Linear Regression Model for Housing Price Prediction

Ahmad Naggayev

Content

Introduction => Problem Statement => Dataset

Data Cleaning = > Plots => Linear Regression Model

Model Evaluation => Result/Interpretation

Introduction

In this project, we will demonstrate analysing results of Ames house Prediction and machine learning model implemented to predict house prices.

Housing sales prices are determined by numerous factors such as material quality, living area square feet, size of the garage, location of the house, number of bedrooms, and so on.

We will apply Linear regression models to predict the prices and find out the most optimal model with the maximum R-squared value.

Understanding the problem

Defining Audience

In this model we will focus on realtors as target audience of the prediction. Realtors struggle with accurate pricing due to numerous factor but correct analysing and effective modeling can help them to razer focus on the most important features and make data driven decisions.

Feature Selection

Some features may have little impact, while others are highly correlated. We will be focused mainly on these features in the project: Overall_quality, Gen. living area, Garage Cars, Garage Area and Sales Price as independent variable

Model Effectiveness

To evaluate effectiveness, we assess underfitting (low R^2, poor pattern capture), overfitting (high training R^2, low test R^2), and ensure a balanced R^2 score for a well-generalized model.

Data used

Two datasets (train.csv and test.csv) were provided (csv format) for the project and both were cleaned and saved in data folder for further analysis

Loading Processed Data

[48]:
```
# Load processed data
train_df = pd.read_csv("../data/cleaned_train.csv")
test_df = pd.read_csv("../data/cleaned_test.csv")
```

Essential data preprocessing for optimal model performance.

Handling Null Values

NaN values were filled with zeros in this project, as it was a logical choice that preserved the integrity of the mean and median.

Column Header Consistency

Column headers were converted to lowercase and spaces were replaced with "_" to ensure consistency and prevent errors when accessing specific columns.

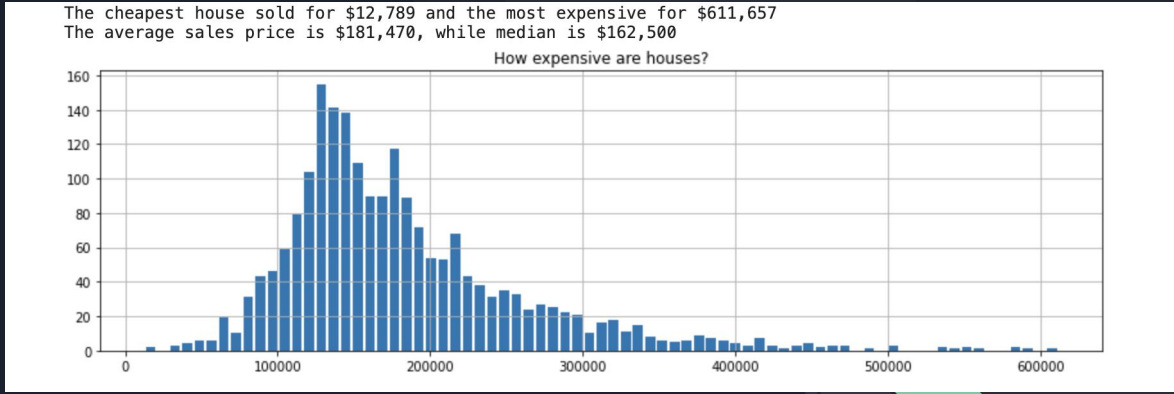
Encoding Categorical Features

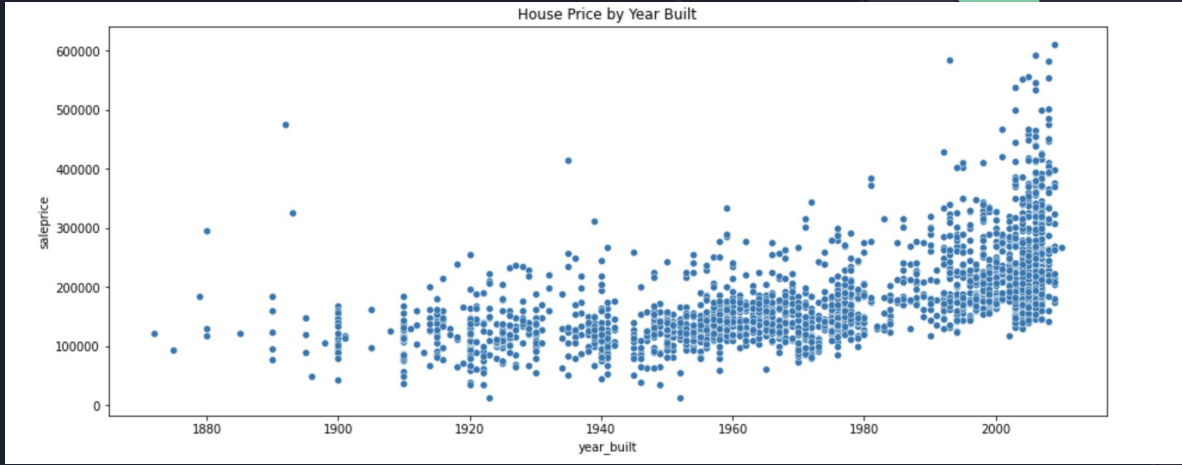
Encoding categorical features with dummy data was not required for this project since the model was based on numerical values for prediction.

