Linear Regression Analysis of an E-Commerce Company's Customer Sales Data

597-AS Final Project
Professor Hari Balasubramanian

Group Members:
Ansel Fernandes
Joel Mathew Varghese

INTRODUCTION

An E-Commerce retail company based out of NYC has released sales data regarding its top 500 customers. The company wants to better understand the factors that influence their revenue. The company, having platforms for their customers in the form of an App as well as a Website, wants to analyze where their customers spend more time, and the use of which platform generates more revenue. It is important to note that the focus of analysis is not on what the customer purchases, but their behavior in terms of loyalty, and average time spent on each of the two platforms.

DATASET

The dataset for our project is downloaded from Kaggle, found in the following link. It includes a total of 8 columns, with 500 rows of data, each row describing the following data for each distinct customer:

Email - the email id of a customer

Address - the delivery address used by a customer

Avatar - a color assigned to a customer (not unique)

Avg. Session Length - The average time in minutes spent in store each session

Time on App - Average time spent on the App in minutes each session

Time on Website - Average time spent on the Website in minutes each session

Length of Membership - How many years the customer has been a member

Yearly Amount Spent - The average amount of money the customer spent in a year

MAIN QUESTIONS FOR OUR ANALYSIS

1. Is there a correlation between Yearly Amount Spent with any of the predictor variables? If yes, which predictor variable or group of predictors form the best model with Yearly Expenditure as our response?
2. Following question 1, how well can we predict the Yearly Expenditure with our best set of predictors?
3. Should the company focus more on their website, or more on their app?

METHODOLOGY

We conduct our analysis using Linear Regression in R-Studios, to build the best models and produce plots and statistical summaries to try and answer the above questions.

Simple intuition on taking a quick glance at the dataset shows us that for Yearly Amt Spent as the response variable, the more likely predictor variables are Avg. Session Length, Time on App, Time on Website and Length of Membership. Thus, **we don't consider Email, Address and Avatar as predictors in our analysis.**

DATA CLEANING AND PREPROCESSING

We start off by cleaning our dataset by primarily inspecting for duplicate and missing values. To do this, we use the following 2 lines of code, and it can be seen from the output displayed that our dataset is clean and contains 0 duplicate or missing values:

```
#Data Cleaning - inspect duplicate and missing values
sum(duplicated(Ecommerce_Customers))
sum(is.na(Ecommerce_Customers))
```

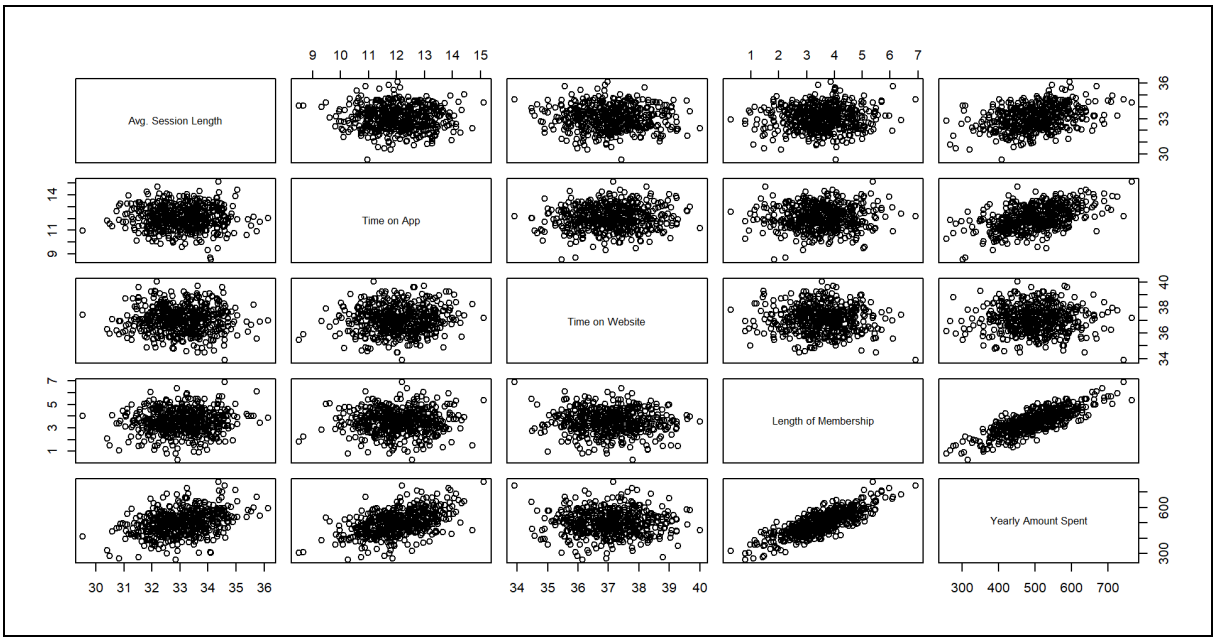
Output:

