A black background with grey leaves

Description automatically generated

**Linear Regression Analysis**

-- Paris Housing Price Prediction

A grey logo on a black background

Description automatically generated

Gang Wei

06/29/2024

Linear Regression Analysis: Paris Housing Price Prediction

**Main Objective:** Use Python linear regression to develop models to predict housing prices in Paris.

**Introduction**

Housing purchase is one of the most important decisions to make in most people’s lives. There are various attributes of houses like number of bedrooms, square footage, location, age, interest rate, and more. This project predicts the housing prices in a specific location (Paris, French) utilizing Linear Regression model. The analysis identifies key factors influencing housing prices, helps investors or home buyers with their purchase decision.

To address the challenge, I propose a solution that leverages advanced data analytics and machine learning models. By analyzing housing attributes like number of previous owners, house buyers could better target the main affecting factors. This data-driven approach develops a formula for buyers to use to estimate the housing prices in Paris.

Various locations might result in different affecting factors, for example Paris residents care the total square meters more, while Beijing residents might prefer newer houses. When it comes to different location or different culture, similar technique could be applied to achieve our goal: develop appropriate housing price prediction formula to help house buyers with estimate prices.

**Data Dictionary**

Data Source: <https://www.kaggle.com/datasets/mssmartypants/paris-housing-price-prediction> from <https://www.kagge.com>

Data Attributes: Table of 10,000 instances and seventeen features (all numerical data)

Input variables:

* squareMeters - total area of the house
* numberOfRooms - total rooms the house has
* hasYard - whether the house has a yard or not
* hasPool - whether the house has a pool or not
* floors - number of floors
* cityCode - zip code
* cityPartRange - the higher the range, the more exclusive the neighborhood is
* numPrevOwners - number of previous owners
* made – year the house built
* isNewBuilt – whether the house is newly built or not
* hasStormProtector – if the house has storm protector
* basement - basement square meters
* attic - attic square meters
* garage - garage size
* hasStorageRoom – where the house has storage room or not
* hasGuestRoom - number of guest rooms

Output variables:

1. Price – The value of the house (kaggle, 2021)

**Results and Model Interpretation**

Based on the dataset we can see that the target variable (price) is continuous, I initialed a regression model for the data. After data analysis and visualization, I decided to use the linear regression for the housing price prediction.

According to David Olive, if the target variable Y is quantitative and that at least one predictor variable x is quantitative, then the multiple linear regression model is often a very useful model to use just like in our case (Olive, 2017).

After implementing the linear regression machine learning algorithms, according to the correlation, I developed a full model and a second model with highly price related variable to find the best model for the housing prediction.

|  |  |  |
| --- | --- | --- |
|  | **Full model** | **2nd model** |
| **Model** | y=4926.76+100x1+… | y=6460.88+99.999x |
| **R square** | 99.99996% | 99.99987% |
| **RMSE** | 1921 | 3269 |
| **Mean** | 4,951,703 | 4,951,703 |

R square is the coefficient of determination, it is a number between 0 and 1 representing the percentage that variance of target variable could be explained by the predictor variables. The closer the R square to one, the more price variance could be explained by the predictor variables; RMSE is short for root of mean squared errors. RMSE measures the average difference between the predicted values and the actual values. It provides an estimation of how well the model can predict the target value (Olive, 2017).

As the table shown above, both of their coefficients of determinant are close to one, almost all the price variances could be explained by the x independent variables; the model 2 RMSE seems too big for the full model, but comparing the average housing price of five million, the RMSE is minor. In a word, the full model is slightly better with higher R square and lower RMSE, but both models could predict the housing price well.

Considering the cost and time to collect all sixteen variable data we may choose the second model instead of the full model.

**Conclusion**

Based on the machine learning algorithm analysis, it is evident that the linear regression model performs well on the Paris housing price prediction with extremely high coefficients of determinants and low root mean squared errors. Therefore, deploying the linear regression model for housing price prediction presents a compelling choice.

# References

kaggle. (2021, june 9). *Paris Housing Price Prediction*. From kaggle.com: https://www.kaggle.com/datasets/mssmartypants/paris-housing-price-prediction

Olive, D. J. (2017). *Linear Regression.* Springer Cham.