Customer Segmentation using Logistic Regression

(Project Report STAT 5120, Regression Analysis)

By

Michael Akinola

Eniola Oladele

Department of Mathematics and Statistics
Austin Peay State University
Clarksville, United States
makinola@my.apsu.edu
eoladele@my.apsu.edu

Course Instructor:

Dr. Matthew Jones

May 6, 2020

***Abstract* — Machine learning has found application in various sectors of human endeavors of which marketing is not an exemption. Marketing is a key part of any business. It requires strategic steps in order to maximize its impact in today's dynamic world using every available medium. In this work, we applied Logistic regression – a classification algorithm in machine learning to achieve this goal. Our goal is to achieve high true positive (TP) and reduce false negative or type 2 error. In other words, we want to have high tolerance on false positive or type 1 error. We used two methods in our choice of algorithm and evaluated them. Both methods gave accuracy of 0.70 but one method outshone the other in terms of error type reduction.**

Keywords — logistic regression, spending score, customer segmentation, marketing

I. INTRODUCTION

The first steps in any data science research or analytics effort, once the business goal has been defined, are to understand and prepare the data of interest. Our data was fetched from the Kaggle. We used a mall customer data set which contains five columns namely; CustomerID, Gender, Age, Annual Income and Spending Score. The preparation processes involve dropping of column that is not needed, transformation of categorical columns, Gender and Spending Score to discrete values for absorption into our model. The Spending Score column was transformed from continuous values to categorical values to have High (main targets) and Low values with set boundaries and then later made discrete. We are lucky to have a clean data to work with in terms as there are no null values or unusual character in the data set. We proceeded to carry out exploratory data analysis (EDA) which took two major steps; statistical checks and visualizations, such as boxplots, histogram and scatter plots, of our data set to know the distribution of our metrics if there is any bias in the distribution of our metrics. The EDA was followed by the application of machine learning algorithm on our data. We took two approaches in the application of Logistic Regression algorithm to our data and the two approaches were evaluated.

II. DATA EXPLORATION

To understand the data, we first load this data into our R Studio, before looking at several instances from the data. Table 1 shows our top six samples. Next, we extracted a statistical summary of the entire data set before generating some visualizations of our data set by various plots. Since this data has already been cleaned we will not need to perform additional tasks.

Table 1: Top six samples from data set

	Gender	Age	Annual_Income	Spending_Score
1	Male	19	15	39
2	Male	21	15	81
3	Female	20	16	6
4	Female	23	16	77
5	Female	31	17	40
6	Female	22	17	76

Table 2: Statistics of data set

Gender	Age	Annual_Income	Spending_Score
Female:112	Min. :18.00	Min. : 15.00	Min. : 1.00
Male : 88	1st Qu.:28.75	1st Qu.: 41.50	1st Qu.:34.75
	Median :36.00	Median : 61.50	Median :50.00
	Mean :38.85	Mean : 60.56	Mean :50.20
	3rd Qu.:49.00	3rd Qu.: 78.00	3rd Qu.:73.00
	Max. :70.00	Max. :137.00	Max. :99.00

Our data set contains spending information for 112 females and 88 males as shown in Table 2. This distribution is not biased. The five number summaries of other predictors are also okay. They are all important consideration in moving forward with our choice of data and EDA.

Table 3 below shows the statistics by gender without high shortcomings in age and annual income distribution.

