Lexus Carton - Business Engineering Management Science

Nov 21, 2023

Intro. Data Science – Fall 2023

Project Update #2

This project has four parts: A random forest Logistic regression, Kmeans and limitations

Random forest:

As I discussed in my previous update, I was focused on mapping binary and categorical variables. I ended up working with the label encoder library to map a dictionary for the strings, and assigned numerical values. There were 22 columns in the data set when I did this for binary and Mult categorical variables. Below are examples. There was another stage of data cleaning. There were two columns that were Personal_Recomendation_frequency, one was renamed to Personalized_Recommendation_Frequency_words. The latter variable was numerical and had a scale of 1-5.

- Personalized_Recommendation_Frequency_words={"No":0, "Yes":1, "Sometimes":2}
- Improvement_Areas = {'Customer service responsiveness':0,"Product quality and accuracy":1,"Reducing packaging waste":2,"Shipping speed and reliability":3}
- Purchase_Categories={'Beauty and Personal Care':0,' Clothing and Fashion':1,"Groceries and Gourmet":2,"Home and Kitchen":3,"others":4}

Then, I was focused on the feature importance; my correlation table did not produce any values above 0.5, so I decided to do a random forest of all the variables. I was also able to create a random forest, one that was created through the Excel file's columns and my own using a random forest classifier. I had to play around with the parameters to get a good estimation for accuracy and precision.

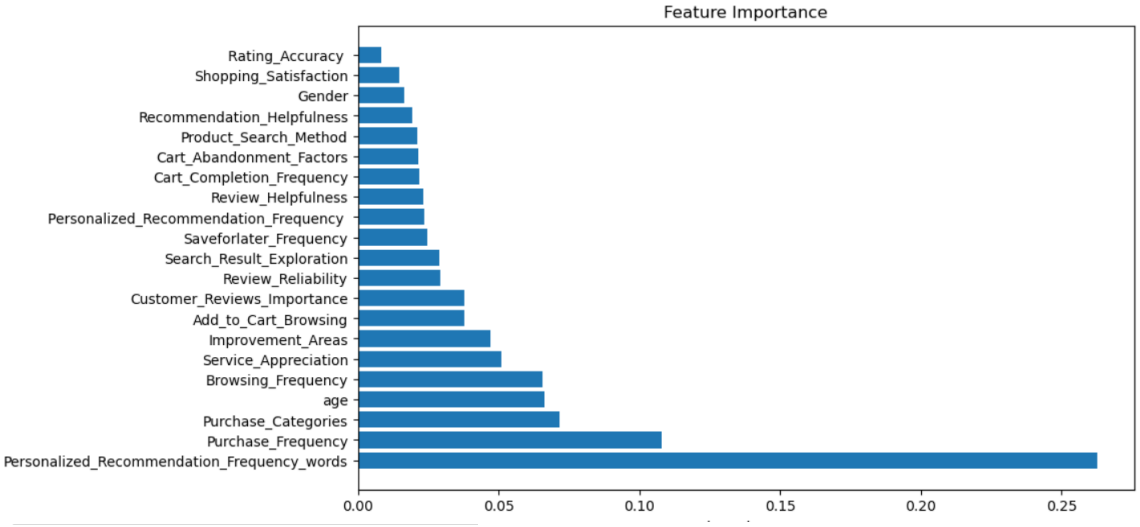
n_estimators=200, # Number of trees ,max_depth=5, min_samples_split=20, min_samples_leaf=15,I was able to see that purchase frequency was the enormous driving factor in consumer habits on Amazon.

```
Trainging accuracy 0.7307692307692307
Overal accuracy 0.6111111111111112
Confusion Matric [[13 26]
 [ 9 42]]
              precision    recall  f1-score   support

           0       0.59      0.33      0.43        39
           1       0.62      0.82      0.71        51

    accuracy                           0.61        90
   macro avg       0.60      0.58      0.57        90
weighted avg       0.61      0.61      0.58        90
```

The model was accurate, or near the mark, 61% of the time. But it had a recall score of 82%. The model correctly identified 13 positives and 42 false negatives. Nonetheless, it falsely identified 26 false positives and 9 false negatives.




```
Feature Importance (sorted by importance):
Personalized_Recommendation_Frequency_words: 0.2626
Purchase_Frequency: 0.1079
Purchase_Categories: 0.0716
age: 0.0662
Browsing_Frequency: 0.0654
Service_Appreciation: 0.0511
Improvement_Areas: 0.0470
Add_to_Cart_Browsing: 0.0379
Customer_Reviews_Importance: 0.0377
Review_Reliability: 0.0291
Search_Result_Exploration: 0.0288
Saveforlater_Frequency: 0.0247
Personalized_Recommendation_Frequency : 0.0237
Review_Helpfulness: 0.0233
Cart_Completion_Frequency: 0.0217
Cart_Abandonment_Factors: 0.0216
Product_Search_Method: 0.0211
Recommendation_Helpfulness: 0.0193
Gender: 0.0164
Shopping_Satisfaction: 0.0148
Rating_Accuracy : 0.0082
```

This displays the most probable way consumers would make their decisions whether or not to leave a review.

