Housing Price Predictions with Various Regression Models

Jessica Israel
Computer Science Department
College of Wooster
Wooster, OH

Quoc Do
Computer Science Department
College of Wooster
Wooster, OH

Abstract

Housing prediction is always a critical task in real estate and urban planning, requiring accurate and reliable models to capture complex market dynamics. In this study, we will analyze housing prices using machine learning prediction from the **Linear Regression** model and the **Polynomial Regression** model. Leveraging a real-world housing dataset, the models are trained and evaluated using key performance metrics, including **Mean Squared Error (MSE)** and **R^2 score**. Our study focuses on identifying key features that influence housing prices and assessing the predictive accuracy of the models. By comparing the results from using our models, we demonstrate how **polynomial regression** can model complex, nonlinear relationships in housing prices, while **linear regression** serves as a baseline for comparison.

Keywords: Linear Regression, Polynomial Regression, Housing Prices

Introduction

In this paper, we will be studying the topic of linear regression in machine learning and will use the California Housing dataset. The dataset labels have values that represent the median house value in thousands of dollars. The size of the data is 20,640 samples, each sample consisting of ten attributes: Longitude, Latitude, Housing Median Range, Total Rooms, Total Bedrooms, Population, Households, Median Income, Median House Value, and Ocean Proximity. Linear regression and polynomial regression will be applied to this dataset.

Background

Housing prices have always been an important aspect of life once a person becomes an adult, and it is important to know and understand the factors that affect housing costs when searching for a place. A way that helps us analyze this kind of data is linear regression. Linear regression is a statistical method that estimates the linear relationship between dependent variables and independent variables. It is also basic and commonly used for data analysis.

We will first implement linear regression and then implement polynomial regression. There are several studies similar to what this paper is researching.

Jiajun Yu writes an article called, "A Multivariate Regression Analysis of Factors Influencing California Housing Prices". Their goal is to find the factors that influence housing prices in California through multiple linear regression and multicollinearity analysis. Through their findings, it is discovered that home age has a pivotal role in housing prices in California, and geographical location also has a determining influence on housing prices. (Yu 2024)

Problem Description

The goal of this research is to use linear regression for an analysis of the factors influencing California housing prices. Once this is complete, the linear regression model will be compared to a polynomial regression model. We plan to see if we get similar results for the factors that influence housing prices as Juajun Yu with our linear regression model and polynomial regression model.

Data Preparation

Here we are using California housing data. This data is about houses found in a given California district and some summary stats based on the 1990 census data. The original data has 10 columns and 20,640 lines of data. Those 10 attributes include:

- **longitude**: A measure of how far west a house is; a higher value is farther west.
- **latitude**: A measure of how far north a house is, a higher value is farther north .
- **housing_median_age**: Median age of a house within a block; a lower number is a newer building.
- **total_rooms**: Total number of rooms within a block.
- **total_bedrooms**: Total number of bedrooms within a block.

- **population**: Total number of people residing within a block.
- **households**: Total number of households, a group of people residing within a home unit, for a block.
- **median_income**: Median income for households within a block of houses (measured in terms of thousands of US Dollars).
- **median_house_value**: Median house value for households within a block measured in US Dollars).
- **ocean_proximity**: Location of the house w.r.t ocean/sea.

Implementation of Linear Regression

The implementation begins with importing libraries: **"pandas"** manipulations, **"scikit-learn"**for machine learning tasks, and specific modules for model evaluation such as **"mean_squared_error"** and **"r2_score"**.

The dataset is stored in a CSV file and is loaded into **"pandas"DataFrame** for further analysis. A key preprocessing step involves encoding the categorical variable **"ocean_proximity"** into numerical values using the replace method to facilitate compatibility with machine learning algorithms. This systematic approach sets the foundation for subsequent tasks, including data exploration, feature engineering, model training, and evaluation.

Once the data is prepared, it is split into **training and testing** with the ratio of 70:30 (70% of the data belongs to the training set, and the rest is for testing) subsets using "train_test_split", ensuring a robust evaluation setup. The workflow provides a clean, structured pipeline for building a linear regression model to predict housing-related outcomes. Next, a linear regression model from scikitlearn is then trained on the processed data. Finally, performance is assessed using metrics like **Mean Squared Error (MSE)** and **R-squared (R^2)** scores to evaluate the model's accuracy and goodness of fit.

Implementation of Polynomial Regression

Polynomial regression is another model that can be used and expands on the limitations of linear regression and can be used for non-linear relationships for the data. We want to see how the models of linear regression and polynomial regression occur. We follow a video by RegenerativeToday that gives a step-by-step process on how to implement polynomial regression.(RegenerativeToday 2023)

The implementation for Polynomial Regression is the same when importing libraries, splitting the data, and evaluating the outcome as the linear regression model with a few differences. One major key difference is when we implemented **StandardScaler** from **"scikit-learn"** to scale our data. Using fit transform for the training data, helps to fit and transform the data and compute the mean and standard deviation for the data. We also import **PolynomialFeatures** which allows us to change the degrees for the polynomial.

Performance is also assessed using the metrics **Mean Squared Error (MSE)** and **R-squared (R^2)** scores to evaluate the model's accuracy and goodness of fit. For the Polynomial specifically, we use **Mean Absolute Error**.

