TASK

Capstone Project – Linear Regression in Action

Visit our website

Introduction

Get ready for an engaging capstone project! In this project, you'll utilise the `ames.csv` dataset to predict sale prices for houses in Ames, Iowa using multiple linear regression. Gain valuable insights into the factors that influence house prices and make informed recommendations to optimise sales strategies in the residential property market.

THE TASK AT HAND

The `ames.csv` dataset provides valuable insights into the housing market in Ames by recording various features of houses, including the size of the living area of a house, the year the house was built, and the sale price of the house. The main objective of this task is to utilise this dataset to predict the sales prices of houses based on the available features. By employing a multiple linear regression model, we aim to uncover the underlying relationships between the independent variables and the dependent variable. Through further exploration and analysis, you will gain valuable insights into the factors influencing sale prices within the housing market. By developing an accurate predictive property value model, real estate agents and others can make informed decisions regarding listing prices, property taxes, and property investment. This task will challenge you to extract meaningful insights and provide recommendations that align with the specific needs of the property industry.

RECAP

When predicting the sale prices of houses, a few essential steps need attention. First, preparing the data set is crucial to avoid a biased and inaccurate analysis. Second, accurately splitting the dataset into appropriate independent and dependent variables ensures relevant model training. Finally, to create an error plot, start by obtaining the predicted values and the corresponding actual values from your model's predictions. Then, utilise a plotting library such as **matplotlib** or **seaborn** to generate a scatter plot.

Practical task

Follow these steps:

1. Read the `ames.csv` file into the Jupyter notebook, **Linear_Regression_Ames.ipynb**.
2. Clean and prepare the dataset if necessary.
3. Perform exploratory data analysis to gain insights into the dataset by visualising the distributions of the dependent variable and independent variables, and identifying any patterns or trends in the data.
4. For the linear regression model, use the following independent variables:
    - **Gr_Liv_Area:** size of above grade, ground living area in square feet.
    - **Garage_Area:** size of garage in square feet.
5. Split the dataset into the independent variables and the single dependent variable.
6. Generate plots to explore the relationships between the independent variables and the dependent variable.
7. Split the data into training and test sets using a split ratio of 75:25.
8. Build a multiple linear regression model using the training set with the selected independent variables 'Gr_Liv_Area' and 'Garage_Area'.
9. Print out the intercept and coefficients of the trained model.
10. Generate predictions for the test set.
11. Evaluate the model's performance by computing the mean squared error (MSE) or root mean squared error (RMSE) on the test set using **`sklearn.metrics`**.
12. Generate an error plot to visualise the differences between the predicted and actual values in the test set.
13. Print the coefficients and interpret them within the context of the median value prediction.

14. Summarise the findings from the analysis, including insights from the exploratory data analysis, model performance, and any notable observations within the notebook.

Optional Challenge

Now that you've built a linear regression model using just two independent variables (`Gr_Liv_Area` and `Garage_Area`), try improving the model by adding more predictors from the dataset.

Follow these steps:

1. Explore the dataset further and identify other features that might influence house prices. Consider variables such as:
    - `Overall_Qual` (overall material and finish quality)
    - `Total_Bsmt_SF` (total basement area)
    - `Year_Built` (original construction date)
    - `Full_Bath` (number of full bathrooms)
    - `Fireplaces` (number of fireplaces)
2. Update your feature set by adding at least 2–3 additional variables that seem relevant.
3. Refit your linear regression model using the extended set of features.
4. Evaluate and compare model performance:
    - Use R^2 and RMSE scores.
    - Generate a residual plot to check for improved prediction patterns.

5. Reflect: Does the model perform better with the additional features? What changes do you observe in the residual distribution or accuracy metrics?

1.

Important: Be sure to upload all files required for the task submission inside your task folder and then click "Request review" on your dashboard.

Share your thoughts

Please take some time to complete this short feedback **form** to help us ensure we provide you with the best possible learning experience.