Hands-on Lab: Final Project: Generative AI for Data Science

Estimated Effort: 60 mins

Project Scenario

You have been employed as a Data Scientist by a consultancy firm. The firm has a client who is a used car dealer. They have a special feature on Ford cars and they want your firm to design a model that can predict the optimum quotation price for the cars in their lot. They provide you with sales data for the past few years. The dataset contains different features of the cars and the price they were sold at.

The tasks assigned to you are as follows.

1. There might be a few duplicate entries and a few missing values in the dataset. Data cleaning will be a part of the assignment.
2. You have to perform exploratory data analysis to draw keen insights on the data and determine the effect of different features on the price. Some specific requests by the client include:
   a. Identify number of sales for each fuel type
   b. Identify which transmission type has more price outliers
3. Compare the models with linear, polynomial and ridge regressions on single and multiple variables to find the best performing model
4. Perform a Grid Search on the Ridge regression model to identify the optimum hyperparameter for the model for best performance.

You decide to use Generative AI to create python codes that can help you analyse the data, determine the best features and create the prediction model as per requirement.

> Disclaimer: This is a fictitious scenario created for the purpose of this project. The dataset being used is publicly available.

About the Dataset

This dataset contains used car sale prices for Ford cars. This is a public dataset available on the Kaggle website as Ford Car Pricing Dataset under the CC0: Public Domain license. The dataset has been slightly modified for the purpose of this project.

Attributes of this dataset have been explained below.

Variable	Description
model	Car model name
year	Year of car make
transmission	Type of transmission (Automatic, Manual or Semi-Auto)
mileage	Number of miles traveled
fuelType	The type of fuel the car uses (Petrol, Diesel, Hybrid, Electric, Other)
tax	Annual Tax payable in USD
mpg	Miles per Gallon that the car runs at
engineSize	Engine Size of the car
price	Price of car in USD

Code execution environment

To test the prompt-generated code, keep the Jupyter Notebook (in the link below) open in a separate tab in your web browser. The notebook has some setup instructions that you should complete now.

Jupyter-Lite Test Environment

> Please note the lab environment above will only work on Windows (Google Chrome or Firefox browser). If you don't have a Windows system with either of these browsers, use the lab environment provided in the next lesson of the module.

The data set for this lab is available in the following URL.

```
URL = "https://cf-courses-data.s3.us.cloud-object-storage.appdomain.cloud/IBMSkillsNetwork-AI0271EN-SkillsNetwork/labs/v1/m3/data/used_car_price
```

Complete the setup in the Jupyter Notebook and then proceed further.

> Important Note: All prompts that are made available have been hidden and the users are encouraged to first try to write their own prompts to create the solutions. Also, the prompts given as solutions have also been maintained as ones which will create generic code structures which you can modify

according to the question at hand.

Importing the Dataset

You can begin by using the Generative AI model to create a python script that can load the dataset to a pandas dataframe. The dataset file already has the headers in the first row.

Write the prompt to generate the said code and test it in the JupyterLite environment. For verification of appropriate loading, include a step for printing the first 5 values of the loaded dataframe.

▶ Click here for the prompt

Data Preparation

Data Cleaning

At this stage, it is required to clean up the data. As has been informed to you, the data may have missing values and duplicate entries. Write a prompt that performs the following tasks

1. Identifies the columns with missing values and fills the blank cells with average value of the columns.
2. Identifies and drops the duplicate entries from the data.

▶ Click here for prompt

Data Augmentation (optional)

Once cleaned, you may choose to augment this dataset with additional samples, created synthetically using Mostly.ai.

Data Insights and Visualization

Write prompts that generate codes to prform the following actions.

1. Identify the 5 attributes that have the highest correlation with the `price` parameter.

▶ Click here for the prompt

2. Count the number of cars under each unique value of `fuelType` attribute.

▶ Click here for the prompt

3. Create a Box plot to determine whether cars with automatic, manual or semi-auto type of `transmission` have more price outliers. Use the Seaborn library for creating the plot.

▶ Click here for the prompt

4. Generate the regression plot between `mpg` parameter and the `price` to determine the correlation type between the two.

▶ Click here for the prompt

Model Development and Evaluation

Write prompts that generate codes to perform the following actions.

1. Fit a linear regression model to predict the `price` using the feature `mpg`. Then calculate the R^2 and MSE values for the model.

▶ Click here for the prompt

2. Fit a linear regression model to predict the `price` using the following set of features.
   `year`, `mileage`, `tax`, `mpg` and `engineSize`.
   Calculate the R^2 and MSE values for this model.

▶ Click here for the prompt

3. For the same set of features as in the question above, create a pipeline model object that uses standard scalar, second degree polynomial features and a linear regression model. Calculate the R^2 value and the MSE value for this model.

▶ Click here for the prompt

4. For the same set of features, split the data into training and testing data parts. Assume testing part to be 20%. Create and fit a Ridge regression object using the training data, set the regularization parameter to 0.1, and calculate the R^2 using the test data.

▶ Click here for the prompt

5. Perform a second order polynomial transform on both the training data and testing data created for the question above. Create and fit a Ridge regression object using the modified training data, set the regularisation parameter to 0.1, and calculate the R^2 and MSE utilising the modified test data.

▶ Click here for the prompt

6. In the question above, perform a Grid Search on ridge regression for a set of values of alpha {0.01, 0.1, 1, 10, 100} with 4-fold cross validation to find the optimum value of alpha to be used for the prediction model.

▶ Click here for the prompt

Conclusion

Congratulations! You have completed this guided project on using Generative AI for different data science tasks.

By the end of this project, you are now capable of using Generative AI for the tasks of:

- **Data preparation**: cleaning, transforming and augmentation
- **Data analysis**: drawing insight, creating visualizations
- **Model development**: creating simple as well as complex prediction models
- **Model refinement**: found the optimum model using Grid Search

Author(s)

Abhishek Gagneja