Predicting Moscow Housing Prices

Final Projrect - Machine Learning

Niv Sade & Tom Regev

Project Overview

- **Objective: Predict apartment prices in Moscow.**
- **Model Used: XGBoost.**
- **Main Focus: Reducing RMSLE by cleaning data, engineering new features, and optimizing model parameters.**
- **Final RMSLE: Achieved a significant reduction compared to the baseline.**

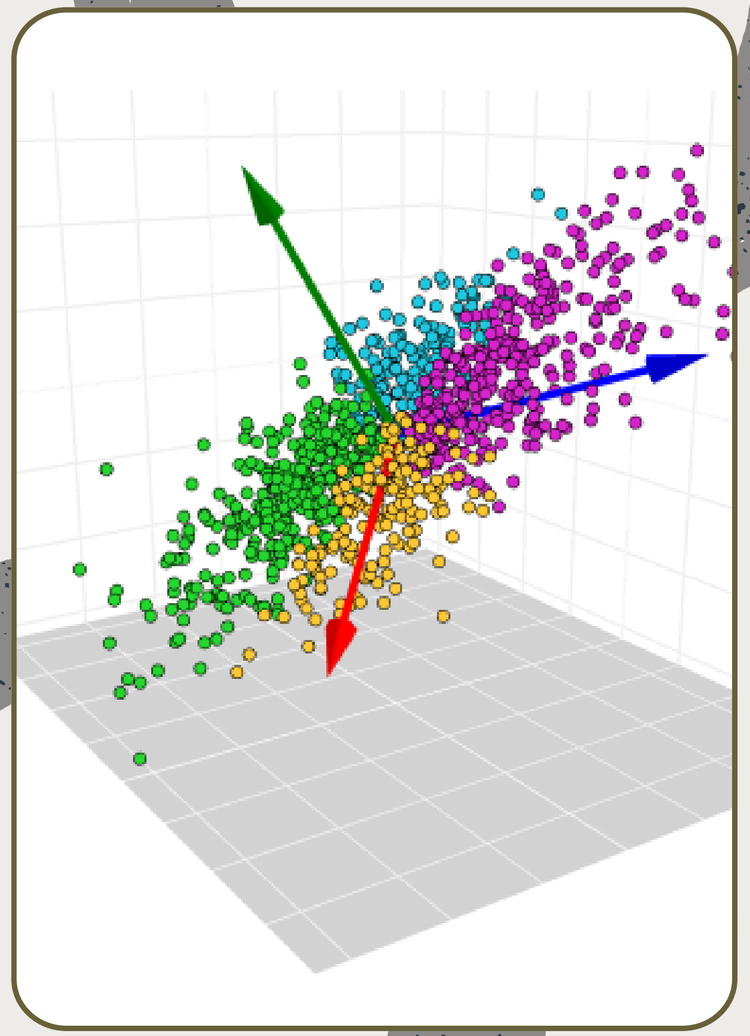
Data Cleaning & Handling Outliers

- Identified and converted impossible values to NA like:
  - life_sq > full_sq
  - floor > max_floor
  - Year values like 2502009

- Filtered unrealistic price values:
  - Example: 150m² apartment priced at only 1 million rubles.

Feature Engineering

- **Categorized features into groups: (e.g., Apartment Structure, Location, Amenities).**
- **Applied PCA on each category to determine the most significant variables.**
- **Correlation Analysis: Selected variables with the highest correlation with price_doc**
- **Final Model: Reduced the number of variables to 45.**



Feature Engineering - New Variables

Neighborhood Segmentation

Categorized neighborhoods as “cheap”, “middle”, and “expensive” based on average price.

Transaction-Building Year Difference

Combined transaction date with building year.

Create new features

like Room Size and kitch size

Handling Missing Data (NA Values)

We tried a model-based approach to fill missing values:

- **Estimated missing values based on the same building's properties.**

- **Features completed:**
  Max floor, Build Year, Apartment condition, Number of Rooms.

Eventually, we applied an imputer function to fill missing values with the mean of each feature, which resulted in a slight improvement in performance.

