Nader Khashab
April 29, 2021

Customer Segmentation: Final Report

GitHub Project Link

Problem Statement

The mall has acquired some basic data about its customers. They data on their age, gender, annual income, and spending score. We are trying to understand out of the data gathered, what will influence a customers spending score so that the marketing team could plan strategically. Can this dataset be broken up into groups? Can we predict the likelihood of someone purchasing an item from the mall? To demonstrate this, I will be mainly using supervised learning so that by the end we know what variables will have the most impact on someones spending score.

Data Wrangling

This part of the project did not require much effort due to the fact that the dataset obtained already had been cleaned. Along with this, there weren't many columns that were not going to be used in the analysis. The only column that would be dropped was ‘Customer ID’ as it did not impact spending score in any way and neither would it come into any use for our particular analysis. I then began to create sorted tables based on individual columns before proceeding to the Exploratory Data Analysis (EDA) portion. The columns to be utilized in my analysis is as follows:

Age
Income
Gender
Spending Score

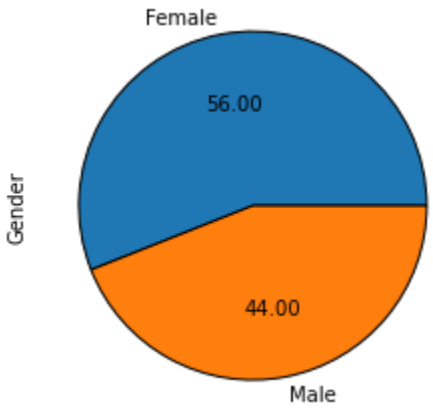
	Gender	Age	Income	SpendingScore
156	Male	37	78	1
158	Male	34	78	1
8	Male	64	19	3
32	Male	53	33	4
30	Male	60	30	4
167	Female	33	86	95
145	Male	28	77	97
185	Male	30	99	97
19	Female	35	23	98
11	Female	35	19	99

Exploratory Data Analysis

In this section I began to explore the data to understand any surface level connections that I could make. To start off the analysis, I found key parameters such as mean, std, minimum, maximum, and etc. of the entire dataset.

	Age	Annual Income (k$)	Spending Score (1-100)
count	200.00	200.00	200
mean	38.85	60.56	50.20
std	13.97	26.26	25.82
min	18.00	15.00	1.00
25%	28.75	41.500	34.75
50%	36.00	61.500	50.00
75%	49.00	78.00	73.00
max	70.00	137.00	99.00

Next, the data set is based on 200 individuals with 44% of the customers being male and 56% being female which can be represented by a pie chart.