Logistic regression

From the top features I found, I wanted to try to run a logistic regression using Review_Left as my outcome variable. I only used the top 5 features to produce better metrics. The features to left are.

X=df[['Personalized_Recommendation_Frequency_words','Purchase_Frequency','Purchase_Categories',"Browsing_Frequency",'Customer_Reviews_Importance']]

Y=df[Reviw_Left]

The features to the right are.

#X = numeric_df.drop(['Review_Left'], axis=1)

#y = numeric_df['Review_Left']

```
Mean Squared Error (MSE): 0.23
Root Mean Squared Error (RMSE): 0.48
R-squared (R2): 0.02
Personalized_Recommendation_Frequency_words: 0.10
Purchase_Frequency: 0.03
Purchase_Categories: -0.05
Browsing_Frequency: 0.04
Customer_Reviews_Importance: 0.01
Intercept: 0.37
```

```
Mean Squared Error (MSE): 0.27
Root Mean Squared Error (RMSE): 0.52
R-squared (R2): -0.14
age: 0.00
Gender: 0.01
Purchase_Frequency: 0.01
Purchase_Categories: -0.05
Personalized_Recommendation_Frequency_words: 0.10
Browsing_Frequency: 0.04
Product_Search_Method: -0.05
Search_Result_Exploration: -0.14
Customer_Reviews_Importance: 0.07
Add_to_Cart_Browsing: -0.01
Cart_Completion_Frequency: 0.02
Cart_Abandonment_Factors: -0.02
Saveforlater_Frequency: 0.01
Review_Reliability: -0.05
Review_Helpfulness: -0.01
Personalized_Recommendation_Frequency : -0.07
Recommendation_Helpfulness: -0.01
Rating_Accuracy : -0.00
Shopping_Satisfaction: -0.05
Service_Appreciation: -0.04
Improvement_Areas: -0.06
Intercept: 0.96
```

One can see from the coefficients that Personalization and recommendation frequency is the dominating factor, and it drives purchase frequency. Ages are in the middle at zero and a few features are negative.

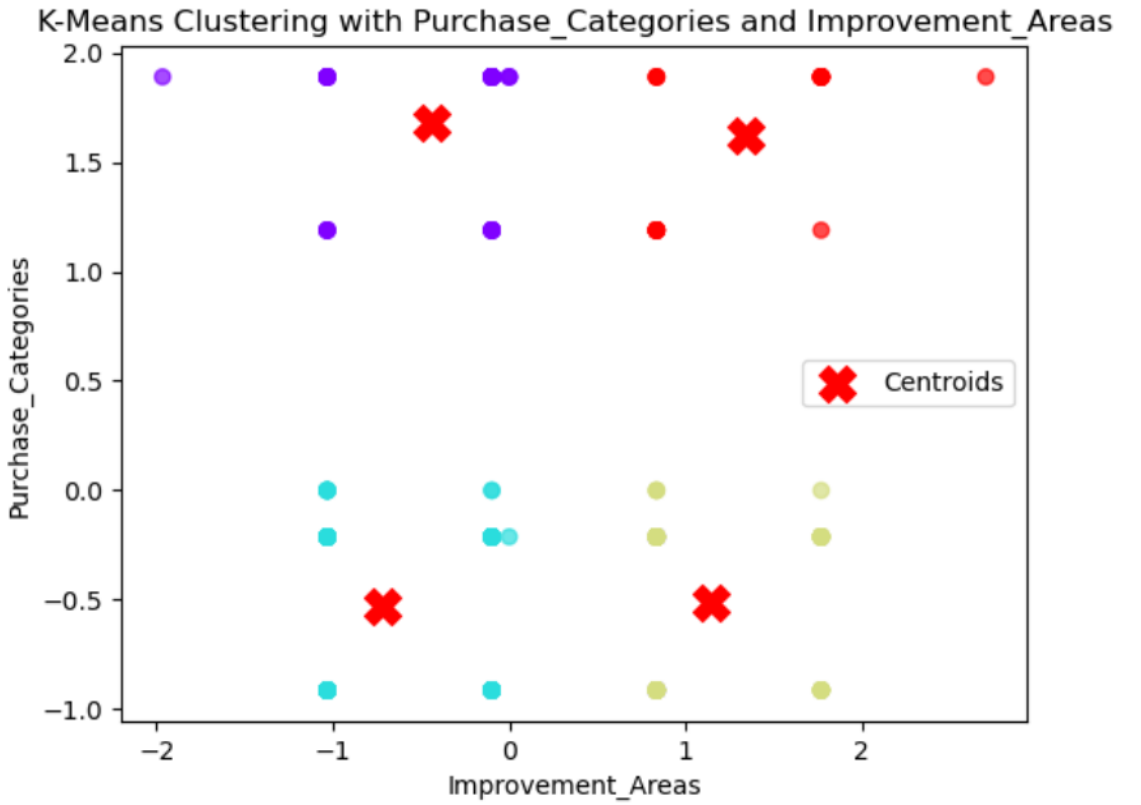
The use value is closer to 0 in the first sample than in the second one. The RMSE is also lower in the first model. The r-squared value is negative in both instances but is smaller in the first sample. This could be due to overfitting, or the model should be attempted with different predictors. In general, it doesn't follow the trend of the data. The first sample with five features has an r squared value of 0.02 which isn't great but it indicates that these features fit the data at a 2% level

