winnycodegurl / dsc-phase-2-projectgroup4

Code Pull requests Actions Projects Wiki Security Insights Settings

View license

0 stars 643 forks 0 watching 6 Branches 0 Tags Activity

Public repository · Forked from learn-co-curriculum/dsc-phase-2-project-v2-3

main 6 Branches 0 Tags Go to file t Go to file Add file Code

This branch is 35 commits ahead of learn-co-curriculum/dsc-phase-2-project-v2-3:main . Contribute Sync fork

winnycodegurl Add files via upload 1 minute ago

.vscode	add images	4 hours ago
data	revisions for v2.3	3 years ago
images	update image	3 hours ago
.canvas	sync .canvas	3 years ago
.gitignore	Initial commit	4 years ago
CONTRIBUTING.md	Initial commit	4 years ago
Group 4 Power Point Presentation...	Add files via upload	32 minutes ago
LICENSE.md	Initial commit	4 years ago
README.md	Merge branch 'main' of github.com:winn...	8 minutes ago
halfway-there.gif	Initial commit	4 years ago
student.ipynb	readme changes	5 hours ago

README License

REAL ESTATE SALES PREDICTION MODEL

Project Overview

In the fast-paced world of real estate, it's crucial for agencies to provide clients with precise information. Clients, whether they're looking to become homeowners or investors, rely on real estate companies for guidance on important decisions such as pricing, market trends, and property evaluations. To meet this need, real estate agencies can benefit from a sophisticated regression-based tool. This tool uses various property variables like the number of bedrooms, year built, floor count, living area, condition, location, and amenities to accurately predict property prices. By employing regression analysis, agencies can offer clients more precise pricing estimates, leading to better-informed decisions. Ultimately, this tool aims to improve client satisfaction, streamline decision-making processes, and drive success for real estate agencies.

Business Problem

Real estate experts in King County need help understanding what factors influence property values and market trends. This study aims to analyze property features, locations, buyer preferences, and market changes over time. By gaining insights from this analysis, real estate professionals can make informed decisions about buying, selling, and positioning themselves in the dynamic King County market. The goal is to provide practical advice to help them succeed in this ever-changing real estate landscape.

The Data Understanding

King County, Washington, situated in the northwest of the United States, is known for its vibrant housing market centered around Seattle. The county has experienced significant growth due to its strong economy and cultural importance, attracting a large number of residents and creating high demand for housing in both urban and suburban areas. Seattle, with its impressive skyline, is especially sought after by tech professionals and city lovers. King County's real estate market is competitive, offering a range of neighborhoods to suit different preferences, from historic areas to modern suburban developments

Target Variable price: Sale price of the house .

Property Characteristics:

	Column_name	Description
0	* `id`	Unique identifier for a house
1	* `date`	Date house was sold
2	* `price`	Sale price (prediction target)
3	* `bedrooms`	Number of bedrooms
4	* `bathrooms`	Number of bathrooms
5	* `sqft_living`	Square footage of living space in the home
6	* `sqft_lot`	Square footage of the lot
7	* `floors`	Number of floors (levels) in house
8	* `waterfront`	Whether the house is on a waterfront
9	* `view`	Quality of view from house
10	* `condition`	How good the overall condition of the house is...
11	* `grade`	Overall grade of the house. Related to the con...
12	* `sqft_above`	Square footage of house apart from basement
13	* `sqft_basement`	Square footage of the basement
14	* `yr_built`	Year when house was built
15	* `yr_renovated`	Year when house was renovated
16	* `zipcode`	ZIP Code used by the United States Postal Service
17	* `lat`	Latitude coordinate
18	* `long`	Longitude coordinate
19	* `sqft_living15`	The square footage of interior housing living ...
20	* `sqft_lot15`	The square footage of the land lots of the nea...

Main Objective:

The primary aim of this project is to develop a predictive regression model to support real estate agencies in advising clients on house prices. This model is intended to anticipate potential changes in property value based on property characteristics, furnishing clients with valuable insights to facilitate informed investment decisions.

#**Statistical Analysis** Statistical analysis plays a crucial role in understanding relationships within datasets, identifying patterns, and gaining insights. In this regression modeling project aimed at predicting property values, several key steps in statistical analysis are essential:

<<<<<<< HEAD Descriptive Statistics Correlation matrix Distribution Analysis Inferential Statistics using Hypothesis Testing and Analysis of Variance MultiColinierity

*Modelling**

Baseline model - simple linear model.

