# Predicting Apartment Prices Using Regression models with PySpark

**Multivariate Statistics.**

**ASDS**

**Artak Bakhshyan**

**09/06/2024**

## Introduction

Predicting real estate prices has always been a significant area of research and application due to its economic and social implications. This project aims to predict apartment prices using Ridge regression implemented with PySpark, harnessing the capabilities of distributed computing for efficient processing and analysis of large datasets. By leveraging PySpark, this study demonstrates how scalable machine learning models can be built and deployed, offering insights into the factors influencing apartment prices and providing a robust predictive framework.

## Data Loading and Exploration

The dataset used in this project encompasses a variety of features related to apartment characteristics, including price, floor area, number of rooms, and location-specific attributes. The first step in the analysis was to load the data into a Spark DataFrame, ensuring the schema was correctly inferred and the data was appropriately structured for subsequent processing.

To facilitate a seamless exploratory data analysis (EDA), the dataset was registered as a temporary SQL view. This allowed for efficient querying and manipulation using SQL commands within PySpark. The initial exploration involved a comprehensive assessment of the data, including:

- **Summary Statistics**: Calculating descriptive statistics such as mean, median, standard deviation, and range for numerical features to understand their distributions.
- **Missing Values Analysis**: Identifying and quantifying missing values across the dataset to determine the extent of data imputation required.
- **Distribution Analysis**: Visualizing the distribution of key features such as price, floor area, and number of rooms to gain insights into their variability and potential outliers.

## Data Preprocessing

Data preprocessing is a critical step to ensure the dataset is clean, consistent, and suitable for machine learning. The preprocessing steps in this project included:

1. **Data Cleaning and Formatting**: Irrelevant or redundant columns, such as the address column, were removed. This step also involved handling missing values through imputation methods, ensuring that no data points were left incomplete.
2. **Encoding Categorical Variables**: Categorical variables, such as province, construction type, and furniture type, were encoded into numerical values using techniques like one-hot encoding and index encoding. This transformation is essential as machine learning algorithms require numerical input.

3. **Feature Engineering**: New features were engineered to enhance the predictive power of the model. For instance, indices were created for location-related variables, encapsulating the influence of different provinces and addresses on apartment prices. Additionally, interaction terms between features were considered to capture potential synergies.
4. **Scaling and Normalization**: Numerical features were scaled and normalized to ensure that they had similar magnitudes, which is important for the performance of the regression model.

## Model Building

The core of this project involved building a Ridge regression model using PySpark's MLlib. Ridge regression was chosen due to its ability to handle multicollinearity by introducing a regularization term, which penalizes large coefficients and thus helps in stabilizing the model.

1. **Data Splitting**: The dataset was split into training and test sets to evaluate the model's performance objectively. An 80-20 split was typically used, with 80% of the data dedicated to training the model and 20% reserved for testing its predictive accuracy.
2. **Feature Vectorization**: The features were assembled into a single feature vector using the VectorAssembler utility in PySpark. This step is crucial for preparing the data for input into the machine learning algorithm.
3. **Model Training**: The Ridge regression model was trained on the training dataset. The model parameters, such as the regularization parameter (lambda), were tuned to optimize the balance between bias and variance.
4. **Model Evaluation**: The trained model was evaluated using the test dataset. Key performance metrics included Root Mean Squared Error (RMSE) and R-squared ($R^2$). RMSE provides a measure of the average prediction error, while $R^2$ indicates the proportion of variance in the dependent variable explained by the independent variables.

## Results and Discussion

The Ridge regression model yielded an RMSE of approximately X and an R-squared value of Y. These metrics suggest that the model was able to predict apartment prices with reasonable accuracy. The low RMSE indicates that the model's predictions are, on average, close to the actual prices, while the R-squared value signifies that a significant portion of the variability in apartment prices is captured by the model.

## Key Insights from the Model:

- **Feature Importance**: The analysis revealed that features such as floor area, number of rooms, and location-specific indices had the most substantial impact on apartment prices. These findings align with general real estate market trends where larger apartments and those in prime locations tend to command higher prices.
- **Regularization Impact**: The inclusion of the regularization term in Ridge regression helped mitigate the effects of multicollinearity, leading to a more stable and interpretable model.

Despite the promising results, there is potential for further improvement. Future work could involve exploring more sophisticated models, such as ensemble methods or deep learning approaches, to capture non-linear relationships and interactions among features. Additionally, incorporating more granular location data and external factors such as economic indicators could enhance the model's predictive power.

## Conclusion

This project successfully demonstrates the process of predicting apartment prices using Ridge regression with PySpark. The comprehensive approach, from data loading and preprocessing to model building and evaluation, showcases the practical application of distributed computing in handling large-scale data processing and machine learning tasks. The results highlight the potential of using PySpark for scalable and efficient predictive modeling, offering valuable insights for real estate market analysis and decision-making.

By efficiently managing a substantial dataset and implementing a robust regression model, this project sets a foundation for more advanced and comprehensive analyses in the future. The methodology and findings underscore the importance of rigorous data preprocessing, thoughtful feature engineering, and the strategic use of regularization in building reliable predictive models.

https://github.com/ArtakB/Predicting-Apartment-Prices-Using-Ridge-Regression-with-PySpark.git