K means clustering:

After both of these, K means clustering would be a good way to display what areas of purchase categories need improvement. There are only five mentioned in this filtered CSV file.

The features were scaled, so -2 for the x-axis is 1; this helped a lot, as when not scaled, all the centroids fell to the bottom of the graph. dictionary and stores values to certain keys. So the Purchase category associated with 3 is around the middle of the y axis of the graph., 0.5

- Purchase_Categories={'Beauty and Personal Care':0,'Clothing and Fashion':1,"Groceries and Gourmet":2,"Home and Kitchen":3,"others":4}
- Improvement_Areas = {'better app interface and lower shipping charges':0,'Customer service responsiveness':1,"Product quality and accuracy":2,"Reducing packaging waste":3,"Shipping speed and reliability":4,'Scrolling option would be much better than going to next page':5,"User interface":6}



- Green:
  - **Purchase_categories:** 'Beauty and Personal Care':0,'Clothing and Fashion':1,"Groceries and Gourmet":2,"
  - **Improvement areas:** 'Scrolling option would be much better than going to next page' User interface":6
- Orange:
  - **Purchase Categories :** ,"Home and Kitchen":3,"others":4}
  - **Improvement areas:** Shipping speed and reliability":4,'Scrolling option would be much better than going to next page':5,"User interface":6
- Purple
  - **Purchase Categories :,"**Home and Kitchen":3,"others":4}
  - **Improvement area :** 'better app interface and lower shipping charges':0,'Customer service responsiveness':1,"Product quality and accuracy":2,
- teal
  - **Purchase Categories :** 'Beauty and Personal Care':0,'Clothing and Fashion':1,"Groceries and Gourmet":2,

- **"Improvement area** : 'better app interface and lower shipping and 'Customer service responsiveness', Improvement areas: ,"Shipping speed and reliability

Limitations:

Consumer's minds are constantly changing and can be influenced by various factors. They can range from product availability, shopping out of boredom, or financial constraints. This Survey was only open for. Additionally, the Survey covered a period of 2023/06/04 to 2023/06/16. Having more time may show more trends in the data. This survey did not show where survey takers lived. If this information was provided then marketers can

While creating the Random Forest, the highest accuracy metric produced was 60% Recall and f1 score was relatively good for finding those with a review, as review =1 in this model. In the beginning, Purchase categories had multiple filters, which may help with the clustering methods as It is some wheat generic at this point.

Given that Personalization recommendation frequency was the highest contributing factor, One may assume that a customer need to accepts cookies, or make a Amazon account, in order to get the full benefit of shopping on the platform.

Future work

The purchase categories were filtered in the csv file into Beauty and Personal Care Beauty Personal, Clothing or Beauty Personal, Home, and Kitchen. Some responses were very exact; There was one product that was Beauty and Personal Care;Clothing and Fashion;Home and Kitchen while another was Beauty and Personal Care;Home and Kitchen;others. Purchase Categories could use this information to narrow in on specified products.

The purchase categories became 5 broad categories just to capture the target market, producers can look at the feature importance and invest resources in strategies such as customer service, marketing towards different generations, and personalization such as browsing habits. An algorithm can be implemented based on emailing and texting habit to offer better promotions to these customers.