Table 3: Statistics by Gender

Gender	Age	Annual_Income	Spending_Score
Female:112	Min. :18.0	Min. : 16.00	Min. : 5.00
Male : 0	1st Qu.:29.0	1st Qu.: 39.75	1st Qu.:35.00
	Median :35.0	Median : 60.00	Median :50.00
	Mean :38.1	Mean : 59.25	Mean :51.53
	3rd Qu.:47.5	3rd Qu.: 77.25	3rd Qu.:73.00
	Max. :68.0	Max. :126.00	Max. :99.00
---	---	---	---
Gender	Age	Annual_Income	Spending_Score
Female: 0	Min. :18.00	Min. : 15.00	Min. : 1.00
Male :88	1st Qu.:27.75	1st Qu.: 45.50	1st Qu.:24.50
	Median :37.00	Median : 62.50	Median :50.00
	Mean :39.81	Mean : 62.23	Mean :48.51
	3rd Qu.:50.50	3rd Qu.: 78.00	3rd Qu.:70.00
	Max. :70.00	Max. :137.00	Max. :97.00

Table 4: Statistics by Annual Income distribution

Gender	Age	Annual_Income	Spending_Score
Female:58	Min. :18.00	Min. : 15.00	Length:102
Male :44	1st Qu.:26.25	1st Qu.: 42.25	Class :character
	Median :31.50	Median : 60.00	Mode :character
	Mean :34.92	Mean : 60.20	
	3rd Qu.:38.75	3rd Qu.: 77.75	
	Max. :70.00	Max. :137.00	
---	---	---	---
Gender	Age	Annual_Income	Spending_Score
Female:54	Min. :18.00	Min. : 15.00	Length:98
Male :44	1st Qu.:34.00	1st Qu.: 40.00	Class :character
	Median :44.50	Median : 62.00	Mode :character
	Mean :42.94	Mean : 60.94	
	3rd Qu.:51.75	3rd Qu.: 78.00	
	Max. :69.00	Max. :137.00	

The table above shows income distribution between the two genders. There are 102 low-income earners and 98 high-income earners with both having a good proportion of both genders. We further explore our data set by plotting various charts as seen in figures 1 to figures 9. Figure 9 showed gave an interesting insight of our data having five different clusters of spending patterns.

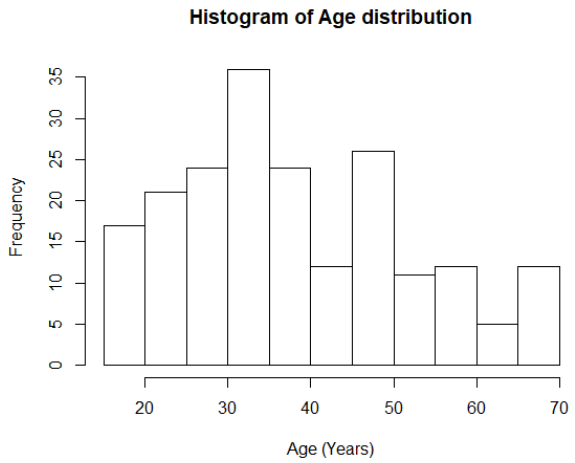

Figure 1: Histogram showing age distribution

Here, we see that most of the customers age ranges from teen to 50 years.

