A picture containing shape

Description automatically generated

**Module 4 Assignment**

Parva Pareshbhai Patel (002195186)

CPS Analytics, Northeastern University

ALY6110 : Data Management and Big Data

Dr. Mohammad Shafiqul Islam

December 10, 2022

**Introduction**

* One of the well-known predictive modelling techniques is the linear regression model. The project is based on the building the linear regression model.
* Ecommerce customers dataset that has the variables such as the time on app, time on website, length of membership, Avg. session length and yearly amount spent. Using the linear regression model yearly time spent is predicted.
* Time on app and time on website is the time spent by the customer on app and website. Whereas, session length is the amount of time on app or website in one go and membership is the subscription period.

**Analysis**

* Initially, before starting any analysis on the data it is necessary to look for the missing values in the dataset. After running the code shown in the figure 1 it is confirmed that there are no null or missing values in the dataset.

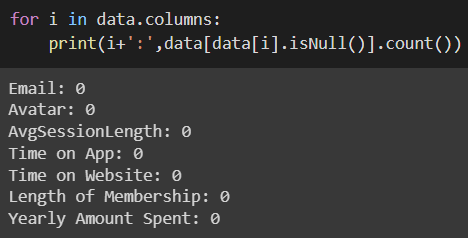
****

Figure 1 : Code to check null values

* Figure 2 is the bar plot of the length of membership. From the bar plot it can be seen that the majority of the customer prefer to take the subscription of around 3 to 4 months. Whereas, there are very less customers who take the membership of more than 6 to 7 months.

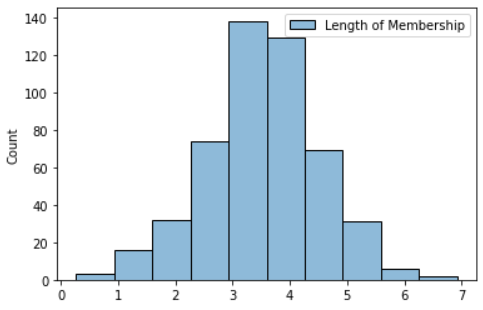


Figure 2 : Bar plot of Length of membership

* The bar plot of avg. session length is shown in Figure 3. It can be concluded that maximum number of customers have avg. session length between the 30 to 35 minutes.

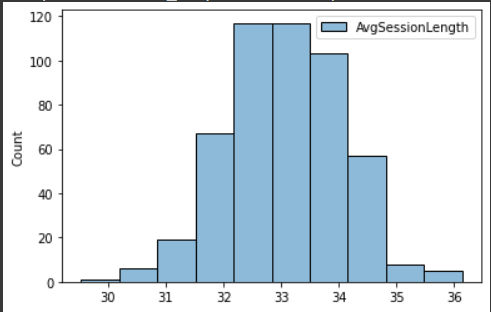


Figure 3 : Bar plot of Avg. session length.

* The scatter plot of membership length vs yearly spending is shown in figure 4. Since the scatter plot's trend is upward and the length of membership is plainly increasing, it can be concluded that membership costs are likewise rising.

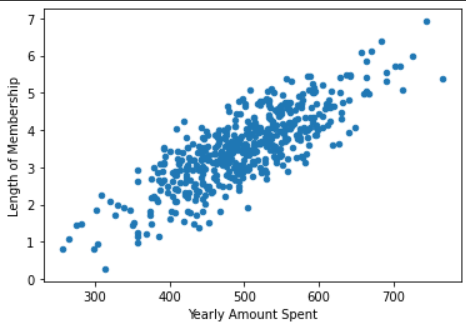


Figure 4 : Length of membership vs yearly amount spent

**Linear Regression Model**

* To built the regression model the variables used were avg. session length, time on app, time on web and length of membership. This all variables were store in one variable called as features. This features variable is then combined with the original dataset and new data frame is created called data1.

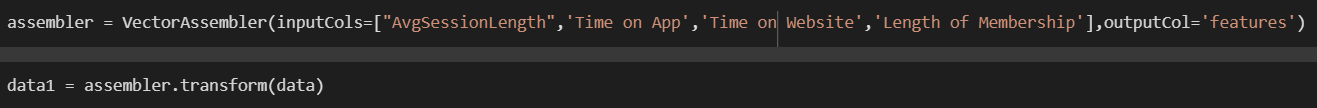


Figure 5 : Creating features variable.

* Following the joining of the feature and the original data, the yearly spending amount and the feature variable are extracted, and a new data frame that will be used to develop the regression model is produced using them.



Figure 6 : Creating data frame to use in regression model.

* The data is spitted into the train and test dataset in 70:30 ratio.