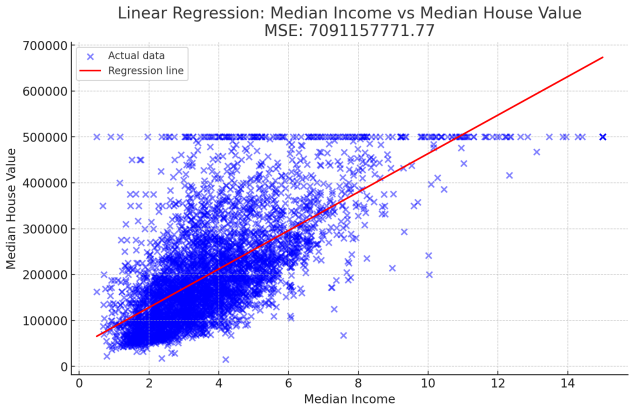

Figure 1: Linear Regression Model

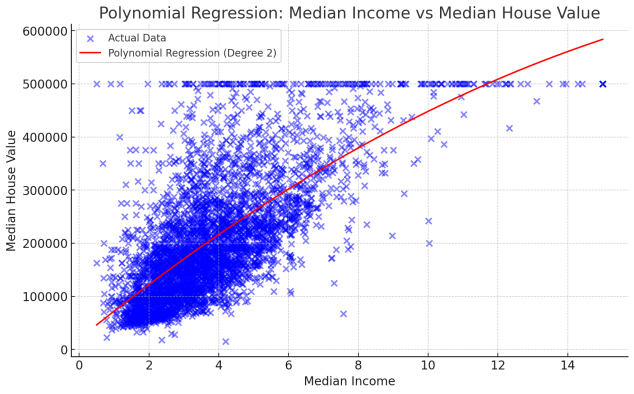

Figure 2: Polynomial Regression Model

Linear Regression Results

Figure 1 illustrates the linear regression model. A good model has low MSE and a high R^2. A closer to 1 R^2 value signifies a better model. The performance of the models was evaluated using the **Mean Squared Error (MSE)** and **R-squared (R^2)** score to determine their accuracy and suitability for predicting housing prices. Three variations of linear regression were implemented:

- **Linear Regression with "ocean_proximity" column:**
  - Mean Squared Error (MSE): 4,646,764,593.33

  - R^2 Score: 0.6514
- **Linear Regression without "ocean_proximity" column:**
  - Mean Squared Error (MSE): 4,738,972,791.40
  - R^2 Score: 0.6445
- **Multi Linear Regression:**
  - Mean Squared Error (MSE): 4,668,669,607.98
  - R^2 Score: 0.6498

Based on the result, **Linear Regression with the "ocean_proximity" column** has the **worst Mean Squared Error** score overall but the **highest R^2 Score** which means this model has the **best prediction** accuracy among those three. On the other hand, **the Multi-Linear Regression model** performed slightly worse, with a marginal increase in MSE and a decrease in R^2 score. Lastly, the worst of the three models is **the model excluding "ocean_proximity"** since it shows a very high MSE and the lowest R^2 Score.

Polynomial Regression Results

Figure 2 shows the polynomial regression model that has the polynomial degree of two that performed the best. In this figure we plot the attributes against the house value.

MSE / R^2 Results:

- **Mean Squared Error (MSE)**: 3,959,342,101.52
- **R-squared (R^2)**: 0.702995

These are the results from **including "ocean_proximity"** in the model.

The MSE of approximately 3.9 billion suggests a substantial average error in the predicted housing prices. This is common in housing price models due to the large range of possible prices, and it may imply that the model may need more feature engineering. Compared to the linear model it is a bit better but not by much.

The R^2 score of about 0.702 shows that the model explains about 70.2 percent of the variance in housing prices, which is a moderate level of accuracy for a basic linear regression model on housing data. While it performs better than the polynomial model, the R^2 score indicates that there is room for improvement, as 30 percent of the variance remains unexplained.

Mean Absolute Error Results:

- **Mean Absolute Error with a polynomial degree of 2:**
  - Test: 45235.386
  - Train: 44842.948
- **Mean Absolute Error with a polynomial degree of 3:**
  - Test: 40534.084
  - Train: 38822.456
- **Mean Absolute Error with a polynomial degree of 4:**
  - Test: 40054.063
  - Train: 35098.504

The results above show the models using different polynomial degrees. The two numbers are testing data and training data results. The closer the numbers are, the better the results. The further they are, the more we overfit. The best model from the three variations is the polynomial model with a degree of two.

Conclusion

Analyzing the result of both linear regression and polynomial regression, we conclude that the polynomial regression model achieves a significantly lower **Mean Squared Error (MSE)** of **3,959,342,101.52**, compared to **4,646,764,593.33** for Linear Regression. Additionally, the Polynomial Regression model has a higher **R^2 score** of **0.703**, indicating it explains approximately **70.3%** of the variance in housing prices, compared to **65.1%** for the Linear Regression model.

The outperforming result of the **Polynomial Regression model** over the **Linear Regression model** underscores the Polynomial Regression model's ability to **capture nonlinear relationships** in the data, making it better suited for datasets with complex patterns like housing price prediction. The models also show that geographical location does impact the housing prices significantly just as it did for Yu's results.

For **housing price prediction**, where such relationships are expected, the additional complexity of Polynomial Regression is justified by its superior performance. As such, it is recommended as the optimal model for this task.

References

RegenerativeToday. 2023. Polynomial Regression in Python - sklearn. https://www.youtube.com/watch?v=nqNdBlA-j4w.

Yu, J. 2024. A Multivariate Regression Analysis of Factors Influencing California Housing Prices. In *Proceedings of the International Conference on Mathematics and Machine Learning*, ICMML '23, 165–169. New York, NY, USA: Association for Computing Machinery.