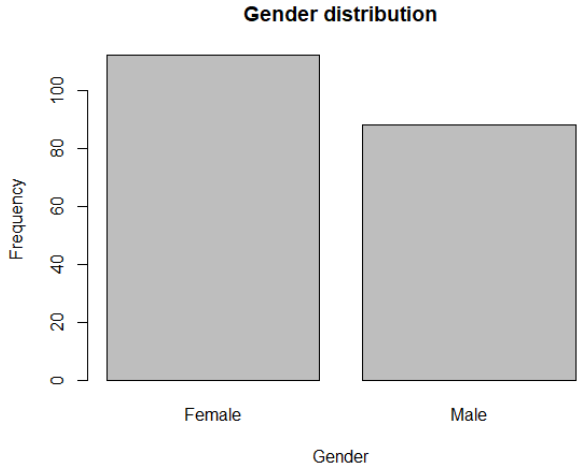

Figure 2: Histogram showing gender distribution

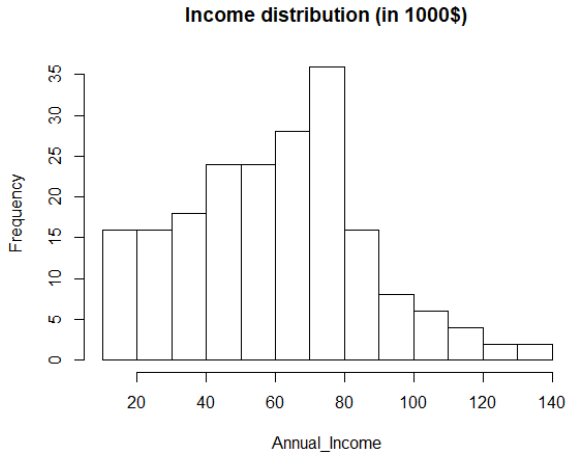

Figure 3: Histogram showing overall income distribution

The income distribution chart in figure 3 shows that most of the customers earn between $15,000 and $80,000 annually.

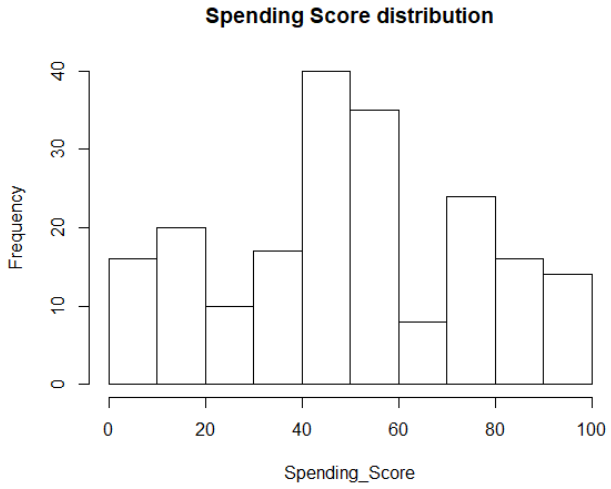

Figure 4: Histogram showing spending score

Spending score range of 40 to 60 are prominent in the data as show in figure 4 above.

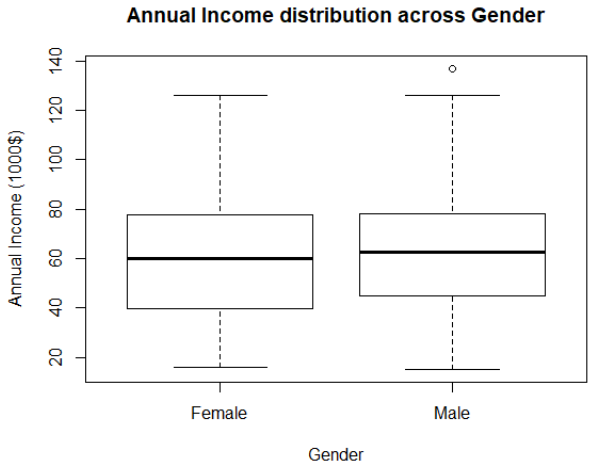

Figure 5: Boxplots showing the annual income distribution by gender

Figure 5 shows that males seem to generally earn more than females also with higher median annual income. Both genders have same highest and lowest annual earning. There is an outlier in the male income distribution of about $140,000 annual income. This is not out of place and we choose to keep it for modelling.

Figure 6: Boxplots showing spending score distribution by gender

It is not surprising that females generally spend more than their male counterparts though they appear to have same median spending score.

Figure 7: Scatterplot of age against spending score

The scatter plot in figure 7 shows that there is no correlation between these two variables. Though there seems to be a sign of clusters. Similarly, figure 8 also shows that there is no correlation between aga and annual income.