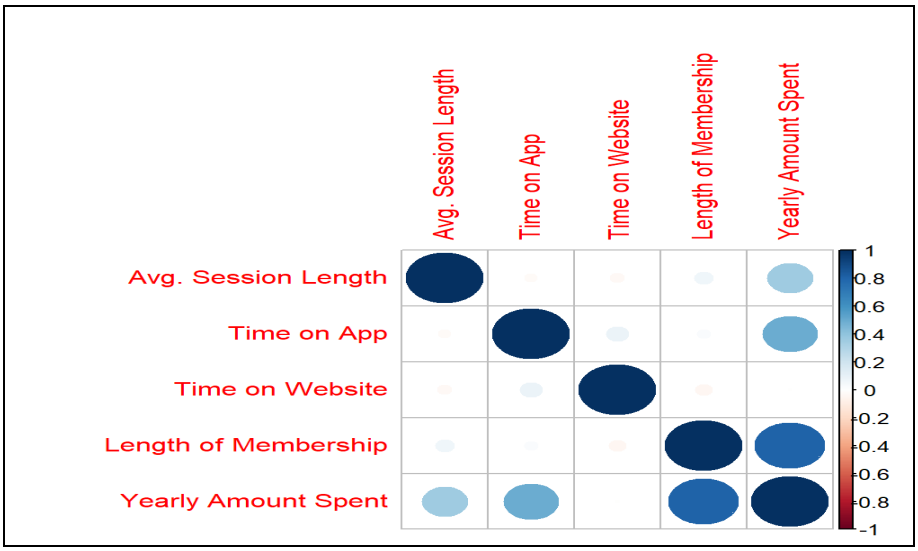
```
> sum(duplicated(Ecommerce_Customers))
[1] 0
> sum(is.na(Ecommerce_Customers))
[1] 0
```

EXPLORATORY DATA ANALYSIS

Correlation Plots

We begin our EDA by first plotting each predictor variable against each other. To do this we first use the pair function, which gives us a matrix of scatter plots, as shown below, and also the corr function, which gives us a correlation heat plot (2nd diagram):





Observations

It can be observed that there seems to be

- A relatively strong correlation between Yearly Amount Spent and the Avg. Session Length
- Stronger correlation between Yearly Amount Spent and Time on App
- And an even stronger correlation between Yearly Amount Spent and the Length of Membership.
- Poor Correlation between Yearly Amount Spent and Time on Website

VARIABLE SELECTION USING LINEAR REGRESSION MODELS

We build every possible combination using our 4 predictor variables, and run a Linear Regression Model on all of them with Yearly Amount Spent as the response. For each combination, we determine the VIF (to check for collinearity), AIC, BIC and we select the best model among these combinations on the basis of their R^2 values.

Observations:

- Every predictor in each of the models below has a Variance Inflation Factor of around 1. We have displayed the approx average VIF in the table below for each model. Since all VIFs are around 1, we can conclude that Collinearity is not a concern here.

- The combinations with the highest R^2 values are models 12 and 15. Model 12 contains all predictors except 'Time on Website', while model 15 contains all 4 predictors. Model 15 has slightly higher AIC and BIC values.
- However, in the summary output of Model 15, the 'Time on Website' predictor has a p-value of 0.35, which is greater than 0.01, meaning we cannot reject the null hypothesis for 'Time on Website'.
- Thus, we select Model 12 as the best model, with the predictors as Avg. Session Length, Time on App, and Length of Membership.

Model	Predictor Variables	R^2	VIF	AIC	BIC
1	Length of Membership	0.65	-	-	-
2	Time on Website	0.002	-	-	-
3	Time on App	0.2478	-	-	-
4	Avg Session Length	0.1243	-	-	-
5	Length of Membership + Time on Website	0.6545	1.002	5265.9	5282.8
6	Length of Membership + Time on App	0.8807	1.008	4734.4	4751.2
7	Length of Membership + Avg. Session Length	0.7478	1.003	5108.6	5125 .4
8	Avg. Session Length + Time on App	0.3831	1.000	5555.5	5572.6
9	Avg. Session Length + Time on Website	0.1227	1.001	5731.9	5748.7
10	Time on App + Time on Website	0.2482	1.006	5654.6	5671.5
11	Avg. Session Length + Time on App + Time on Website	0.3829	1.001	5556.9	5578