=======

1. Descriptive Statistics
2. Correlation matrix
3. Distribution Analysis
4. Inferential Statistics using Hypothesis Testing and Analysis of Variance
5. MultiColinierity

Modelling

1. Baseline model - simple linear model.

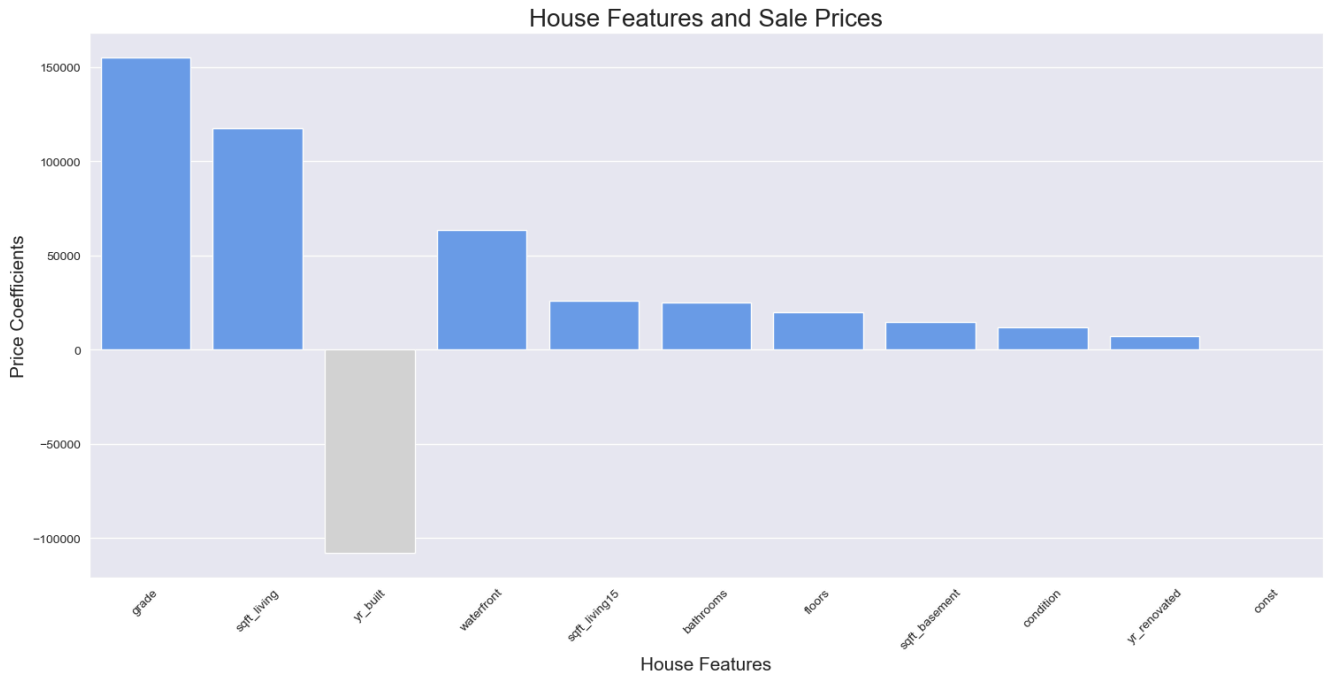
0a440037de0b20dcd0b9a308c1a7d88dff9cfac8

2. log transformation.
3. Multiple Linear Regression
4. Residual modelling.