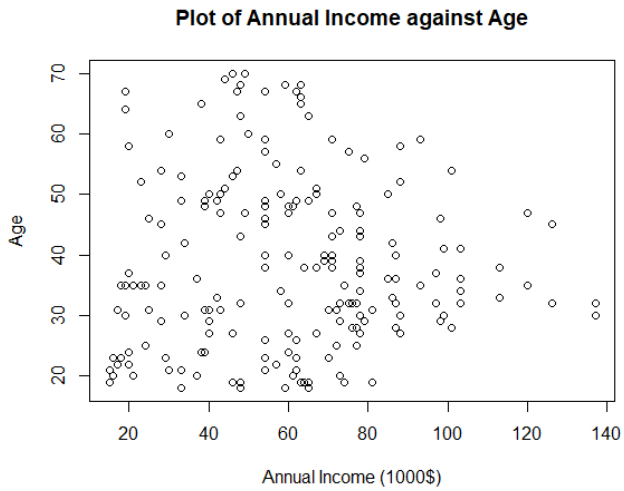

Figure 8: Scatter plot of age against annual income

Figure 9: Scatterplot of annual income against spending score

We can see five different clusters in the scatter plot shown in figure 9 above. This gives an insight that there are possibly five different types of shoppers in our data. Had it been we want to perform an unsupervised learning, We could have given names to the different clusters shown above but we are interested in separating the shoppers into two classes; high-spending and low-spending customers.

III. METHOD

Machine learning algorithms can be classified by the method in which they are constructed into supervised and unsupervised learning. Supervised learning methods use training data to build a model, which is subsequently applied to additional data while unsupervised learning methods seek relationships among data points that can be leveraged to construct a model that is subsequently applied to the data of interest. As earlier stated, we used the logistic regression algorithm - a supervised learning method in this work.

Logistic regression is a classification algorithm in which the predictors are numerical values, and response is binary ('Yes' or 'No', 'Positive' or 'Negative'). We started by transforming our data into usable format for our model. The Gender column was transformed to have dummy variables where 'Male' is assigned a value of '1' and 'Female' a value of '0'. Then our Spending Score column was transformed to 'High' and 'Low' with values greater or equal to 50 assigned to the former and values less than 50 were assigned to the latter. Eventually the 'High's' and 'Low's' were assigned dummy variables of '1' and '0' respectively as shown in the Table 5 below.

Table 5: Top five samples after encoding of categorical variables

	Gender	Age	Annual_Income	Spending_Score
1	1	19	15	0
2	1	21	15	1
3	0	20	16	0
4	0	23	16	1
5	0	31	17	0
6	0	22	17	1

While there are many logistic model transformations, we chose the logit transformation for our modelling. The model was initially applied to the entire data set then to a split of our data and both were evaluated.

Logit Function

The logit function is defined as the logarithm of the odds (i.e., $p/(1-p)$), which is also known as the log-odds. Thus, the logit function can be written for a probability of success p.

$$\text{logit}(p) = \log(p/(1-p)); \text{ where } 0 \leqslant p \leqslant 1 \quad \textit{(Equation 1)}$$

This relationship can be inverted to obtain the logistic function, which for a parameter α is defined by the following expression:

$$\text{logit}^{-1}(\alpha) = \text{logistic}(\alpha) = 1/(1+\exp(-\alpha)) \quad \textit{(Equation 2)}$$

The logit function is an S shaped curve that converts real numbers into a probability. The logit function is related to the sigmoid function, but is centered at the origin (0, 0). The diagram below is the plot of the logistic function, or the inverse of the logit function.

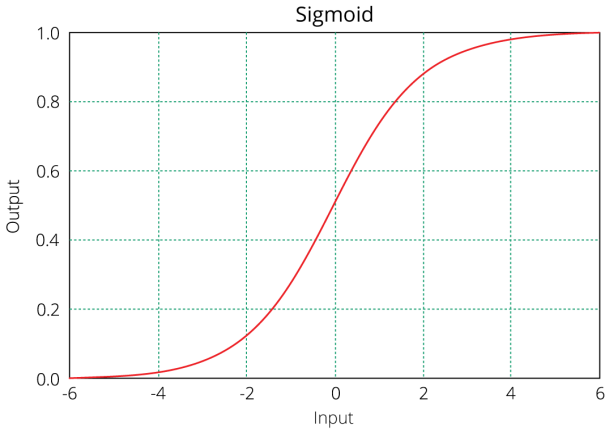

Image credit: (https://www.hackerearth.com/practice/machine-learning/machine-learning-algorithms/logistic-regression-analysis-r/tutorial/)

