MODELING 2022-2023 SEM 2

Linh K. T. Vo (1661618)

HAN University of Applied Sciences

Lecturer: Tijmen Weber

Date: 07/04/2023

Abstract

To improve the business performance of an online shop for next periods, customer behavior of making online purchases on the website is taken into account. They were asked to fill in a short survey at the end of their purchases. Some statistical analyses are carried out by Python to find out data-driven insights about customer behaviors. After that, using a multiple regression model is a great way to predict the purchase amount of a customer.

This report is presented in APA format.

Contents

Description of the data and cleaning

Table 1 presents data dictionary of the dataset with the name of variables and the description of variables.

Table 1

Data dictionary

Variable Name	Description
Purchase_Amount	Total amount spent by the customer on all products purchased (dependent variable)
Time_Spent_on_Website	Number of seconds spent by the customer on the website before final payment
Number_of_products_browsed	Number of products the customer browsed through before final payment
Pictures	Average number of pictures on the pages of the products purchased by the customer
Shipping_Time	Average number of days it took to ship each item bought by the customer
Review_rating	Average rating of the products purchased by the customer (ranging from 1 to 5 stars)
Find_website	How the customer found the website: Through a search engine, through friends or family, through a social media advertisement, or other
Ease_of_Purchase	How the customer rated the ease of their purchase on a scale of 1 (worst) to 5 (best)
Age	Age of the customer
Device	The device (PC or mobile) used by the customer to make the purchase

Models without missing values

Cleaning missing values

As can be seen from Table 2, there are 2 variables that have missing values: `Find_website` and `Ease_of_purchase`. For this part, these missing values are removed from the dataset. After cleaning missing values, there are 1553 rows left in a dataset.

Table 2

Missing values in each variable

Variables	Number of missing values
Purchase_Amount	0
Time_Spent_on_Website	0
Number_of_products_browsed	0
Pictures	0
Shipping_Time	0
Review_rating	0
Find_website	135
Ease_of_purchase	226
Age	0
Device	0

Cleaning negative values

There are 35 rows that have negative values in two variables `Purchase_Amount` and `Time_Spent_on_Website` which are unreasonable. These rows are removed before further analysis. After cleaning, there are 1518 rows left in the dataset.

Creating dummy variables

In the original dataset, `Find_website` and `Device` are categorical variables. In addition to that, the values in `Ease_of_purchase` are also cateogrical because they are ordinal data. Not like `Ease_of_purchase`, the values in `Review_rating` can be averaged and produce meaningful insight. Therefore, `Find_website`, `Device` and `Ease_of_purchase` are used to make dummy variables.

Table 3

Dummy variables

Find_website	Device	Ease_of_purchase
Search_Engine	PC	2.0
Social_Media_Advertisement	Mobile	3.0
Friends_or_Family		4.0
Other		5.0

Checking outliers

89 rows are Outliers by using CooksD. The detail code lines are explained in Python script. By the rule of thumb, only good reasons are taken into account to remove all the outliers.

In this case, all the missing values and negative values are removed. With the outliers spotted by CooksD, there is no other valid reasons to remove them because outliers can be good outlieres and useful for further analysis. The dataset is therefore kept the same and includes all the Outliers after analyzing the effect of two models with/without outliers in the **Model analysis** section below.

Checking multicollinearity

To check the multicollinearity between, VIF and correlation matrix are implemented.

Table 4

VIF factor of variables

VIF Factor	Features
189.37	Time_Spent_on_Website
191.56	Number_of_products_browsed
10.60	Pictures
9.91	Shipping_Time
29.10	Review_rating
18.40	Age
1.62	Find_website_Social_Media_Advertisement
1.03	Find_website_Other
1.14	Find_website_Friends_or_Family
1.41	Device_Mobile
1.00	Ease_of_purchase_2
1.24	Ease_of_purchase_3
1.24	Ease_of_purchase_5

From Table 4, it may be worth considering dropping one of the highly correlated variables (`Time_Spent_on_Website`, `Number_of_products_browsed`, `Review_rating`, or `Age`) to address the issue of multicollinearity.

Table 5

Correlation matrix between four variables

	Sample size	p-value	Correlation coefficient
Time spent vs Number of products	1518	0.000	0.976
Review rating vs Age	1518	0.827	0.006

	Sample size	p-value	Correlation coefficient
Picture vs Shipping Time	1518	0.693	0.010

Based on the data from Table 5, there is only one positive correlation between the two variables: $r(1518) = 0.976$, $p = 0$. The more the number of products are browsed, the more time customers spend on the website. There are no statistically significant correlations between two other pair of variables: $r(1518) = 0.006$, $p = 0.827$ and $r(1518) = 0.010$, $p = 0.693$.