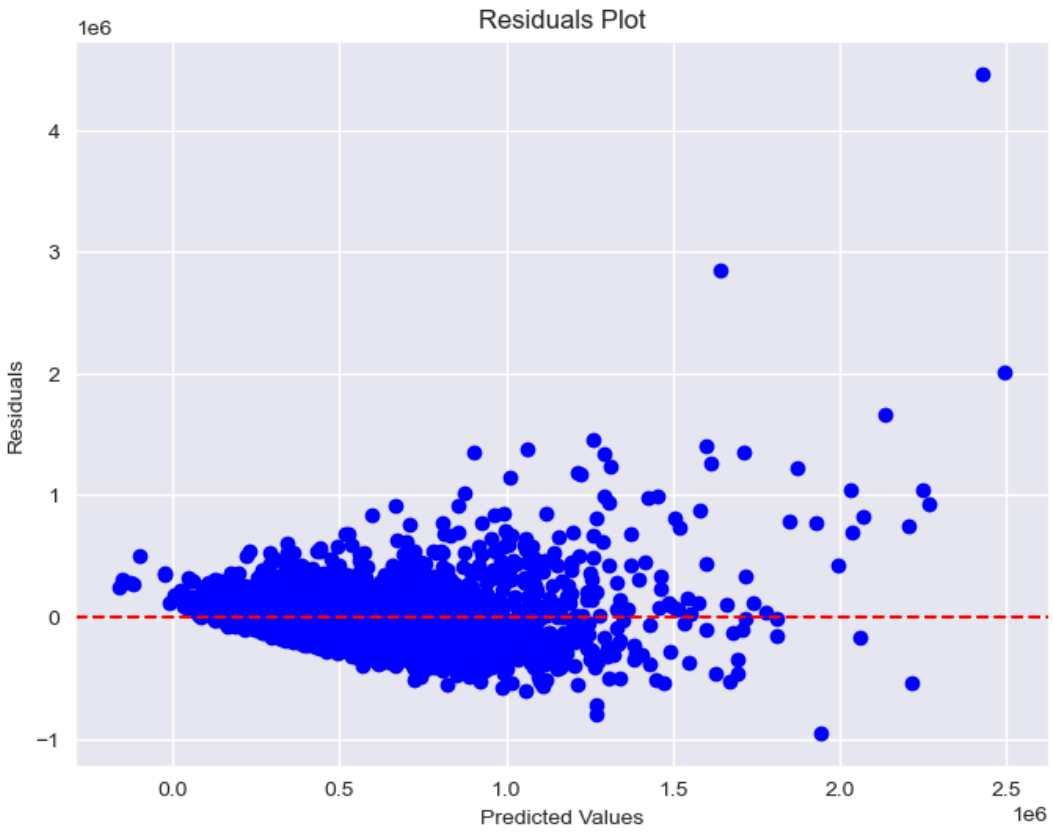
REGRESSION RESULTS

SIMPLE LINEAR REGRESSION R-squared (0.48): Indicates that approximately 48% of the variability in house prices is explained by the square footage of living space. It measures how well the model captures patterns in the data. Mean Squared Error (MSE) (68845100756.11): Represents the average squared difference between actual and predicted house prices. Lower values indicate better accuracy, but here, the MSE is quite large, suggesting room for improvement. Root Mean Squared Error (RMSE) (262383.50): This is the square root of the MSE, providing a measure of typical deviation between predicted and actual house prices. The RMSE is approximately 262,383.50 units. Intercept (540631.16): Estimated house price when all independent variables are zero. It's around 540,631.16 units, suggesting a baseline value. Coefficient (259767.82): Represents the change in house prices for a one-unit increase in square footage of living space, with other variables held constant. For every one-unit increase in square footage, house prices are expected to increase by approximately 259,767.82 units.