Figure 10: The logit function

Model Evaluation

After building a classification model, there are various ways to evaluate the model. In this work, we utilized the confusion matrix metrics to measure the performance of our model. A confusion matrix is a table layout that displays a quantitative visualization of the performance of an algorithm. We have a special case in this project where there are only two classes, high (positive) and low (negative). The confusion matrix is shown in Figure 11 below with 4 different combinations of actual and predicted values.

The four different combinations of actual and predicted values are true negative (TN), false negative (FN), true positive (TP) and false positive (FP).

Spending Score		Actual	
		Low (0)	High (1)
Prediction	Low (0)	True Negative (TN)	False Negative (FN)
	High (1)	False Positive (FP)	True Positive (TP)

Figure 11: Confusion matrix

True Positive (TP) is the total number of predicted positives that are actually positive.

True Negative (TN) is the total number of predicted negatives that are actually negative.

False Positive (FP) aka Type 1 Error is the total number of predicted positives that are actually negative.

False Negative (FN) aka Type 2 Error is the total number of predicted negatives that are actually positive.

All classification metrics are calculated with the predicted values and the actual values. The predicted values are generated by applying the classification model on the testing dataset and the observed values are the true outcome of testing dataset.

Accuracy Score

Accuracy score is the starting point of classification evaluation. It is calculated as the proportion of correct predictions over all predictions. For example, if a classification model yields 100 predictions, 85 out of 100 are predicted correctly, then the accuracy score will be 0.85 or 85%.

Precision

Precision is the proportion of the prediction that is actually correct. It shows the proportion of precision of class 0 predictions that are actually class 0. Likewise, it shows the proportion of class 1 predictions that are actually class 1.

Recall

Recall is the proportion of actual class of a label that is identified correctly.

$$AccuracyScore = \frac{CorrectPredictions}{AllPredictions} = \frac{TP+TN}{TP+TN+FP+FN}$$

$$PositivePrecision = \frac{TruePositive}{PredictedPositive} = \frac{TP}{TP+FP}$$

$$NegativePrecision = \frac{TrueNegative}{PredictedNegative} = \frac{TN}{TN+FN}$$

$$PositiveRecall = \frac{TruePositive}{ActualPositive} = \frac{TP}{TP+FN}$$

$$NegativeRecall = \frac{TrueNegative}{ActualNegative} = \frac{TN}{TN+FP}$$

Equations 3 – 7

IV. Results

A. Full Data Set Model

The data was modelled as a whole and then evaluated.

Spending Score		Actual	
		Low (0)	High (1)
Prediction	Low (0)	61	37
	High (1)	23	79

Figure 12: Confusion matrix of full data set model

Model evaluation

Accuracy Score:

$$AccuracyScore = \frac{CorrectPredictions}{AllPredictions} = \frac{TP+TN}{TP+TN+FP+FN}$$

Accuracy Score = 140/200
Accuracy Score = 0.7 or 70 %

Precision:

$$PositivePrecision = \frac{TruePositive}{PredictedPositive} = \frac{TP}{TP+FP}$$

$$NegativePrecision = \frac{TrueNegative}{PredictedNegative} = \frac{TN}{TN+FN}$$

Positive Precision = 79/102
Positive Precision = 0.77 or 77 %

Negative Precision = 61/98
Negative Precision = 0.62 or 62 %

Recall:

$$Positive\,Recall = \frac{True\,Positive}{Actual\,Positive} = \frac{TP}{TP + FN}$$
$$Negative\,Recall = \frac{True\,Negative}{Actual\,Negative} = \frac{TN}{TN + FP}$$

Positive Recall = 79/116
Positive Recall = 0.68 or 68 %

Negative Recall = 61/84
Negative Recall = 0.726 or 72.6 %

B. Split Model

Data was split into train test of 70 % : 30 %

Spending Score		Actual	
		Low (0)	High (1)
Prediction	Low (0)	12	0
	High (1)	18	30

Figure 13: Confusion matrix of the Split Model

The vertical represents the actual outcome while the horizontal represents the model prediction.