Model Optimization

- **Hyperparameter tuning for XGBoost:**
  - **Adjusted max_depth, learning_rate, colsample_bytree, and reg_lambda to best fit our dataset.**

- **Split data into two models:**
  - **Investment apartments vs. Owner-occupied apartments.**
  - **Different features affect each type differently.**
  - **Improved accuracy by handling them separately.**

Results & Impact on RMSLE

Baseline RMSLE: 0.535

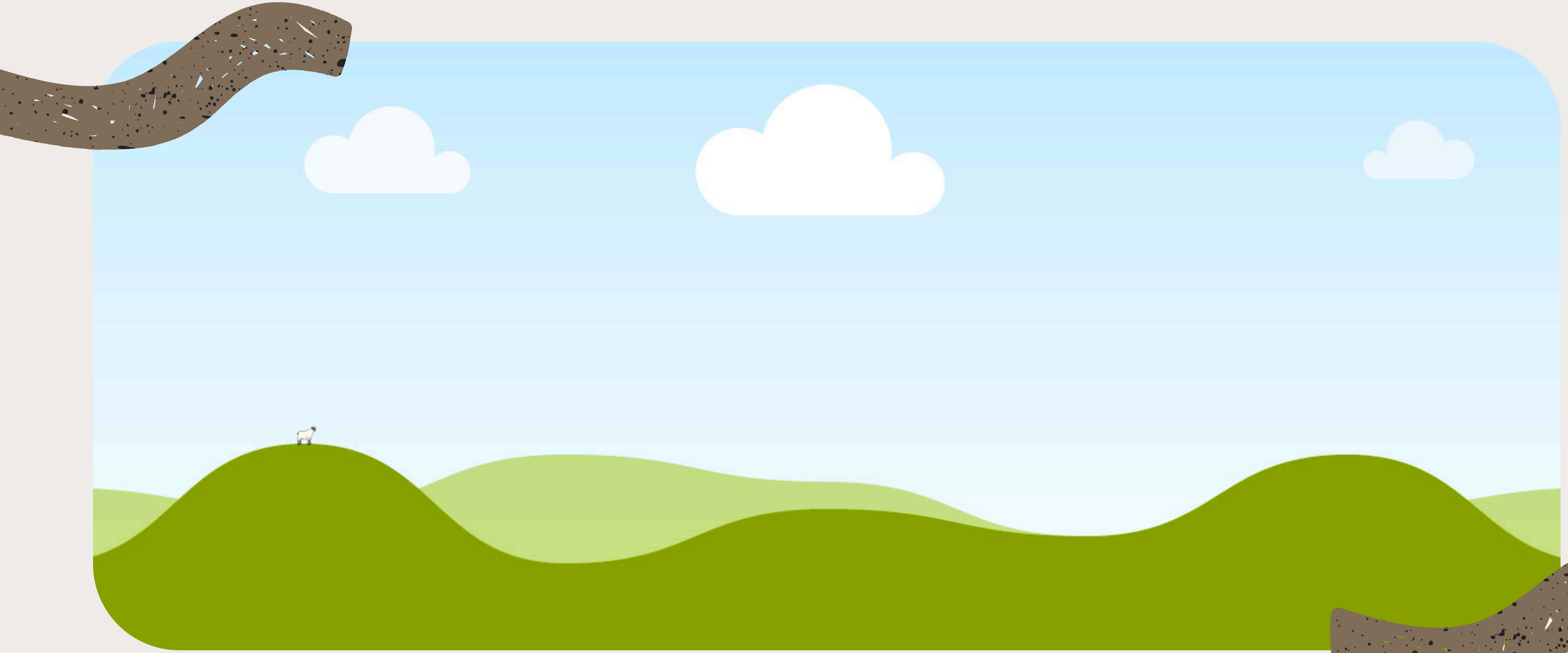
Final RMSLE: 0.316

Biggest gains came from:

- **Adding domain-specific features.**
- **model-based approach to fill missing values:**
- **Removing unrealistic values.**
- **Removing observations with unrealistic prices.**

”אנחנו לא יכולים להוריד את מחירי הדירות, אבל לפחות הצלחנו להוריד את RMSLE! 🚀🏡“

-Niv Sade-



Moscow's skyline **shows** stunning architecture!