Table 6

The effect of multicollinearity

	Model 1a	Model 1b	Model 1c
Intercept	380.95***	384.33***	379.86***
	(5.41)	(5.60)	(5.36)
Time_Spent_on_Website	0.29***		0.25***
	(0.03)		(0.01)
Number_of_products_browsed	-1.58	10.44***	
	(1.17)	(0.26)	
Observations	1,518	1,518	1,518
R^2	0.54	0.51	0.54
Adjusted R^2	0.54	0.51	0.54
Residual Std. Error	69.97	72.47	69.99
F Statistic	892.69***	1560.75***	1782.57***

*p<0.1; **p<0.05; ***p<0.01

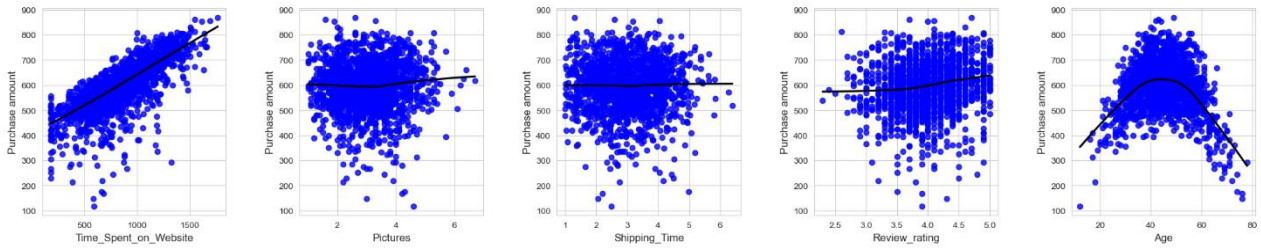
As can be seen from Table 6, `Number_of_products_browsed` and `Time_Spent_on_Website` are all statistically significant in model 1b and model 1c with *p-value* < 0.01. However, when putting them into the same model 1a, `Number_of_products_browsed` is not statisticlly significant. This can be explained by it high correlation. Therefore, one of them should be removed, which is `Number_of_products_browsed`.

Checking non-linear relationships

By using regression plot from seaborn package and parameter `lowess = True`, non-linear relationships between `Age` and `Purchase_Amount` is found out as shown in Figure 1. Adding the squared term `Age2` into the model assumes that the relationship between `Age` and `Purchase_Amount` is non-linear and follows a quadratic curve. This may improve the model's fit and accuracy.

Figure 1

Regression plots



While `Age` modeled linearly is significant, the polynomial model has a higher (adjusted) R-squared. And the coefficient with the highest term (in this case is `Age2`) is negative so it has an inverted U-shape. `Age2` is used to replace `Age` from now on.

Table 7

The effect of Age and Age2

	Model 1d	Model 1e
Intercept	644.91***	-191.25***
	(12.28)	(31.92)
Age	-1.15***	37.29***
	(0.26)	(1.41)
Age2		-0.42***
		(0.02)
Observations	1,518	1,518
R^2	0.01	0.34

Adjusted R^2	0.01	0.34
Residual Std. Error	102.61	83.74
F Statistic	18.84^{***}	394.76^{***}

*p<0.1; **p<0.05; ***p<0.01

Conclusion

After cleaning process, there are some few changes to the original dataset.

1. Removing all the missing values
2. Creating dummy variables for `Find_website`, `Device` and `Ease_of_purchase`
3. Removing all the negative values in `Time_Spent_on_Website` and `Numbe_of_products_browsed`
4. Outliers: keeping all the outliers
5. Multicolllinearity: removing `Number_of_products_browsed` from the model
6. Non-linear relationship: replacing `Age` with `Age2` (squared variable of `Age`)

Model with missing values

Cleaning negative values

To prepare model that includes all missing values, cleaning negative values is necessary. After cleaning, 1890 rows from the original dataset decreases to 1855 rows (35 rows are removed).

Creating dummy variables

The same goes for this model. Dummy variables are created based on three variables `Find_webiste`, `Device` and `Ease_of_purchase`. However, because this model has NA values (missing values), there is a slight change in the dummy variables as shown in Table 8.

Table 8

Dummy variables

Find_website	Device	Ease_of_purchase
Search_Engine	PC	2.0
Social_Media_Advertisement	Mobile	3.0
Friends_or_Family	Nan	4.0
Other		5.0
Nan		Nan

Checking multicollinearity

To check the multicollinearity between variables in the dataset that have all missing values, VIF and correlation matrix are implemented.

