PYTHON MINOR PROJECT - 02

A Minor Project Report

Submitted to the School of Management- IIT Mandi

in partial fulfillment of requirements of the 1st sem Assignment work for

MBA course

in

Data Science and AI

by

Priya Rani (MB23041)

Adith Sagar (MB23013)

Submitted to: Dr. Manoj Thakur

SCHOOL OF MANAGEMENT

INDIAN INSTITUTE OF TECHNOLOGY, MANDI

SALES PREDICTION USING LINEAR REGRESSION

1. OBJECTIVE

The objective of the provided Python code is to perform a linear regression analysis on an advertising dataset to understand and predict the impact of advertising expenditures (TV, Radio, and Newspaper) on product sales.

The dataset used in this report is: Click here

2. TOOLS AND LIBRARIES USED

- **Pandas:** It's used for data manipulation, reading the dataset, and performing initial data exploration and preparation.
- **NumPy:** It is used for handling arrays and numerical operations required for calculations in data preparation and model evaluation.
- **Seaborn and Matplotlib:** These libraries are used for data visualization to understand data distribution and relationships.
- **Scikit-learn (sklearn):**It is used for machine learning algorithms, metrics, and model selection.
- **StatsModels:** Employed for statistical modeling, in this case, for fitting an Ordinary Least Squares (OLS) regression model.

3. BACKGROUND

1. Sales prediction means predicting how much of a product people will buy based on factor such as the amount you spend to advertise your product, the segment of people you advertise for, or the platform you are advertising on about your product.
2. Typically, a product and service-based business always need their Data Scientist to predict their future sales with every step they take to manipulate the cost of advertising their product.

4. PYTHON PROGRAM

- **STEP 1: Installing and Importing Necessary Libraries:** (i.e. pandas, numpy, seaborn, matplotlib, scikit-learn, and statsmodelshere.)

[1]:
```
pip install pandas numpy seaborn matplotlib scikit-learn statsmodels
```

[1]:
[2]:
```
import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt
from sklearn.linear_model import LinearRegression, Ridge, Lasso,␣
↪ElasticNet
from sklearn.metrics import mean_squared_error, r2_score
from sklearn.model_selection import train_test_split
from sklearn.model_selection import cross_val_score
import statsmodels.api as sm
```

- **STEP 2: Loading the Dataset:** This step involve loading the dataset into the Python environment using the Pandas library.

[5]:
```
dataset = pd.read_csv("C:\\Users\\hp\Desktop\\Advertising.csv")
```

- **STEP 3: Data Inspection and Preprocessing:** In this we renamed the column 'Unnamed: 0' to 'Index' for better understanding.

[13]:
```
# Rename the column 'Unnamed: 0' to 'Index'
dataset.rename(columns={'Unnamed: 0': 'Index'}, inplace=True)
```

- **STEP 4: Preparing the Data for Model Building:** In this the dataset was divided into predictor variables (x: TV, Radio, Newspaper) and the target variable (y: Sales). A train-test split was performed to separate the data for model training and evaluation.

[12]:
```
#Preparing model
x = dataset.drop('Sales', axis=1)
y = dataset[["Sales"]]
```

```
x_train, x_test, y_train, y_test = train_test_split(x, y, test_size=0.
↪20, random_state=46)
```

- **STEP 4: Fitting the Linear Regression Model:** In this a linear regression model was fitted using the **StatsModels** library to explore the relationship between advertising expenditures and sales.

[14]:
```
# Linear Regression Model
lm = sm.OLS.from_formula("Sales ~ TV + Radio + Newspaper",␣
↪data=dataset).fit()
```

- **STEP 5: Model Evaluation and Analysis:** Through this step we extracted fitted regression model's summary and coefficients to understand the influence of each advertising medium on sales.

[15]:
```
# Print the summary of the regression model
print(lm.summary())
# Print the coefficients of the model
print(lm.params, "\n")
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                  Sales   R-squared:                       0.897
Model:                            OLS   Adj. R-squared:                  0.896
Method:                 Least Squares   F-statistic:                     570.3
Date:                Mon, 25 Dec 2023   Prob (F-statistic):           1.58e-96
Time:                        12:33:45   Log-Likelihood:                -386.18
No. Observations:                 200   AIC:                             780.4
Df Residuals:                     196   BIC:                             793.6
Df Model:                           3
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept      2.9389      0.312      9.422      0.000       2.324       3.554
TV             0.0458      0.001     32.809      0.000       0.043       0.049
Radio          0.1885      0.009     21.893      0.000       0.172       0.206
Newspaper     -0.0010      0.006     -0.177      0.860      -0.013       0.011
```

```
==============================================================================
Omnibus:                       60.414   Durbin-Watson:                   2.084
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              151.241
Skew:                          -1.327   Prob(JB):                     1.44e-33
Kurtosis:                       6.332   Cond. No.                         454.
==============================================================================
Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly
specified.
Intercept    2.938889
TV           0.045765
Radio        0.188530
Newspaper   -0.001037
dtype: float64
```

RESULTS EVALUATION

1. **Model Fit:** The model explains approximately 89.7% (R-squared: 0.897) of the variation in sales.
2. **Overall Significance:** The regression model is statistically significant in predicting sales(i.e. because of a high F-statistic (570.3) and a very low p-value (1.58e-96)).
3. **Impact of Advertising Mediums:** TV and Radio advertising show positive impacts on sales, with TV having a smaller impact (coefficient: 0.0458) compared to Radio . Newspaper advertising shows a negligible negative impact (coefficient: -0.0010), and its influence on sales might not be statistically significant (high p-value: 0.860).

- **STEP 6: Model Performance Evaluation:** It is done by fitting the model to the training data, making predictions on the test data, and by calculating the Root Mean Squared Error (RMSE).

[18]:
```
# Evaluating the model
results = []
names = []
models = [('LinearRegression', LinearRegression())]
# Loop through each model, fit it to the data, and calculate the Root␣
 ↪mean squared error
```

```
for name, model in models:
    model.fit(x_train, y_train)
    y_pred = model.predict(x_test)
    result = np.sqrt(mean_squared_error(y_test, y_pred))
    results.append(result)
    names.append(name)
    msg = "%s: %f" % (name, result)
    print(msg)
```

```
LinearRegression: 1.703648
```

- **STEP 7: Making Predictions on New Data** In this we utilized the fitted regression model to make predictions on new data containing advertising spending information to forecast sales.

[19]:
```
# Make predictions on new data
new_data = pd.DataFrame({'TV': [110], 'Radio': [60], 'Newspaper': [20]})
predicted_sales = lm.predict(new_data)
print("Predicted Sales:", predicted_sales)
```

```
Predicted Sales: 0    19.264052
dtype: float64
```

RESULTS EVALUATION

1. **Model Performance:** The RMSE value came approximately 1.70 indicates that on average, the model's predictions are approximately 1.70 units away from the actual sales values.
2. **Predicted Sales on New Data:** The model estimated that sales are expected to be approximately 19.26 units(based on the specified advertising expenditures).