Report
2023

08/01/2023
IT Outsource
Author: Maksat Bairamov

Predicting House Prices with Regression Model:

An Exploratory Study and Model Evaluation

1. Introduction

The objective of this project is to create a robust and accurate model that predicts house prices based on various features. The prediction of house prices is crucial for both real estate professionals and potential buyers or sellers. By developing a reliable prediction model, we aim to provide valuable insights and assist in making informed decisions in the real estate market.

We will begin by exploring and preparing the "California Housing dataset" available in the scikit-learn library. The dataset contains features such as median income, house age, average number of rooms, and geographical coordinates, among others. To ensure the dataset's integrity, we conducted preliminary data analysis and handled missing values by imputing them with the mean values of the corresponding features.

2. Data Preparation

The "California Housing dataset" served as the foundation for our analysis. We carefully examined the data distribution, checked for any anomalies or outliers, and applied necessary data transformations to prepare it for model training. By addressing missing values, we aimed to create a comprehensive dataset that would facilitate the accurate training of our models.

3. Model Selection

In our pursuit of accurate house price prediction, we evaluated several regression models: Linear Regression, Random Forest, XGBoost, and LASSO. Each model brings unique characteristics and advantages to the task. We selected these models to encompass a diverse range of regression approaches and evaluated their performance using appropriate metrics.

4. Model Training and Evaluation

The selected models underwent training on the training dataset and subsequent evaluation on the test dataset. We used key metrics such as Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and R-squared (R2) to assess their performance. Additionally, we employed cross-validation techniques to gauge the models' robustness and generalizability on different subsets of the data.

5. Results Visualization

To gain further insights into the models' predictions, we visualized the actual house prices against the predicted values using scatter plots. These visualizations provided an intuitive comparison, allowing us to understand the models' accuracy and deviations in predicting house prices. Additionally, we visualized the importance of features in the XGBoost model to identify the most influential factors affecting house prices.

6. Discussion of Results

Based on the performance metrics and feature importance analysis, we deliberated on the best-performing model for predicting house prices. The XGBoost model, after careful hyperparameter tuning using GridSearchCV, emerged as the top performer. We further investigated the key features that significantly influence house prices, revealing the importance of 'MedInc' (median income of residents), 'Latitude,' and 'Longitude' in driving house price predictions.

7. Conclusions

Our project's outcome was the successful development of reliable models capable of predicting house prices with acceptable accuracy. The XGBoost model's impressive performance, coupled with its interpretable feature

importance, offers valuable insights for real estate stakeholders. We believe that the predictive power of our models will aid in better decision-making within the real estate market.

8. Future Possibilities

While our project achieved promising results, there are potential avenues for further improvement. We recognize the need for an exhaustive hyperparameter search to fine-tune the models and potentially enhance their predictive capabilities. Additionally, incorporating more sophisticated techniques for handling missing data may further improve model accuracy. Furthermore, creating new features based on domain knowledge could provide valuable contextual information and boost the overall predictive performance.

In conclusion, our project serves as a valuable foundation for future developments in the realm of house price prediction using regression models. The insights gained from this project can assist various stakeholders in the real estate market, enabling more informed decisions and optimizing the overall real estate experience.