Table 9

VIF factor of variables

VIF Factor	Features
188.45	Time_Spent_on_Website
190.52	Number_of_products_browsed
10.40	Pictures
10.11	Shipping_Time
29.85	Review_rating
18.11	Age
1.61	Find_website_Social_Media_Advertisement
1.03	Find_website_Other
1.13	Find_website_Friends_or_Family
1.14	Find_website_nan
1.41	Device_Mobile
1.00	Ease_of_purchase_2
1.24	Ease_of_purchase_3
1.24	Ease_of_purchase_5
1.21	Ease_of_purchase_nan

From Table 9, it may be worth considering dropping one of the highly correlated variables (`Time_Spent_on_Website`, `Number_of_products_browsed`, `Review_rating`, or `Age`) to address the issue of multicollinearity.

Table 10

Correlation matrix between four variables

	Sample size	p-value	Correlation coefficient
Time spent vs Number of products	1855	0.000	0.975

	Sample size	p-value	Correlation coefficient
Review rating vs Age	1855	0.928	0.002
Picture vs Shipping Time	1855	0.540	0.014

Based on the data from Table 10, there is only one positive correlation between the two variables: $r(1855) = 0.975$, $p = 0$. The more the number of products are browsed, the more time customers spend on the website. There are no statistically significant correlations between two other pair of variables: $r(1855) = 0.002$, $p = 0.928$ and $r(1855) = 0.014$, $p = 0.540$.

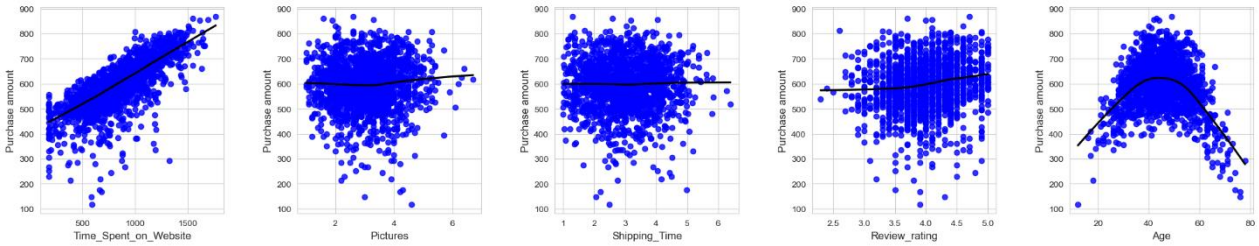
The result is the same to the dataset that excludes all missing values. Therefore, `Number_of_products_browsed` is removed from the model.

Checking non-linear relationships

By using regression plot from seaborn package and parameter `lowess = True`, non-linear relationships between `Age` and `Purchase_Amount` is found out as shown in Figure 2. Adding the squared term `Age2` into the model assumes that the relationship between `Age` and `Purchase_Amount` is non-linear and follows a quadratic curve. This may improve the model's fit and accuracy. `Age2` is used to replace `Age` from now on.

Figure 2

Regression plots



Conclusion

Cleaning process for dataset that has missing values is mostly similar to the dataset that has no missing values. To prepare data with missing values, there are some cleaning steps implemented:

1. Creating dummy variables for `Find_website`, `Device` and `Ease_of_purchase`
2. Removing all the negative values in `Time_Spent_on_Website` and `Numbe_of_products_browsed`
3. Multicolllinearity: removing `Number_of_products_browsed` from the model
4. Non-linear relationship: replacing `Age` with `Age2` (squared variable of `Age`)

Model presentation and explaination

Model with Outliers and without Outliers

Model 1 and model 2 with all independent variables. `Find_website_Search_Engine`, `Device_PC` and `Ease_of_purchase_4` are reference categories.

Table 11

Models with or without Outliers

	Model 1	Model 2
Intercept	342.53***	330.49***
	(18.28)	(13.77)
Time_Spent_on_Website	0.31***	0.31***
	(0.03)	(0.02)
Number_of_products_browsed	-2.25**	-2.63***
	(1.12)	(0.82)
Pictures	-3.08*	-1.98
	(1.81)	(1.32)
Shipping_Time	-1.18	-2.58**
	(1.76)	(1.30)
Review_rating	30.04***	32.01***
	(3.49)	(2.55)
Age	-1.56***	-1.12***
	(0.17)	(0.14)
Find_website_Search_Engine (ref)		
Find_website_Social_Media_Advertisement	4.51	6.90**
	(3.68)	(2.68)
Find_website_Other	-11.52	-25.93*
	(14.70)	(13.76)
Find_website_Friends_or_Family	14.13**	17.28***
	(6.55)	(4.88)
Device_PC (ref)		
Device_Mobile	-9.25**	-8.13***
	(3.77)	(2.75)
Ease_of_purchase_4 (ref)		
Ease_of_purchase_2	-26.16	0.00**
	(47.10)	(0.00)
Ease_of_purchase_3	-6.39	-9.34***