Model evaluation

Accuracy Score:

$$Accuracy\,Score = \frac{Correct\,Predictions}{All\,Predictions} = \frac{TP + TN}{TP + TN + FP + FN}$$

Accuracy Score = 42/60
Accuracy Score = 0.7 or 70 %

Precision:

$$Positive\,Precision = \frac{True\,Positive}{Predicted\,Positive} = \frac{TP}{TP + FP}$$
$$Negative\,Precision = \frac{True\,Negative}{Predicted\,Negative} = \frac{TN}{TN + FN}$$

Positive Precision = 30/30
Positive Precision = 1 or 100 %

Negative Precision = 12/30
Negative Precision = 0.286 or 28.6 %

Recall:

$$Positive\,Recall = \frac{True\,Positive}{Actual\,Positive} = \frac{TP}{TP + FN}$$
$$Negative\,Recall = \frac{True\,Negative}{Actual\,Negative} = \frac{TN}{TN + FP}$$

Positive Recall = 30/48
Positive Recall = 0.625 or 62.5 %

Negative Recall = 12/12
Negative Recall = 1 or 100 %

V. Discussions

Generally, it is better we split our data set so we can evaluate our model using data from our original data set but from our result, we could see some advantage of using the full data set over splitting our data set.

Among other information provided by our evaluation metrics, we chose to discuss the details of the confusion metrics as it relates to business goal defined at the beginning of our report. It is important to note that accuracy score is very intuitive but not always reliable. In our model, we got same accuracy score of 70 %. This is not bad and can be improved upon.

Among the four terms, false positive or type 1 error and false negative or type 2 error are more crucial. The chances of committing these two types of errors are inversely proportional - that is, decreasing type 1 error rate increases type 2 error rate, and vice versa. In different situations, we try to avoid either type 1 error or type 2 error. Figure 14 below shows these four terms in a confusion matrix format.

Spending Score		Actual	
		Low (0)	High (1)
Prediction	Low (0)	True Negative (TN)	Type II Error
	High (1)	Type I Error	True Positive (TP)

Figure 13: Confusion matrix showing error type

Our goal is to achieve high true positive (TP) and high recall on positive class, or high spending scores class. In other words, we are willing to sacrifice customers with low spending scores as high spending ones i.e. sacrifice positive precision to achieve high positive recall. Another way to describe our goal is, we want to reduce false negative (FN) or type 2 error, to have high tolerance on false positive or type 1 error. This means we don't mind our model having high likelihood of classifying a data point or any given customer to a high spending class so we can attract or reach them to special offers and other advertisement.

VI. Conclusion

We conclude by saying our full data set model works fine with an accuracy score of 70 % and high positive recall rate of 68 % against 62.5 % of the data set split model. Adjusting some hyper-parameter may improve our split model. Also, having more data samples and probably more predictor variables can generally improve our models.

References

Julian J. Faraway (2015). *Linear Models with R.* University of Bath, United Kingdom.

https://www.kaggle.com/vjchoudhary7/customer-segmentation-tutorial-in-python#Mall_Customers.csv

https://www.rdocumentation.org/packages/caret/versions/3.45/topics/confusionMatrix

https://www.hackerearth.com/practice/machine-learning/machine-learning-algorithms/logistic-regression-analysis-r/tutorial/

https://www.r-bloggers.com/how-to-perform-a-logistic-regression-in-r/

stackoverflow.com

** all links are valid as at 5/3/2020