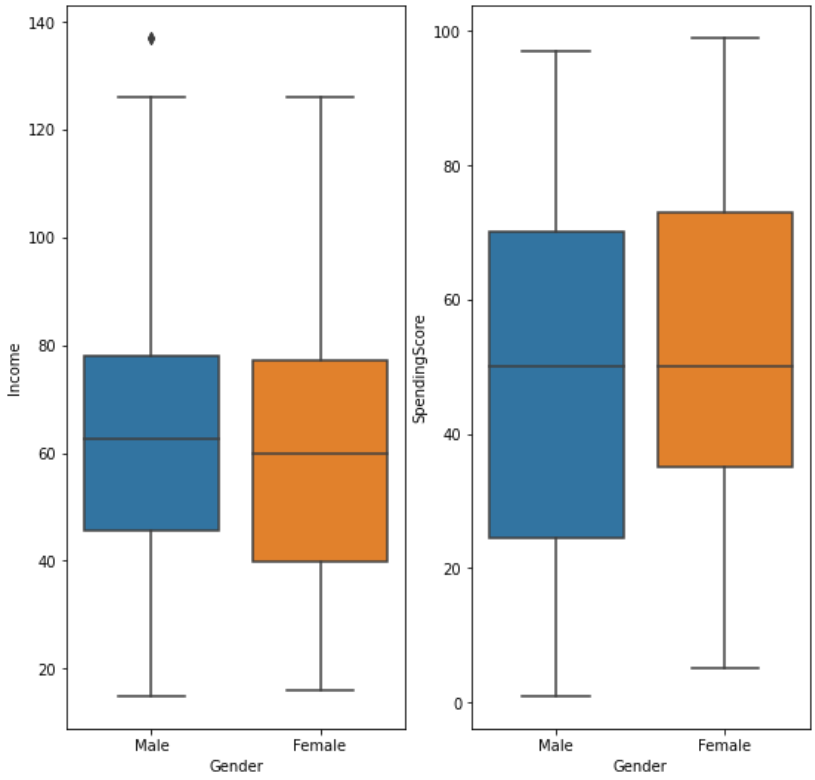


With this data I further broke it down and I was able to calculate the average age, income, and spending score based on the gender of the customer which is shown as follows.

Male Average Breakdown	
Age	39.806818
Income	62.227273
SpendingScore	48.511364

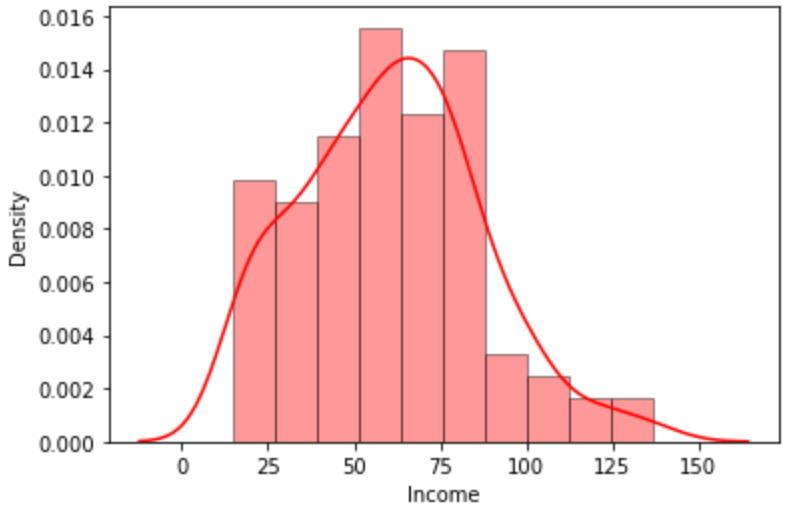
Female Average Breakdown	
Age	38.098214
Income	59.250000
SpendingScore	51.526786

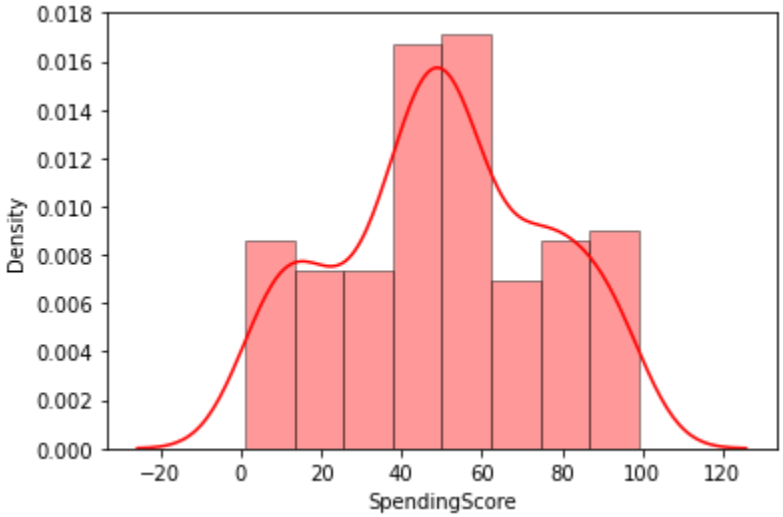
On average, Male tend to have a higher income but a lower spending score while females tend to make less but have a higher spending score. From this we can expect a female walking into our mall to come out with a purchase a little over 50% of the time. This data can also be visualized with a box and whiskers plot:

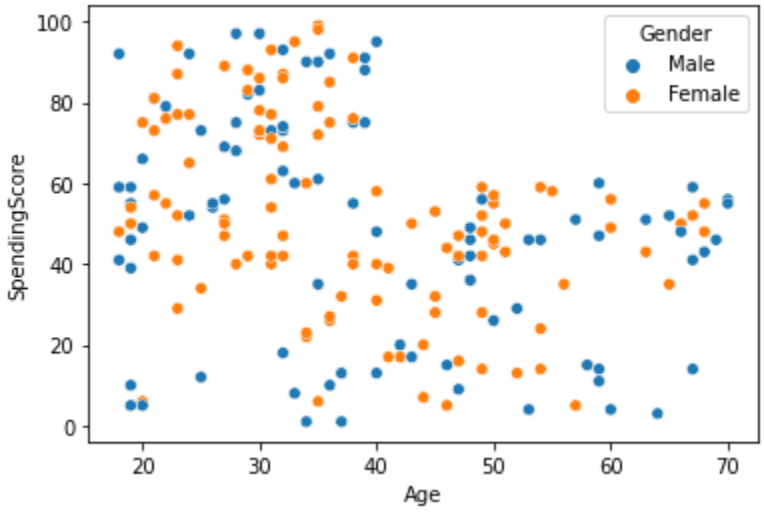


The box and whiskers plots Gender vs. Income and Gender vs. Spending Score in order to better visualize the correlation between the variables. We can see that the Males typically have a wider spread of spending score as opposed to the females which seems to have a tighter observation between the first quartile and third.

I furthered my analysis of the data set to understand the distribution of income and spending score as a whole by creating a distribution plot.


0.016
0.014
0.012
0.010
0.008
0.006
0.004
0.002
0.000
Density
0
25
50
75
100
125
150
Income


0.018
0.016
0.014
0.012
0.010
0.008
0.006
0.004
0.002
0.000
Density
−20
0
20
40
60
80
100
120
SpendingScore


100
80
60
40
20
0
SpendingScore
20
30
40
50
60
70
Age
Gender
Male
Female

Lastly, when plotting Age vs Spending Score We can clearly see a correlation. Younger individuals ranging from the ages of 18 to 40 tend to have a higher spending score on average than those above the age of 40. This means, typically when an individual walks into the mall between the ages of 18 to 40, they are most likely to walk out with something in their hands. The plot does also show that those in the range of ages 18 to 40 that are most likely to have a higher spending score lean towards the female side rather than the males.

Machine Learning Analysis

Before trying any machine learning analysis, I first had to prepare the data to ensure they were fit for model analysis. This dataset was small and clean so this task was not difficult. I ensured that each of the columns were actual numeric values. I also dropped the gender column as I did not believe it would have any impact on the modeling analysis. Because I was focusing on supervised learning it was important to split the data into training and test data sets.

Once I completed this task I thought out of some machine learning analysis to be used for this modeling. The four that I utilized are Linear Regression, KNN Regressor, Random Forest Regressor, and XGBoost.

To evaluate the models I used root mean square error (RMSE). The equation used to calculate RMSE is as follows:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N} (Predicted_i - Actual_i)^2}{N}}$$

This model is measuring the difference between the actual values and the predicted values. The better model will have a lower RMSE which means the predicted values were closer to the actual values resulting in a more accurate model for future purposes.

Linear Regression

Linear regression is an important statistical algorithm to learn the correlation between dependent and independent variables.A linear relation will result in the dependent variable increasing or decreasing depending on the change of the independent variables. I calculated the RMSE of the linear regression model to be 25, which is on the higher end of the spectrum. This essentially means when predicting a value it will be in the range of +- 25 from the actual value.

KNN Regressor

KNN Regression approximates the association between the independent variable and the continuous outcome by averaging the observations in the same neighborhood. These neighborhoods are chosen by the analysis or through cross validation. For this analysis, I had the RMSE done from k values that ranged between 1 and 20. These are the results:

K Value	RMSE
1	28.50
2	24.63
3	23.21
4	21.54
5	21.37
6	20.53
7	20.45
8	20.77
9	20.41
10	20.55
11	21.28
12	21.38

13	21.37
14	21.65
15	21.53
16	21.79
17	22.02
18	22.04
19	22.25
20	22.50

From the RMSE calculations, the lowest RMSE was 20.41 with a k-value of 9. This means that predictions are +- 20.41 on average from the actual values and has a better prediction model than the linear regression.