12	Avg. Session Length + Time on Website + Length of Membership	0.9842	1.004	3724.7	3745
13	Avg. Session Length + Time on App + Length of Membership	0.7494	1.003	5106	5127
14	Time on Website + Time on App + Length of Membership	0.8804	1.006	4736	4757
15	Yearly Amount Spent + Avg. Session Length + Time on App + Time on Website + Length of Membership	0.9842	1.01	3725	3751

TRAINING, TESTING AND PREDICTING

- 80-20 split in the dataset for Training and Testing: Having selected Model 12 as the best model, we train it using 80% of our dataset, and test this model on the remaining 20%.
- On running the following lines of code, we get an output as shown below:

```
#Can the model also be used to predict where more customers would spend most of their time on?
set.seed(10)
sample <- sample(c(TRUE,FALSE), nrow(Ecommerce_Customers), replace=TRUE, prob=c(0.8,0.2))
train <- Ecommerce_Customers[sample,]
test <- Ecommerce_Customers[!sample,]

#Training the model using our best predictors
train.EC <- lm(train$`Yearly Amount Spent`~train$`Avg. Session Length`+train$`Time on App`+train$`Length of Membership`)

#predicting on test dataset
predict.EC <-predict(train.EC, test)

#MSE and MAE
mean((predict.EC-test$`Yearly Amount Spent`)^2)
MAE(predict.EC,test$`Yearly Amount Spent`)
```

Output:

```
> #MSE and MAE
> mean((predict.EC-test$`Yearly Amount Spent`)^2)
[1] 12046.71
```

```
> MAE(predict.EC,test$`Yearly Amount Spent`)
[1] 86.58884
```

- As seen in the output above, we get a Mean Squared Error of 12046.71 and a Mean Absolute Error of 86.58.

CONCLUSION & SUMMARY

We conclude by circling back to the three main questions we set out to answer with our analysis:

1. *Is there a correlation between Yearly Amount Spent with any of the predictor variables? If yes, which predictor variable or group of predictors form the best model with Yearly Expenditure as our response?*

- As seen above in the correlation plots obtained, we can conclude that Yearly Amount Spent is correlated most strongly with Avg Session Length, Time on App and Length of Membership (in increasing order of correlation strength). As far as which combination of predictors is the best, we have seen that Model 12, with the same 3 predictors mentioned here perform the best together.

2. *Following question 1, how well can we predict the Yearly Expenditure with our best set of predictors?*

- We obtain a Mean Squared Error and a Mean Absolute Error of 12046.71 and 86.6 respectively, using the Model 12. Linear Regression on Model 12 provides an impressive R^2 value of 0.98 along with the lowest AIC and BIC values.

3. *Should the company focus more on their website, or more on their app?*

- From the correlation plot, which shows that the Yearly Amount Spent is more strongly correlated with Time On App, while there is almost no correlation with Time on Website. This can be further proved from the Summary results of running their Linear Regression Model.

Yearly Amount Spent V/S Time on App

```
Call:
lm(formula = Ecommerce_Customers$`Yearly Amount Spent` ~ Ecommerce_Customers$`Time on App`)

Residuals:
     Min       1Q   Median       3Q      Max
-225.217  -41.391   -1.604   46.310  238.741

Coefficients:
                                Estimate Std. Error t value Pr(>|t|)
(Intercept)                       19.209     37.457   0.513    0.608
Ecommerce_Customers$`Time on App`  39.834      3.097  12.861   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 68.79 on 498 degrees of freedom
Multiple R-squared:  0.2493,    Adjusted R-squared:  0.2478
F-statistic: 165.4 on 1 and 498 DF,  p-value: < 2.2e-16
```

Yearly Amount Spent V/S Time on Website

```
Call:
lm(formula = Ecommerce_Customers$`Yearly Amount Spent` ~ Ecommerce_Customers$`Time on Website`)

Residuals:
     Min       1Q   Median       3Q      Max
-242.833  -54.494   -0.574   50.159  266.225

Coefficients:
                                 Estimate Std. Error t value Pr(>|t|)
(Intercept)                      506.9961   130.4001   3.888 0.000115 ***
Ecommerce_Customers$`Time on Website`  -0.2073     3.5173  -0.059 0.953029
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 79.39 on 498 degrees of freedom
Multiple R-squared:  6.974e-06, Adjusted R-squared:  -0.002001
F-statistic: 0.003473 on 1 and 498 DF,  p-value: 0.953
```

- It can be clearly observed that not only is the Coefficient for Time on Website close to 0, but also its p-value is greater than 0.05, indicating that the correlation is poor as well as the null hypothesis cannot be rejected.
- Whereas, for Time on App, the p-value is well below 0.01 and the coefficient is about 40.
- This shows that the more the customers spend time on the Company's app, the more they spend, resulting in higher revenue for the company.
- Thus, it would be more profitable for the company to spend more time bettering the App, than the Website.