AT - Lesson 69 - Project_Question Copy

March 21, 2023

0.0.1 Instructions

Goal of the Project This project is designed for you to practice and solve the activities that are based on the concepts covered in the following lessons:

1. Multiple linear regression - Introduction
2. Car Prediction - Feature Encoding

Getting Started:

1. Click on START Project on the panel and follow the instructions given below.n
2. Create a duplicate copy of the Colab file as described below.

- Click on the **File menu**. A new drop-down list will appear.
- Click on the **Save a copy in Drive** option. A duplicate copy will get created. It will open up in the new tab on your web browser.

3. After creating the duplicate copy of the notebook, please rename it in the **YYYY-MM-DD_StudentName_Project69** format.
4. Now, write your code in the prescribed code cells.

0.0.2 Problem Statement

Implement multiple linear regression to create a predictive model capable of predicting the price of diamonds on the basis of various factors such as its cut, color, clarity, depth etc.

0.0.3 List of Activities

Activity 1: Analysing the Dataset

Activity 2: Data Preparation

Activity 3: Feature Encoding

Activity 4: Train-Test Split

Activity 5: Model Training using `statsmodels.api`

Activity 1: Analysing the Dataset

- Create a Pandas DataFrame for **Diamonds** dataset using the below link. This dataset consists of following columns:

Field	Description
carat	weight of the diamond
cut	quality of the cut
color	diamond colour, from J (worst) to D (best)
clarity	a measurement of how clear the diamond is (I1 (worst), SI2, SI1, VS2, VS1, VVS2, VVS1, IF (best))
depth	total depth percentage = z / mean(x, y) = 2 * z / (x + y)
table	The width of the diamond's table expressed as a percentage of its average diameter
price	price in US dollars
x	length in mm
y	width in mm
z	depth in mm

Dataset Link: https://student-datasets-bucket.s3.ap-south-1.amazonaws.com/whitehat-ds-datasets/diamonds.csv

- Print the first five rows of the dataset. Check for null values and treat them accordingly.
- Remove the unnecessary column `Unnamed: 0` as it is of no use.

[ ]:
```
# Import modules

# Load the dataset

# Print first five rows using head() function
```

[ ]:
```
# Check if there are any null values. If any column has null values, treat them accordingly
```

[ ]:
```
# Print the information about the dataset
```

[ ]:
```
# Drop 'Unnamed: 0' column
```

Activity 2: Data Preparation Extract numerical attributes from the dataset and create a heatmap to identify correlation among various numerical attributes.

[ ]:
```
# Extract all the numeric (float and int type) columns from the dataset.
```

[ ]:
```
# Draw a correlation heatmap between the numeric features.
```

Q: Which features are highly correlated with `price`?

A:

Q: Is there multicollinearity in the dataset?

A:

Activity 3: Feature Encoding The dataset contains certain columns that are categorical. However for linear regression, we need all numerical variables. Perform **one-hot encoding** to obtain numeric values from non-numeric categorical values.

[ ]:
```
# Create a new dataframe having dummy variables for all the categorical type columns of the dataset using 'get_dummies()' function.
```

[ ]:
```
# Print the information of the new dataframe obtained after one-hot encoding
```

Activity 4: Train-Test Split We need to predict the value of `price` variable, using other variables. Thus, `price` is the target or dependent variable and other columns except `price` are the features or the independent variables.

Split the dataset into training set and test set such that the training set contains 70% of the instances and the remaining instances will become the test set and keep the `price` as the target variables.

[ ]:
```
# Split the 'df' Dataframe into the train and test sets.
```

[ ]:
```
# Create separate data-frames for the feature and target variables for both the train and test sets.
```

Activity 5: Model Training using `statsmodels.api` Now build a multiple linear regression model using the `statsmodels.api` module. Also, print the summary of the linear regression model built.

[ ]:
```
#  Build a linear regression model using all the features to predict insurance charges.
```

[ ]:
```
# Print the summary of the linear regression report.
```

Q: What is the Adjusted R^2 value?

A:

0.0.4 Submitting the Project:

1. After finishing the project, click on the **Share** button on the top right corner of the notebook. A new dialog box will appear.
2. In the dialog box, make sure that '**Anyone on the Internet with this link can view**' option is selected and then click on the **Copy link** button.
3. The link of the duplicate copy (named as **YYYY-MM-DD_StudentName_Project69**) of the notebook will get copied
4. Go to your dashboard and click on the **My Projects** option.
5. Click on the **View Project** button for the project you want to submit.
6. Click on the **Submit Project Here** button.
7. Paste the link to the project file named as **YYYY-MM-DD_StudentName_Project69** in the URL box and then click on the **Submit** button.