Multiple Linear Regresion



RESIDUALS

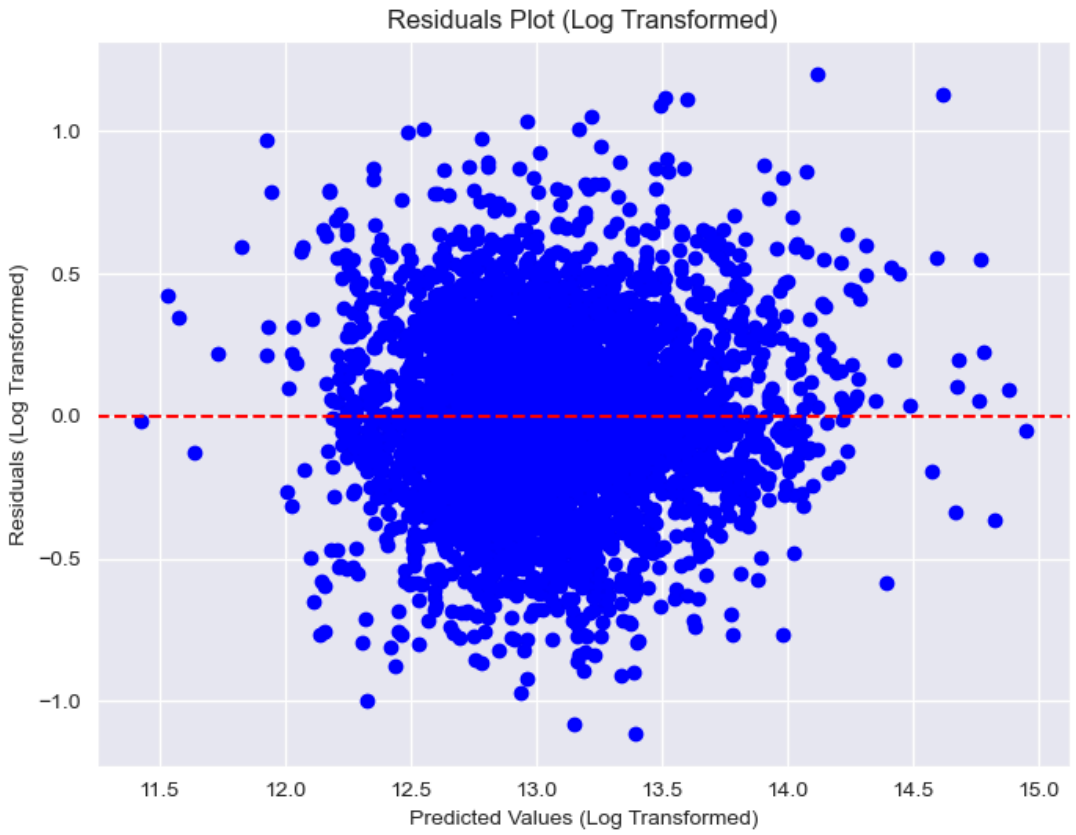


A p-value of 3.975373048166964e-258, which is extremely close to zero, indicates strong evidence against the null hypothesis of homoscedasticity. In other words, there is a significant presence of heteroscedasticity in your data. We can see the positive correlation between the house prices and all other features except the year built which indicates a negative correlation.

Log transformation.

Transforming the dependent variable or one or more independent variables can sometimes stabilize the variance. Common transformations include taking the natural logarithm, square root, or reciprocal of the variables.

Log transformation of the multiple linear regression.



REGRESSION Results

From all the models observed,We can see the positive correlation between the house prices and all other features except the year built which indicates a negative correlation.

Polynomial Regression is the preferred model beacuse from the evaluation it has the highest R-squared value of 0.73

The features below impact price such that an increase will cause an increase in the price of the property. 'bedrooms','bathrooms', 'sqft_living','floors', 'waterfront','view''condition', 'grade', 'sqft_above','sqft_basement', 'yr_renovated', 'sqft_living15','renovated', 'basement'

<<<<<<< HEAD Independence of errors: The errors (residuals) from the regression model should be independent of each other. In other words, the residual for one observation should not predict the residual for another observation.

Normality of errors: The errors are normally distributed. This implies that the residuals should follow a normal distribution with a mean of zero.

No perfect multicollinearity: There should be no perfect linear relationship between the independent variables. In other words, no independent variable should be a perfect linear combination of other independent variables.

Limitations

Despite its effectiveness in predicting property prices, the model has several limitations that need to be addressed:

Limited Property Characteristics: The dataset may lack comprehensive property-based characteristics, potentially limiting the model's ability to capture the full range of factors influencing housing prices.

Multicollinearity Concerns: The presence of correlated predictors, such as square footage and number of bedrooms, can introduce multicollinearity issues. This makes it difficult to discern the individual impact of each feature accurately, potentially affecting the model's reliability.

Assumption Violations: Polynomial regression relies on the assumption of linearity between predictors and the target variable. However, in reality, this assumption may not always hold true, leading to biased estimates and less dependable predictions.Also heteroscedasticity was present.

Overfitting Risk: Polynomial regression models, especially those with high degrees, are prone to overfitting. This occurs when the model fits the training data too closely, capturing noise rather than underlying patterns. As a result, the model may struggle to generalize well to unseen data, impacting its predictive performance.

0a440037de0b20dcd0b9a308c1a7d88dff9cfac8

Limitations

1. The dataset could have more property based characteristics
2. Multicollinearity:The presence of correlated predictors (e.g., square footage and number of bedrooms) can lead to multicollinearity issues, making it challenging to interpret the individual effects of each feature accurately
3. Assumption Violations:Polynomial regression assumes linearity between predictors and the target variable, which may not hold true in all cases. Violations of this assumption can lead to biased estimates and unreliable predictions.
4. Overfitting: Polynomial regression models, particularly those with high degrees, are susceptible to overfitting, where the model fits the training data too closely and may not generalize well to unseen data. Overall the model was the best fit model for this predictions

RECOMENDATIONS

1.Invest in Larger Properties: Investors seeking maximum returns should focus on larger houses, as there's a positive correlation between total square footage and price. Such properties have the potential for higher profits upon resale or rental.

2.Upgrade Existing Properties: Homeowners can increase their property's value by investing in upgrades that increase square footage, such as adding extra rooms or expanding living spaces.

3.Optimize Bedroom and Bathroom Ratios: It's essential to find the right balance between bedrooms and bathrooms to maximize property value. Consulting with real estate professionals can help determine the optimal ratio based on market trends and buyer preferences.

4.Focus on Quality Over Quantity: Prioritize quality improvements that enhance functionality and aesthetics, such as renovating bathrooms with modern fixtures or upgrading kitchen appliances, to add perceived value to the property.

5.Highlight Features in Listings: Emphasize the number of bedrooms and bathrooms in property listings to attract buyers who prioritize space and convenience. Highlight unique features that add versatility to the property.

6.Differentiate Marketing Strategies: Tailor marketing strategies based on property condition and grade ratings. Highlight the benefits of higher-grade properties to attract premium buyers, while emphasizing renovation potential for

Releases

No releases published
Create a new release

Packages

No packages published
Publish your first package

Languages

- **Jupyter Notebook** 100.0%