XGBoost

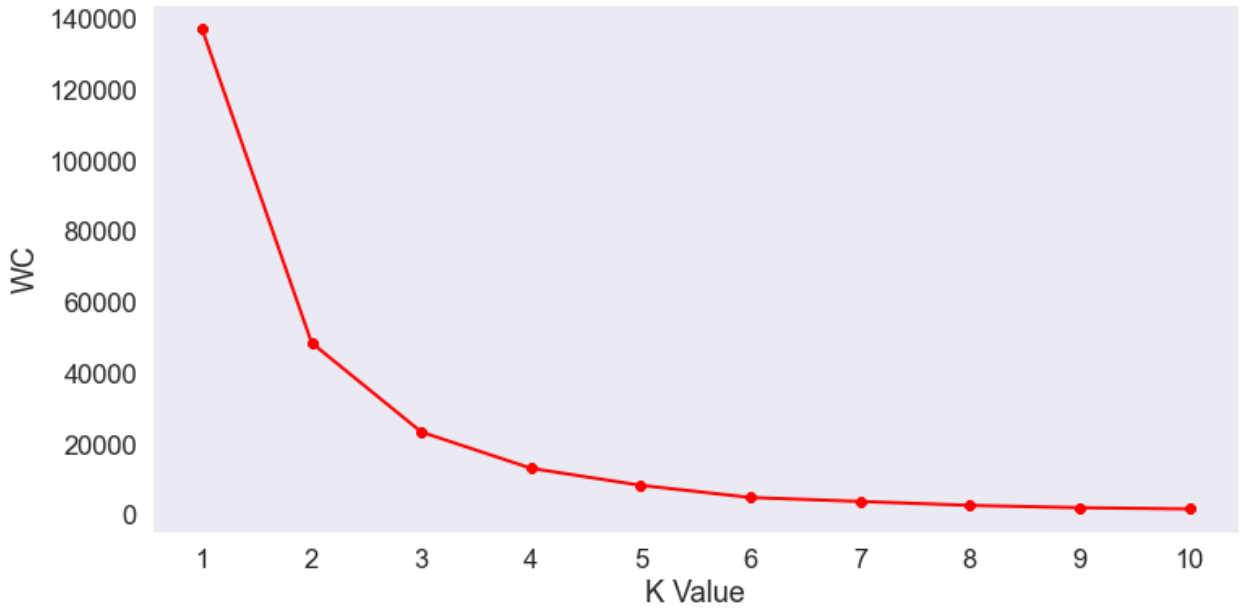
XGBoost has been a favorite modeling technique by kaggle competitors in the data science industry. It is scalable and accurate. It was created to to have improved model performance at an exceptional computing speed. Using this model I predicted the Spending Score from our independent variables and was able to calculate an RMSE of 21.57. This is higher than the RMSE of the KNN Regressor which keeps KNN Regressor as the top modeling technique for this data set so far.

