IS-603 Decision-Making Support System

Topic: Recency Frequency Monetary Analysis

Submitted to: Dr. Faisal Quader

Submitted by:

Ashish Bhargav Gampa

Esha Singh

FNU Ayesha

Taruna Sanjay Pakhare

Tiago Souza Asakawa

TABLE OF CONTENT

1. ABSTRACT:

This project aims to analyze customer behavior and preferences in the e-commerce domain by employing the RFM (Recency, Frequency, Monetary Value) framework and machine learning techniques. The dataset utilized is obtained from an e-commerce platform and includes transactional data and customer attributes.

The project commences with essential data preprocessing steps, such as data cleaning, handling missing values, and formatting the dataset appropriately for analysis. Exploratory data analysis techniques are then applied to gain initial insights into the data, including visualizations of distributions, correlations, and summary statistics. The analysis primarily focuses on RFM calculation, which involves determining the Recency (time since last purchase), Frequency (number of purchases), and Monetary Value (total spending) for each customer. The RFM values are computed using suitable methods and algorithms. The dataset is subsequently subjected to k-means clustering, an unsupervised learning algorithm, to segment customers based on their RFM values and other pertinent features. Through experimentation, the optimal number of clusters is determined, and the resulting segments are thoroughly examined and compared to gain a deeper understanding of customer behavior and preferences within each segment. To assess the significance of different features in predicting customer behavior, decision tree algorithms like CART or Random Forest are utilized.

These algorithms analyze the importance of RFM variables and other attributes in influencing outcomes such as repeat purchase or churn. The insights obtained from the analysis are translated into actionable strategies, such as targeted marketing campaigns and customer retention programs. The project concludes with an evaluation of the analysis outcomes, considering metrics like customer retention rate, repeat purchase rate, and revenue growth. It is recommended to continuously monitor and adapt the strategies based on changes in customer behavior and market dynamics.

In summary, this project provides a comprehensive approach to understanding customer behavior and preferences in the e-commerce domain through RFM analysis, k-means clustering, and decision tree algorithms. The insights gained from this analysis can assist businesses in optimizing their marketing efforts, enhancing customer satisfaction, and driving growth in the competitive e-commerce landscape.

2. INTRODUCTION:

RFM analysis allows companies to categorize their clients according to how frequently, how much, and how recently they make purchases. This assists organizations in better understanding their clients, allowing them to develop stronger marketing tactics and keep their customers returning. RFM analysis is crucial since it gives organizations an understanding of the preferences and behavior of their customers.

Businesses can determine which of their most valued and devoted customers as well as those who are in danger of leaving by segmenting their customer base depending on recency, frequency, and financial value. To keep current clients and draw in new ones, this information can then be used to design focused marketing tactics and customized promotions. In the end, RFM analysis can benefit firms by helping them retain more clients, boost client happiness, and boost sales.

The RFM model is based on three quantitative factors: Recency: How recently a customer made a purchase Frequency: How often a customer makes a purchase Monetary value: How much a customer spends on a purchase. These three RFM factors can be used to reasonably predict how likely (or unlikely) it is that a customer will do business again with an organization.RFM analysis allows one to compare potential suppliers and customers. This helps the organization understand how much revenue can be generated from repeating customers (versus new customers) and what steps can be taken to make the new customers; satisfied and repeat customers.

Recency, Frequency, and Monetary (RFM) analysis is a significant topic in the field of data science and marketing for several reasons:

- Improved Customer Segmentation: RFM analysis enables a business to segment its customers based on certain characteristics, such as purchasing behavior of customers. This allows a business to identify the most valuable customer segment and alter its marketing strategies based on these segments, which increases customer retention and loyalty.
- Increased customer retention and loyalty: As mentioned earlier, targeting the right segment of customers enables a business to offer targeted promotions and incentives, which increases the repetition of customers and increases overall sales, this ensures that customer loyalty and retention increases.

- Improved marketing: RFM analysis ensures that the marketing strategies are tailored according to the targeted segments of customers which ascertains that the marketing strategy applied is effective and this increases overall profitability.
- Real time Analysis: RFM analysis enables a firm to perform analysis in real time, which ensures that a business quickly identifies changes in customer behavior and alter their marketing strategies accordingly. By doing so, businesses can stay ahead of the competition and respond and adapt quickly to the changes in the market.

Overall, RFM analysis is a significant topic in the field of data science and marketing because it allows businesses to better understand their customers' behavior, increase customer retention and loyalty, improve marketing ROI, and stay ahead of the competition.