	(4.78)	(3.51)
Ease_of_purchase_5	6.81	3.85
	(4.72)	(3.45)
Observations	1,518	1,432
R^2	0.59	0.73
Adjusted R^2	0.59	0.73
Residual Std. Error	66.35	47.15
F Statistic	166.70***	319.69***

*p<0.1; **p<0.05; ***p<0.01

In general, R-squared of model 2 without outliers is much higher than that of model 1 with outliers. There are a lot differences between two models. There are some variables that become statistically significant in the model 2 such as `Shipping_Time`, `Find_website_Social_Media_Advertisement`… As explained above, there is no valid reason to remove al the outliers from the original dataset except for negative values in two variables `Time_Spent_on_Website` and `Number_of_products_browsed`. Therefore, dataset with outliers is still good for further analysis.

Model after data cleaning

After cleaning steps, this is the regression model of model 1 with the dataset exlcuding missing values and negative values. As shown in Table 12, generally, R-squared of this model has been improved (62%) from the model before cleaning with some changes as described above. This number means that 62% variance in purchase amount is explained by the variance of all these independent variables. Some of the variables are significant except average number of pictures, shipping time, find website through social media and other tools and ease of purchase. Most of the effects have negative influence and the rest has positive influence on the purchase amount spent on the website.

Indeed, review rating has the largest impact on the purchase spent. It has strong effect as 30.45 coefficient value with *p-value*<0.01. With the same pattern, the more time spent on the website, the more money customers spend.

The shipping time and the average number of pictures have no significant effect on the purchase amount spent while age is statistically significant. After adjusting to squared variable, the effect of this variable is negative as -0.03 coefficient.

From the categorical variables, only Friends_or_Family and device_PC have positive influence on money spent. Customers spend less money on the website if they browse more products on Mobile. It means that customers spend more money on the website if they use PC. This tool must be the most efficient one to increase customers' reach. Beside that, ease of purchase with all ratings are not statistically significant.

Table 12

Model 1 prediction purchase amount

	Model 1
Intercept	322.19***
	(16.41)
Time_Spent_on_Website	0.26***
	(0.01)
Pictures	-2.76
	(1.73)
Shipping_Time	-1.01
	(1.69)
Review_rating	30.45***
	(3.36)
Age2	-0.03***
	(0.00)
Find_website_Search_Engine (ref)	
Find_website_Social_Media_Advertisement	4.31
	(3.55)
Find_website_Other	-11.76
	(14.16)

Find_website_Friends_or_Family	14.24**
	(6.31)
Device_PC (ref)	
Device_Mobile	-8.58**
	(3.63)
Ease_of_purchase_4 (ref)	
Ease_of_purchase_2	-12.20
	(45.37)
Ease_of_purchase_3	-6.27
	(4.60)
Ease_of_purchase_5	6.68
	(4.55)
Observations	1,518
R^2	0.62
Adjusted R^2	0.62
Residual Std. Error	63.90
F Statistic	204.42***

*p<0.1; **p<0.05; ***p<0.01

Model with standardized variables

Four continous variables are standardized: `Time_Spent_on_Website`, `Pictures`, `Shipping_Time`, `Review_rating`. Standardization does not work for non-linear effects. Based on the data in Table 13, the R-squared stays the same in two models. The only difference is the coefficient of standardized variables. In model 1 standardized, `Time_Spent_on_Website` has the largest effect on purchase amount spent by customers. The effect of `Review_rating` is less than half after being standardized. `Pictures` and `Shipping_Time` stay similar in two models. Therefore, model 1 standardized has higher accuracy between the effects of variables and is used for prediction purchase amount of a new customer.

