**Customer Demographic for Blackwell Electronics Part 2**

For this project we continued working with the demographic data for Blackwell Electronics. We began diving in deeper into analyzing the data about customer spending patterns in 4 different regions. To further understand the customer demographics and explain the spending patterns for Blackwell Electronics we utilized Machine Learning to build predictive models in order to help us understand whether:

* There is a difference in the age of customers between regions? If so, can we predict the age of a customer in a region based on other demographic data?
* There is any correlation between the age of a customer and if the transaction was made online or in the store? Are there any other factors that can aide in predicting if a customer will buy online or in our stores?

We were working with the same data sample of 80,000 observations. Collected data was divided by 4 different regions: 1- North, 2- South, 3- East and 4 -West. All the transactions were done either 1-In- store or 0-Online. In addition, each transaction had customer age, number of items purchased, and amount spent.

To answer the first question of whether or not there is a difference between the age of customers in different regions, we explored the data and found that the average age of customers is slightly different among the 4 different regions. In region 2-South, the average age of the customers was 56 years old, the highest among all of the regions. The lowest average age was in region 4-West, at 38 years old. The average age of the customers in region 1-North and region 3-East was similar, standing at 43 and 45 years old respectively. This information can be found in the following table:

Region Average Age

1 43.704132

2 56.609083

3 45.646944

4 38.752424

Name: age, dtype: float64

If we look at the minimum and maximum age of the customers in different regions, we see that in region 2- South the oldest customer was 85 years old and the youngest was 28 years old. This indicates some difference between the oldest and youngest customer in region 4- West, at 68 and 18 years old respectively. In region 1- North and 3- East, the age of the oldest and youngest customer is almost identical, where the youngest age for region 1-North is 18 and region 3-East is 19 and the oldest age is 74 for both regions. This information can be found in the table below.

| **Region** | **1-North** | **2-South** | **3-East** | **4-West** |
| --- | --- | --- | --- | --- |
| **Age max** | 74 | 85 | 74 | 63 |
| **Age min** | 19 | 28 | 18 | 18 |

Also, the visual representation of the difference in customers age in different regions can be found in the following graph.

Logo

Description automatically generated

To dive deeper into understanding whether the age of a customer can be predicted in a region based on the other demographic data, we found that using machine learning predictive models, the accuracy of our model stays relatively low. We discretized our age data in different ways and found that that there is little accuracy between our demographic data of independent variables such as region, age, amount spent, items and dependent variable age in all regions.

When we discretize age in 4 different bins like it is represented in the graph below:

Chart, bar chart

Description automatically generated

we achieved accuracy of our predictive DT model of only .41, which is exceptionally low. The information is represented in the table below:

precision recall f1-score support

17-35 0.38 0.16 0.22 7209

35-55 0.43 0.80 0.56 10447

55-68 0.21 0.05 0.08 3911

68-86 0.24 0.08 0.12 2427

accuracy 0.41 23994

macro avg 0.31 0.27 0.25 23994

weighted avg 0.36 0.41 0.34 23994

We ran 2 additional algorithms for our data with age as a dependable variable. We found that accuracy for Random Forest Classifier, when we split age in 4 different bins, was lower at .35. Gradient Boosting Classifier accuracy was at. 43.

We experimented with different variations of our model and found that if we split age in more bins or groups, then the accuracy of our age prediction decreases dramatically. For example, when we discretize age into 8 bins, the accuracy of the Decision Tree was at .19. The accuracy of our model for Random Forest Classifier and Gradient Boosting Classifier was at .17 and .20 respectively.

On another note, when we divided our age group between two bins (16-50 year old and 51-86 years old), we found that the accuracy of age prediction was at .68. See the table below:

Chart, histogram

Description automatically generated

precision recall f1-score support

51-86 0.57 0.44 0.50 7860

l6-50 0.76 0.84 0.79 16134

accuracy 0.71 23994

macro avg 0.66 0.64 0.65 23994

weighted avg 0.69 0.71 0.70 23994

When running the Random Forest Classifier, when we split age into 2 bins, we got an accuracy of .60. The Gradient Boosting Classifier accuracy was at .68, which was the same as the Decision Tree.

We still consider this accuracy very low to use for marketing purposes for Blackwell electronics. Whether or not dividing customers age into 2 different groups will prove useful for the company to market their products is up to the marketing department to decide.

To answer our second question of whether or not there is any relation between the age of the customers and whether transaction was done (in-store or online), we can look at the correlation table. From the results we see that there is a weak correlation between customers age and whether transactions were made online or in-store. The results can be found in the table below:

in-store age items amount region

in-store 1.000000 -0.178180 -0.003897 -0.085573 -0.133171

age -0.178180 1.000000 0.000657 -0.282033 -0.235370

items -0.003897 0.000657 1.000000 0.000384 -0.001904

amount -0.085573 -0.282033 0.000384 1.000000 0.403486

region -0.133171 -0.235370 -0.001904 0.403486 1.000000

We also looked at whether customers made purchases online or in-store can be predicted based on other demographic data. After running Decision Tree for this model, we found an accuracy of our model of .88. Therefore, we can tell that if we look at all the independent variables like amount, region, number of items purchased and age (all together), we can predict if the transaction was made online or in-store with a good accuracy of .88. See the table below:

precision recall f1-score support

0 0.96 0.79 0.87 11874

1 0.83 0.97 0.89 12120

accuracy 0.88 23994

macro avg 0.90 0.88 0.88 23994

weighted avg 0.90 0.88 0.88 23994

When running Random Forest Classifier, for in-store as a dependent variable, we got an accuracy of .86. Gradient Boosting Classifier was at .89 (almost the same as Decision Tree).

In conclusion, we found that there is a difference in age among customers in the 4 different regions. Region 2-South has the oldest average customers at 56 years old and region 4- West has the lowest average age of 38. Regions 1-North and region 3- West have an average age of 43 and 45 respectively. We found that we cannot predict a customer's age in different regions based on other demographic data. Our models show a small accuracy of .41. Therefore, we do not recommend using it for marketing purposes.

We also found no correlation between customer’s age and whether the purchase was made online or in-store. However, if we look at all of the independent variables like number of items purchased, amount spent, region and age together, we can predict if a purchase was made in-store or online with an accuracy of .88.

I think overall this exercise presents little information that can be used for marketing purposes for Blackwell Electronics, and we think that more information about customers should be collected such as household income, credit history, availability of product selection among different regions and in-store or online, etc.