3. BACKGROUND & RELATED WORK:

There have been many studies and research papers published on the topic of RFM analysis, exploring different aspects of the technique and its effectiveness. For instance, a study published in the International Journal of Research and Analytical Reviews in 2019 examined a case study of a retail company that used RFM analysis to segment its customers and create targeted marketing plans. The study discovered that the strategy led to appreciable increases in client retention and sales. Another study published in the Proceedings of the 2018 International Conference on Management, Education, Information and Control presented a theoretical framework for applying RFM analysis to customer value analysis. In addition to these studies, there are a variety of platforms and software solutions available that can automate the RFM analysis process and make it simpler for organizations to use the methodology. For instance, many e-commerce platforms include integrated RFM analysis tools that let companies segment their clientele depending on consumer behavior and create specialized marketing strategies. RFM analysis is a well-known marketing strategy that has been extensively used and researched in both the academic and corporate arenas. The usage of RFM analysis is anticipated to grow increasingly more popular and sophisticated as companies continue to gather more data on their customers, and new methods and tools are being created to boost its efficacy.

4. RESEARCH METHODS:

K-means Clustering

RFM analysis, as mentioned earlier, is a technique used in marketing to segment customers based on their purchase behavior. Meanwhile, K-means clustering is a popular machine-learning approach for combining comparable data points. . Therefore, we can combine these two techniques and create meaningful customer segments based on their RFM scores.

To use k-means clustering in RFM analysis, we first need to calculate the RFM scores for each customer in our dataset. Recency refers to how recently a customer made a purchase, Frequency refers to how often a customer makes a purchase, and Monetary refers to how much money a customer spends on each purchase. Once we have calculated the RFM scores for each customer, we can use the k-means clustering algorithm to group similar customers together based on their RFM scores.

Decision Tree

Decision trees are a powerful tool in data science and machine learning that can be used in RFM analysis to help identify patterns in customer behavior and segment customers based on their RFM scores. A decision tree is a tree-like model that represents a set of decisions and their possible consequences.

To use decision trees in RFM analysis, we can build a decision tree model that predicts customer behavior based on their RFM scores. The decision tree can be used to segment customers into different groups based on their RFM scores, which can then be used to tailor marketing strategies to each group.

Logistic Regression

Analyzing the relationship between a dependent variable and one or more independent variables is achieved statistically by using logistic regression. When there are just two possible values (binary) for the dependent variable in binary classification problems are common.

To use logistic regression in RFM analysis, based on a customer's RFM scores we can determine whether they are likely to make a purchase in the future. To do this, the RFM scores of the

customers would serve as the independent variables, and the dependent variable would be a binary variable indicating whether the consumer made a purchase during the specific time period. This logistic regression model will generate a probability score for each consumer indicating their likelihood of making a purchase and will help detect a probable customer churn.

Customer Lifetime Value(CLV)

Customer lifetime value is an estimation of the overall contribution to the company over the course of their lifetime. It considers the entire customer's lifetime previous transactions, the purchase amounts, and the frequency of the customer.

We can build the CLV model using RFM analysis by calculating the average purchase value, purchase frequency, customer lifespan, and retention rate. By multiplying all values with each other. businesses can more clearly identify which customers are the most valuable by calculating the CLV for each one using RFM research, and they can then concentrate their marketing and sales efforts on those clients. Over time, this may result in more sales and customer satisfaction.

5. METHODOLOGY:

I. **Data Preprocessing:** Download the dataset from the provided Kaggle link and save it in a compatible format for Weka, such as CSV or ARFF. Load the dataset into Weka using the Preprocess tab or the command-line interface.

II. **Exploratory Data Analysis:** Utilize Weka's data visualization capabilities to explore the dataset, including histograms, scatter plots, and summary statistics. Identify any missing values or outliers in the data and determine the appropriate handling method, such as imputation or removal.

III. **RFM Calculation:** Utilize Weka's data manipulation features, such as AttributeFilter or AttributeExpression, to calculate the RFM values for each customer. Create new attributes for Recency, Frequency, and Monetary Value based on the provided formulas.

IV. **Data Transformation:** If required, preprocess the data further by applying transformations like normalization or standardization to enhance model performance. Weak offers various filter options, such as Normalize or Standardize, for data transformation.

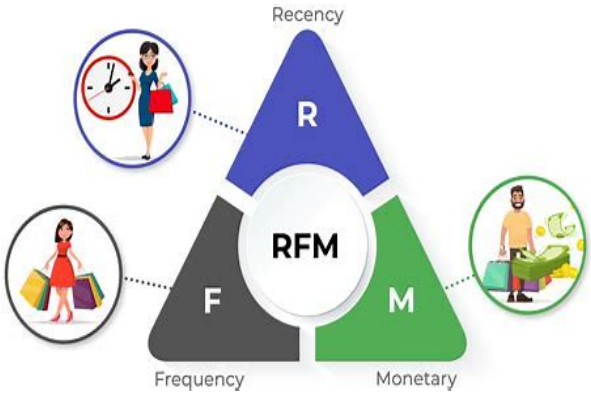
V. **Segmentation Analysis:** Apply clustering algorithms available in Weka, such as k-means or hierarchical clustering, to segment customers based on their RFM values. Experiment with different numbers of clusters and assess their quality using metrics like silhouette score or within-cluster sum of squares.