Table 13

Models with and without standardized variables

	Model 1	Model 1 standardized
Intercept	322.19***	648.33***

	(16.41)	(4.66)
Time_Spent_on_Website	0.26***	77.10***
	(0.01)	(1.65)
Pictures	-2.76	-2.62
	(1.73)	(1.65)
Shipping_Time	-1.01	-0.98
	(1.69)	(1.64)
Review_rating	30.45***	14.90***
	(3.36)	(1.64)
Age2	-0.03***	-0.03***
	(0.00)	(0.00)
Find_website_Search_Engine (ref)		
Find_website_Social_Media_Advertisement	4.31	4.31
	(3.55)	(3.55)
Find_website_Other	-11.76	-11.76
	(14.16)	(14.16)
Find_website_Friends_or_Family	14.24**	14.24**
	(6.31)	(6.31)
Device_PC (ref)		
Device_Mobile	-8.58**	-8.58**
	(3.63)	(3.63)
Ease_of_purcahse_4 (ref)		
Ease_of_purchase_2	-12.20	-12.20
	(45.37)	(45.37)
Ease_of_purchase_3	-6.27	-6.27
	(4.60)	(4.60)
Ease_of_purchase_5	6.68	6.68
	(4.55)	(4.55)
Observations	1,518	1,518
R^2	0.62	0.62
Adjusted R^2	0.62	0.62
Residual Std. Error	63.90	63.90
F Statistic	204.42***	204.42***

*p<0.1; **p<0.05; ***p<0.01. All continuous variables are standardized.

Model for prediction of purchase amount

Model 1 with standardized variables is good for prediction. However, according to Parsimony, it is best to pick the simplest model that fits the purpose. To make a better prediction, it is necessary to leave out some variables that are not relevant. As can be seen from Table 14, there are only 5 independent variables left. But the R-squared and Adjusted R-Squared stay the same as the model with full independent variables. It means that leaving out some irrelevant variables doesn't affect the outcome of the result of regression model. Therefore, the prediction of purchase amount of customers is calculated with this model.

Table 14

Model for prediction

	Model 1
Intercept	650.24***
	(4.29)
Time_Spent_on_Website	76.85***
	(1.64)
Review_rating	14.93***
	(1.64)
Age2	-0.03***
	(0.00)
Find_website_Friends_or_Family	12.69**
	(6.16)
Device_Mobile	-8.77**
	(3.63)
Observations	1,518
R^2	0.62
Adjusted R^2	0.62
Residual Std. Error	63.98
F Statistic	487.24***

*p<0.1; **p<0.05; ***p<0.01. All continuous variables are standardized

Table 15

The information of a customer

Time spent	Age	Review rating	Find website	Device
723	35	4.5	Friends or Family	PC

With the information a new customer as presented in Table 15, the model predicts that the customer pays 631.18 after standardizing variables and calculating squared Age of new inputs.

Model with missing values

After cleaning process as described above, model with missing values is implented with Mice and compared with model without missing values as shown in Table 16. The number of observations increases from 1518 to 1855.

In the new model, the variables that are not statistically significant in the model without missing values are still the same. New dummy variables for missing values does not have influence on the purchase amount spent on the website. These two variables are just missing values at random.

There is not much difference between models. Multiple imputation has not produced any strange results. The conclusions still stay the same so multiple imputation can be a good solution.

Table 16

Models with and without missing values

	Mode 1 with missing values	Mode l without missing values
Intercept	644.83***	648.33***
	(4.24)	(4.66)
Time_Spent_on_Website	75.10***	77.10***
	(1.47)	(1.65)
Pictures	-1.65	-2.62
	(1.47)	(1.65)
Shipping_Time	-1.38	-0.98
	(1.46)	(1.64)
Review_rating	14.85***	14.90***
	(1.46)	(1.64)
Age2	-0.02***	-0.03***
	(0.00)	(0.00)

Find_website_Search_Engine (ref)		
Find_website_Social_Media_Advertisement	5.89	4.31
	(3.28)	(3.55)
Find_website_Other	-12.27	-11.76
	(12.79)	(14.16)
Find_website_Friends_or_Family	12.50**	14.24**
	(5.93)	(6.31)
Find_website_nan	-3.70	
	(5.81)	
Device_PC (ref)		
Device_Mobile	-7.87**	-8.58**
	(3.25)	(3.63)
Ease_of_purcahse_4 (ref)		
Ease_of_purchase_2	-11.20	-12.20
	(44.72)	(45.37)
Ease_of_purchase_3	-5.74	-6.27
	(4.38)	(4.60)
Ease_of_purchase_5	6.40	6.68
	(4.34)	(4.55)
Ease_of_purchase_nan	1.50	
	(4.61)	
Observations	1,855	1,518
R^2		0.62
Adjusted R^2		0.62
Residual Std. Error		63.90
F Statistic		204.42***

*p<0.1; **p<0.05; ***p<0.01. All continuous variables are standardized.