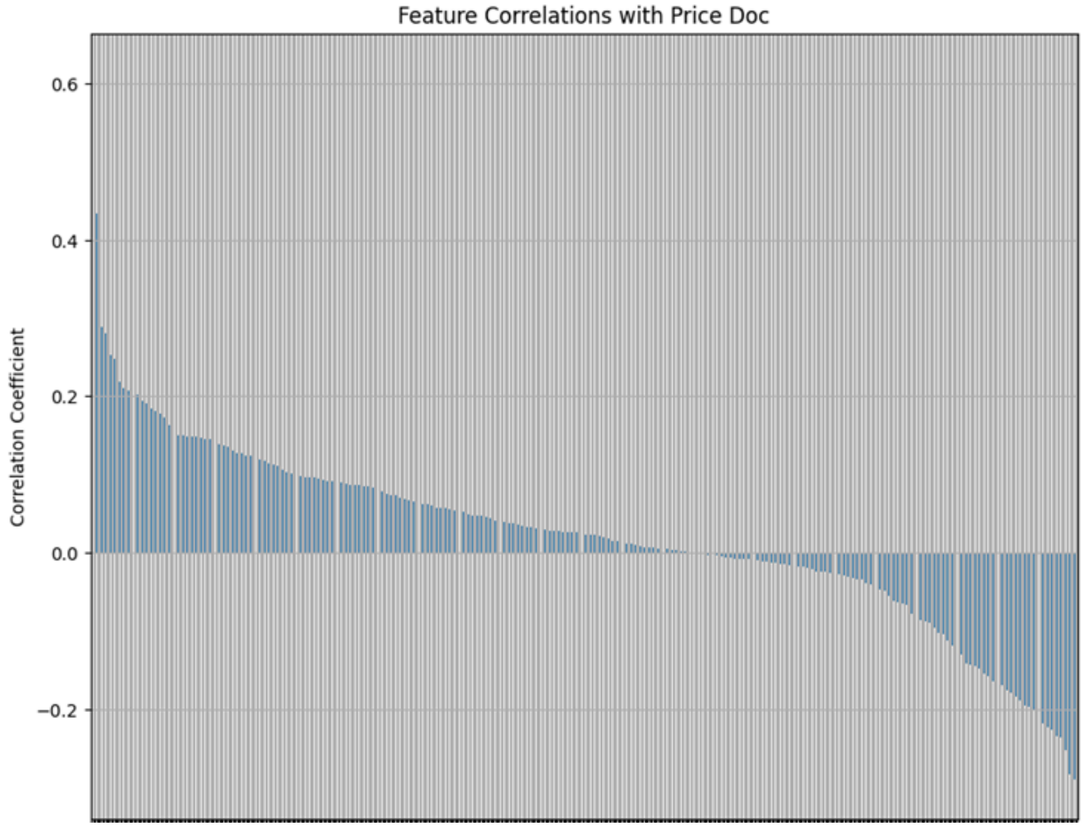
XGBoost Results:
•Mean Squared Error: 0.2002513448708 4675
•Mean Absolute Error: 0.270292806433237 5
•R-squared: 0.411206972595863 7
•Root Mean Squared Error: 0.447494519375206 2

LightGBM shows a marginally better performance than XGBoost with a lower Mean Squared Error and a higher R-squared, indicating closer predictions and a better fit to the data, respectively.

Importance Graphs

All the features with correlation coefficient greater then 0.2





Task2 and Final Project results- comparison

This is the result of Task2

This is the result of the Final Project

The final project's use of LightGBM with more features outperformed Task 2's XGBoost with fewer features, highlighting the benefits of a richer feature set and the effectiveness of the chosen model.

Tuning Parameters of LGB

•colsample_bytree (0.3): Limits each tree to using 30% of the features, reducing overfitting.

•learning_rate (0.05): Slows down the learning process for better generalization.

•max_depth (5): Sets trees to moderate complexity, preventing overfitting.

•n_estimators (350): The number of trees, balancing accuracy and computation.

•subsample (0.2): Each tree trains on 20% of data, helping to avoid overfitting.

Key Insights Gained

•Analysis revealed critical insights into market dynamics. Features like life_sq, full_sq and num_room were the most significant features.

We improved our model by first using features with any positive correlation to our target, and then narrowing down to those with a higher correlation, which enhanced our results, showcasing the power of selective feature inclusion.

Our initial approach to impute missing values was grounded in the belief that using KNN based on 'sub_area' would mirror the nuances of real estate pricing. We anticipated that the locality-specific imputation would enhance our model by incorporating regional trends. However, this method led to a disappointing increase in RMSLE from 0.35 to 0.66, indicating a deterioration compared to our baseline model. This setback prompted us to pivot to simpler imputation methods like mean and median, which ultimately yielded more reliable results.

Summary of Key Actions and Final Model Evaluation

Action Recap:

Data Preprocessing:
Employed robust methods for handling missing data, normalizing distributions, and encoding categorical variables to prepare a clean and analysis-ready dataset.

Feature Engineering:
Enhanced model input through the creation of polynomial features, interaction terms, and careful selection based on feature importance analyses.

Model Development:
Chose LightGBM for its efficiency, fine-tuning its parameters to minimize RMSE and enhance model performance.

Outcome and Evaluation:
Surpassed initial Kaggle benchmarks, with the final LightGBM model yielding improved scores, affirming the effectiveness of our streamlined approach.

Conclusion and Reflections

Project Learnings:

Our project's evolution embodies Occam's Razor, particularly through the selection of LightGBM for its efficiency and precision. We moved from complex to simpler imputation techniques and a minimal yet impactful feature set. This streamlined approach not only simplified our model but also significantly boosted its predictive power, reaffirming that in predictive analytics, often less is more.

Further Applications:

Findings from this study pave the way for applying our refined modeling techniques to broader domains, offering potential advancements in predictive analytics across various industries facing similar data challenges.

Thank You