VI. **Association Rules**: Utilize Weka's association rule mining algorithms, such as Apriori or FPGrowth, to discover frequent itemsets and association rules from customer purchases. Adjust the support and confidence thresholds to extract meaningful rules that reveal patterns in customer behavior.

VII. **Predictive Modeling:** Apply supervised learning algorithms provided by Weka, such as decision trees, random forests, or logistic regression, to predict customer behavior or preferences. By using logistic regression we are calculating RFM rank of the customer and then calculating the probability score of purchases of the customerUtilize the RFM values as features and the target variable (e.g., repeat purchase, churn) as the class label. Evaluate and compare the performance of different models using appropriate evaluation measures like accuracy, precision, recall, or AUC.

VIII. **Retention Strategies:** Based on the insights gained from the analysis, develop targeted retention strategies for each customer segment. Utilize Weka's classification capabilities to classify new customers into appropriate segments and apply the corresponding retention strategies.

IX. **Evaluation and Monitoring:** Evaluate the effectiveness of the retention strategies using appropriate metrics such as customer retention rate, repeat purchase rate, or revenue growth. Continuously monitor the segments and update the strategies as needed based on changes in customer behavior.



6. RESULTS:

File Edit View — ARFF-Viewer - C:\Users\14432\Downloads\ecom_data_rfm.csv

ecom_data_rfm.csv

Relation: ecom_data_rfm

No.	1: Numeric	2: CustomerID Numeric	3: Frequency Numeric	4: Recency Numeric	5: Monetary Numeric	6: rankR Numeric	7: rankF Numeric	8: rankM Numeric	9: groupRFM Numeric	10: Country Nominal	11: **Customer_Segment** Nominal
1	1.0	12346.0	2.0	358.0	2.08	2.0	1.0	1.0	211.0	United Kin...	Lost Lowest
2	2.0	12347.0	182.0	35.0	481.21	5.0	4.0	3.0	543.0	Iceland	Loyal Customers
3	3.0	12348.0	31.0	108.0	178.71	5.0	1.0	2.0	512.0	Finland	Potential Loyalist
4	4.0	12349.0	73.0	51.0	605.1	5.0	2.0	4.0	524.0	Italy	Recent High Spender
5	5.0	12350.0	17.0	343.0	65.3	2.0	1.0	1.0	211.0	Norway	Lost Lowest
6	6.0	12352.0	95.0	69.0	2211.1	5.0	2.0	5.0	525.0	Norway	NULL
7	7.0	12353.0	4.0	237.0	24.3	3.0	1.0	1.0	311.0	Bahrain	About To Sleep
8	8.0	12354.0	58.0	265.0	261.22	3.0	1.0	2.0	312.0	Spain	About To Sleep
9	9.0	12355.0	13.0	247.0	54.65	3.0	1.0	1.0	311.0	Bahrain	About To Sleep
10	10.0	12356.0	59.0	55.0	188.87	5.0	2.0	2.0	522.0	Portugal	Potential Loyalist
11	11.0	12357.0	131.0	66.0	438.67	5.0	3.0	3.0	533.0	Switzerland	Potential Loyalist
12	12.0	12358.0	19.0	34.0	157.21	5.0	1.0	1.0	511.0	Austria	New Customers
13	13.0	12359.0	254.0	40.0	2225.11	5.0	5.0	5.0	555.0	Cyprus	Loyal Customers
14	14.0	12360.0	129.0	85.0	457.91	5.0	3.0	3.0	533.0	Austria	Potential Loyalist
15	15.0	12361.0	10.0	320.0	33.35	2.0	1.0	1.0	211.0	Belgium	Lost Lowest
16	16.0	12362.0	274.0	36.0	1083.29	5.0	5.0	5.0	555.0	Belgium	Loyal Customers
17	17.0	12363.0	23.0	142.0	53.17	4.0	1.0	1.0	411.0	Unspecified	Promising
18	18.0	12364.0	85.0	40.0	162.37	5.0	2.0	1.0	521.0	Belgium	Potential Loyalist
19	19.0	12365.0	23.0	324.0	698.0	2.0	1.0	5.0	215.0	Cyprus	NULL
20	20.0	12367.0	11.0	37.0	35.2	5.0	1.0	1.0	511.0	Denmark	New Customers
21	21.0	12370.0	167.0	84.0	467.65	5.0	3.0	3.0	533.0	Cyprus	Potential Loyalist
22	22.0	12370.0	167.0	84.0	467.65	5.0	3.0	3.0	533.0	Austria	Potential Loyalist
23	23.0	12371.0	63.0	77.0	244.08	5.0	2.0	2.0	522.0	Switzerland	Potential Loyalist
24	24.0	12372.0	52.0	104.0	156.07	5.0	1.0	1.0	511.0	Denmark	New Customers
25	25.0	12373.0	14.0	344.0	64.15	2.0	1.0	1.0	211.0	Austria	Lost Lowest
26	26.0	12374.0	33.0	58.0	139.25	5.0	1.0	1.0	511.0	Austria	New Customers
27	27.0	12375.0	18.0	35.0	119.6	5.0	1.0	1.0	511.0	Finland	New Customers
28	28.0	12377.0	77.0	348.0	209.35	2.0	2.0	2.0	222.0	Switzerland	Lost Lowest