Figure 7 : Splitting data in train and test.

* To build the linear regression model the linear regression function from the pyspark.ml.regression package is used. The model is trained using the training data.

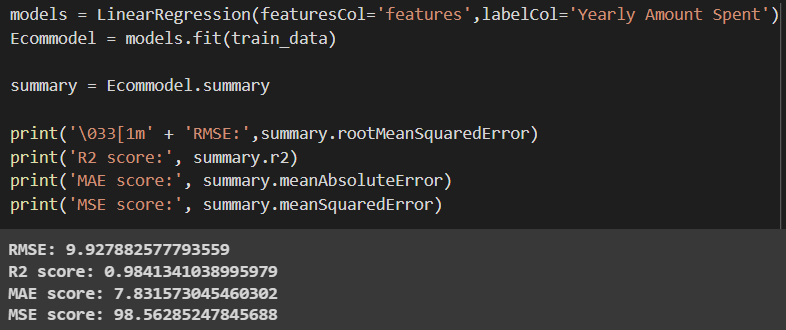


Figure 8 : Linear Regression model

* After the model had been created, its accuracy score was calculated by adding the R2 score and RMSE value. The model's R2 score is 0.984, which indicates that it is 98% accurate. Whereas, the RMSE value is 9.92.

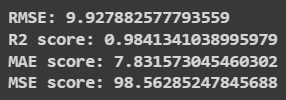


Figure 9 : R2 score and RMSE

* After the prediction has been made using the model, the results are compared to the starting values. Figure 10 illustrates how closely the anticipated value resembles the original number.

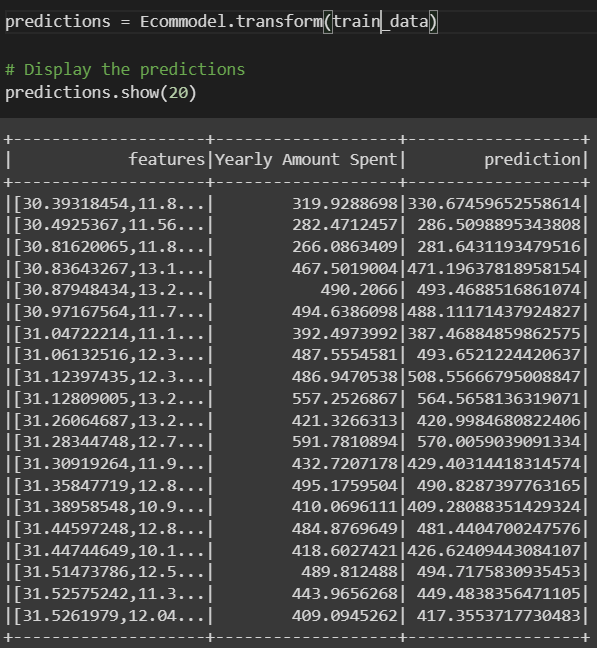


Figure 10 : Comparing original and predicted value.

* The model is now tested using the test data.



Figure 11 : Testing the model

* To calculate the accuracy of the model when tested on test data the R2 score and RMSE is calculated. The R2 score of model on testing data came 0.983 and RMSE value is 10.40.

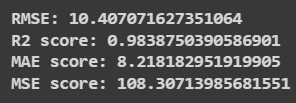


Figure 12 : R2 score and RMSE value of test data

**Conclusion**

* The yearly amount spent was forecast using the linear regression model. Additionally, the model's accuracy was 98%. The model may be trusted for predictions because of its excellent accuracy.
* Therefore, assuming the model's input variables are accessible, it is possible to determine the yearly amount that a consumer spends using that website or e-commerce app. The results of this model can be used by the business owner to get assist in future business decisions.

**Reference**

1. ASKARI. (n.d.). Ecommerce customers-LinearRegression. Kaggle. Retrieved December 8, 2022, from <https://www.kaggle.com/code/pouyaaskari/ecommerce-customers-linearregression/notebook>
2. N. (2022, June 22). Create a DataFrame From Vectors in R. Spark by {Examples}. Retrieved December 8, 2022, from <https://sparkbyexamples.com/r-programming/create-dataframe-from-vectors-in-r/>
3. Scatter Plot. (n.d.). spark.apache.org. Retrieved December 8, 2022, from <https://spark.apache.org/docs/3.2.1/api/python/reference/pyspark.pandas/api/pyspark.pandas.DataFrame.plot.scatter.html>
4. seaborn.barplot. (n.d.). seaborn.pydata.org. Retrieved December 8, 2022, from <https://seaborn.pydata.org/generated/seaborn.barplot.html>