Random Forest Regression

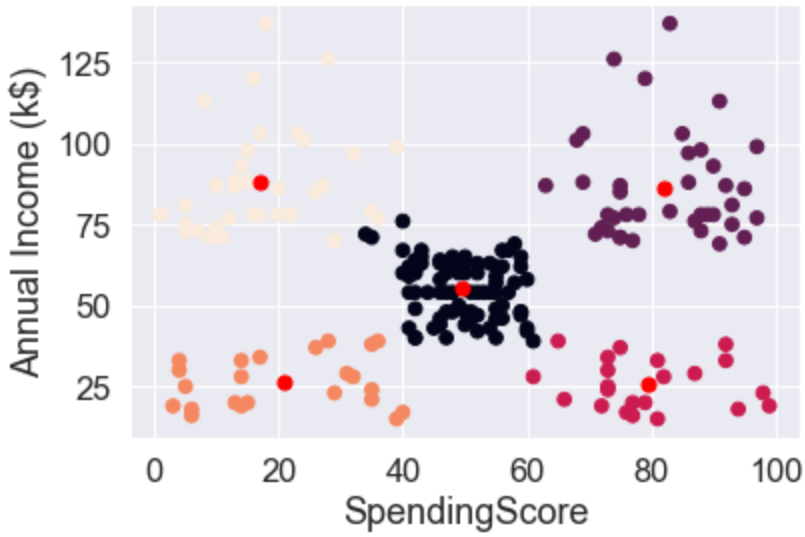
Random Forest Regression is a little different from the other modeling techniques. Instead of searching for the most important feature while splitting the node up, it instead will find the best feature in a given subset. This typically can result in a better model, which in this case it does. When calculating the RMSE for my random forest model I achieved a value of 18.92. Due to this model having the lowest RMSE, I continued the analysis of customer segmentation with this modeling technique.

Clusters

Although I used regression as the modeling technique for this project, I also tried an unsupervised learning model, K-Means Clustering. Clustering is a technique used to find groups that are not clearly labeled in the data. In order to find the number of clusters that this data set would most likely fit, I used the elbow technique.



From this graph we can see the kick off point is around a K value of 4 and 5. I chose a K value of 5 which represents the number of groups this data is most likely split into. I now continued to implement K-Means clustering analysis and this was my outcome:



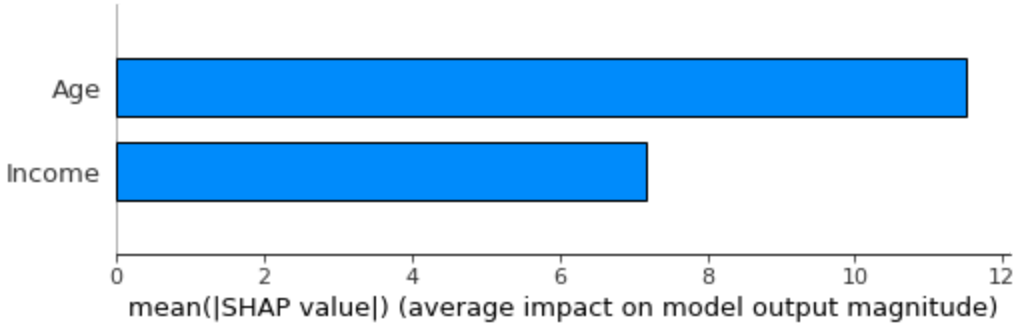
This clustering method is creating groups between income and spending score. From this graph I can see that 5 was the correct value for K and our data is split into the following groups:

1) Low Spending Score, Low Income
2) Low Spending Score, High Income
3) Average Spending Score, Average Income
4) High Spending Score, Low Income
5) High Spending Score, High Income

This data can be used by the mall if they ever wanted to target a product to a specific group. The mall would have to further breakdown the individual groups to understand what type of people they are composed of in order to better target the audience.

SHAP

Shap values is a helpful tool when working with a model that inputs features in order to predict an outcome. It allows us to understand what decisions the model is making and which feature has the biggest impact on the prediction of the outcome. The following bar graph shows us the features ranked in order that helped predict the spending score:



The plot shows that Age had more of an impact on spending score than income. Which correlates to the previous scatter plot that showed this correlation with the ages of 18-40 typically having a higher spending score than those of the ages 40+. Shap is a great tool and can help with the understanding of more complex data sets.

Conclusion

Through the life of the project and utilizing the modeling techniques I found that Random Forest Regression was the best modeling technique. It had the lowest RMSE meaning it was the best model to predict and future values depending on the independent variables. With further analysis using SHAP, it solidified my previous thoughts on the correlation between age and spending score which was visually represented by the scatter plot. Age had more of an impact on spending score than any other feature, but despite This now allows us to answer my initial question in trying to target what audience is most likely to walk out of the mall with a purchase. We can successfully target individuals between the ages of 20-40 and most likely a female