Machine Learning Final Project: Predicting House Prices:

Authors: Eliyahu Elimelech & Segev Cohen

Project Overview:

The objective of the project was to predict house prices using the Sberbank dataset, consisting of 30,471 records and 391 features. The primary evaluation metric used was RMSLE (Root Mean Squared Logarithmic Error).

Dataset Overview:

- Target Column: price_doc (house price).

- Initial Analysis: Explored price distributions, yearly trends, and correlations.

- Challenges Identified: Missing values and outliers in key columns.

Data Cleaning:

- Logical Corrections: Ensured full_sq > life_sq and removed inconsistent build years (e.g., >1500).

- Handling Missing Values:

  - Columns with excessive missing data were removed.

  - Missing values were imputed using averages or KNN imputation methods.

- Outlier Removal: Removed extreme outliers in price_doc (lower and upper 1900 values).

Feature Engineering:

- Label Encoding: Converted categorical and binary features into numerical values.

- New Features Created:

  - Average room size (life_sq / num_room).

  - Relative kitchen size (kitchen_sq / full_sq).

  - Temporal features (month and day).

- Dimensionality Reduction: Applied PCA to improve efficiency and handle multicollinearity.

Model Development:

- Model Exploration:

  - Tried Linear Regression, Polynomial Regression, Lasso, Ridge, and Stepwise feature selection.

  - Advanced models included Random Forest and XGBoost.

- Final Model: XGBoost, chosen for its lowest RMSLE.

- Training Process: Utilized 5-fold cross-validation to optimize performance.

Results and Improvements:

- Baseline RMSLE: 0.469

- Sequential Improvements:

  1. After data cleaning and feature engineering: RMSLE = 0.469.

  2. Incorporating outlier handling: RMSLE = 0.334.

  3. Adding KNN imputation: RMSLE = 0.331.

  4. Hyperparameter tuning: RMSLE = 0.330 (final result).

Key Takeaways:

- Effective feature engineering and data cleaning significantly enhanced model performance.

- XGBoost proved to be the most effective model for this dataset.

- Hyperparameter tuning, though marginal, contributed to additional improvement in RMSLE.

Conclusion: The project showcased the importance of robust preprocessing, thoughtful feature engineering, and advanced modeling techniques in achieving accurate house price predictions.