Plots

Most houses are priced between $100,000 and $200,000, with a right-skewed distribution indicating a few high-value properties.

Newer houses tend to be more expensive, showing an upward trend in prices over time.





Plots

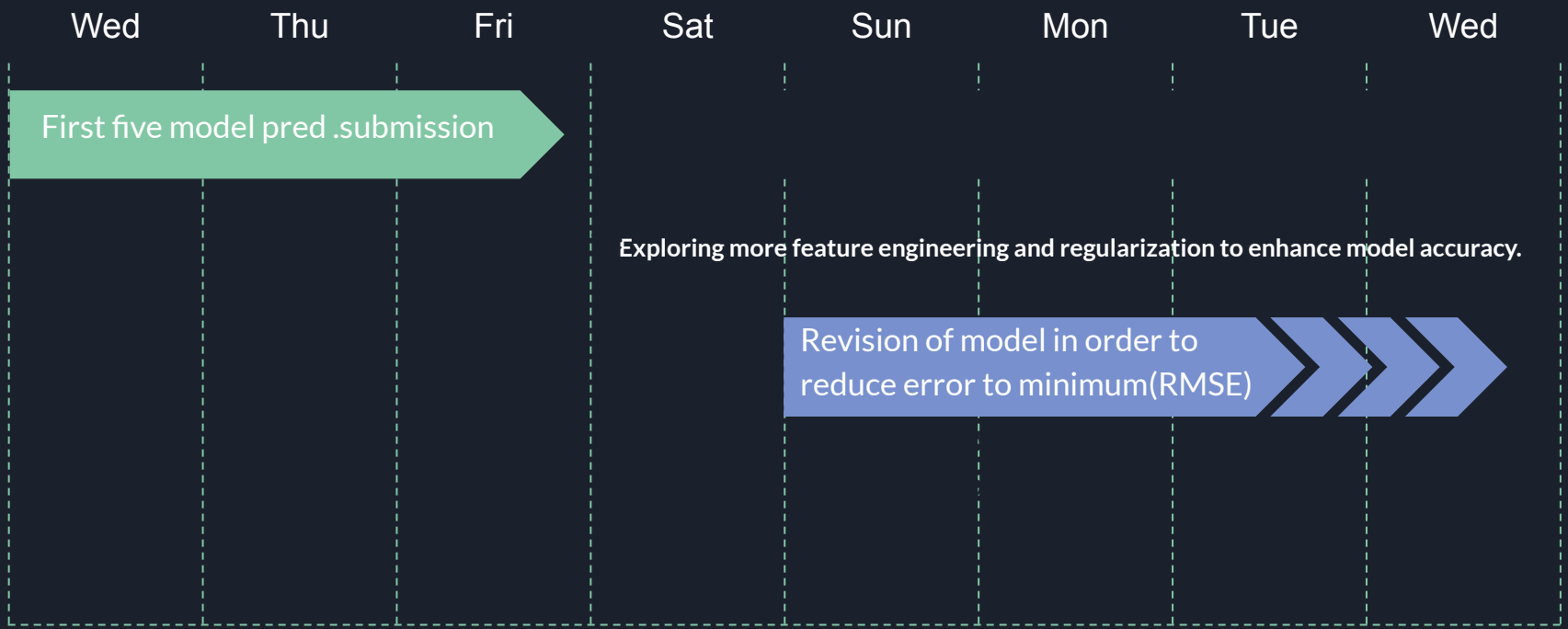
Living Area vs. Sale Price Regression: A strong positive correlation exists, meaning larger homes generally sell for higher prices, with some variance at extreme sizes.

Interpreting Linear Regression Coefficients:

R2 score = (R^2 Score: 0.7935 (79.35%))

Overall Quality	• A one-unit increase in quality raises the house price by $27,392, making it the most significant factor.
Above Ground Living Area	• Each additional square foot adds $48.59 to the price, though less impactful than quality.
Garage Capacity	• Each extra car space increases the price by $7,477.
Garage Area	• Each square foot of garage space adds $55.48, with a lower impact than garage capacity.

Next steps


Wed
Thu
Fri
Sat
Sun
Mon
Tue
Wed
First five model pred .submission
Exploring more feature engineering and regularization to enhance model accuracy.
Revision of model in order to reduce error to minimum(RMSE)

Thank you!