AT - Lesson 68 - 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 - Data exploration

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_Project68** format.
4. Now, write your code in the prescribed code cells.

0.0.2 Problem Statement

The most important factor for an Insurance Company is to determine what premium charges must be paid by an individual. The charges depend on various factors like age, gender, income, etc.

Build a model that is capable of predicting the insurance charges a person has to pay depending on the given features using multiple linear regression.

0.0.3 List of Activities

Activity 1: Analysing the Dataset

Activity 2: Feature Encoding

Activity 3: Exploratory Data Analysis

Activity 4: Train-Test Split

Activity 5: Model Training using `statsmodels.api`

Activity 1: Analysing the Dataset

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

Field	Description
age	Age of primary beneficiary
sex	Insurance contractor gender, female or male
bmi	Body mass index
children	Number of children covered by health insurance/number of dependents
region	Beneficiary's residential area in the US, northeast, southeast, southwest, northwest
charges	Individual medical costs billed by health insurance

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

- Print the first five rows of the dataset. Check for null values and treat them accordingly.
- Create a regression plot with `age` on X-axis and `charges` on Y-axis to identify the relationship between these two attributes.

[ ]:
```
# 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
```

Activity 2 : Feature Encoding The `sex` and `region` columns are categorical attributes. Convert these attributes into numerical ones so that they can be used for linear regression analysis using `map()` function.

- Map the following values for the `sex` column:

  - `male` to 0
  - `female` to 1

- Map the following values for the `region` column:
- `southeast` to 1
- `southwest` to 2
- `northeast` to 3
- `northwest` to 4

[ ]:
```
# Count the occurrence of each value in the 'sex' column.
```

[ ]:
```
# Use the 'map()' function to replace values in 'sex' column to their␣
↪corresponding numeric values.
```

[ ]:
```
# Again count the occurrence of each value in the 'sex' column to verify␣
↪whether all values are correctly mapped
```

[ ]:
```
# Count the occurrence of each value in the 'region' column.
```

[ ]:
```
# Use the 'map()' function to replace a value in the 'region' column to their␣
↪corresponding numeric values.
```

[ ]:
```
# Again count the occurrence of each value in the 'region' column to verify␣
↪whether all values are correctly mapped
```

Activity 3: Exploratory Data Analysis Create the heat-map to look into the correlation of the features

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

Activity 4: Train-Test Split Split the dataset into training set and test set such that the training set contains 67% of the instances and the remaining instances will become the test set and keep the `charges` 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 R^2 value?

A:

Q: Is there multicollinearity in the model?

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_Project68**) 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_Project68** in the URL box and then click on the **Submit** button.