Search — 8:45 PM 4/17/2023

Fig. Dataset

Preprocess Classify Cluster Associate Select attributes Visualize

Classifier

Choose **SimpleLogistic** -I 0 -M 500 -H 50 -W 0.0

Test options

- Use training set
- Supplied test set — Set...
- Cross-validation — Folds 10
- Percentage split — % 66

More options...

(Nom) Customer_Segment

Start Stop

Result list (right-click for options)

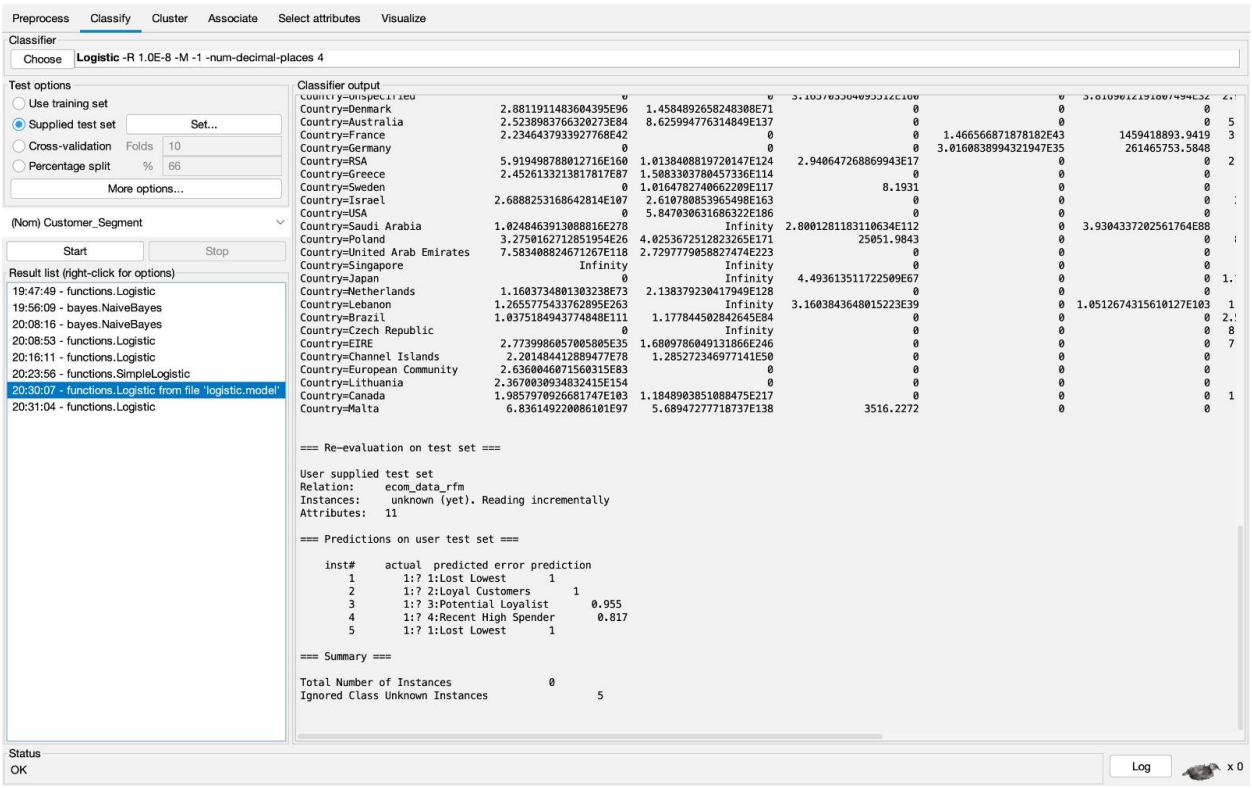
- 19:47:49 - functions.Logistic
- 19:56:09 - bayes.NaiveBayes
- 20:08:16 - bayes.NaiveBayes
- 20:08:53 - functions.Logistic
- 20:16:11 - functions.Logistic
- 20:23:56 - functions.SimpleLogistic

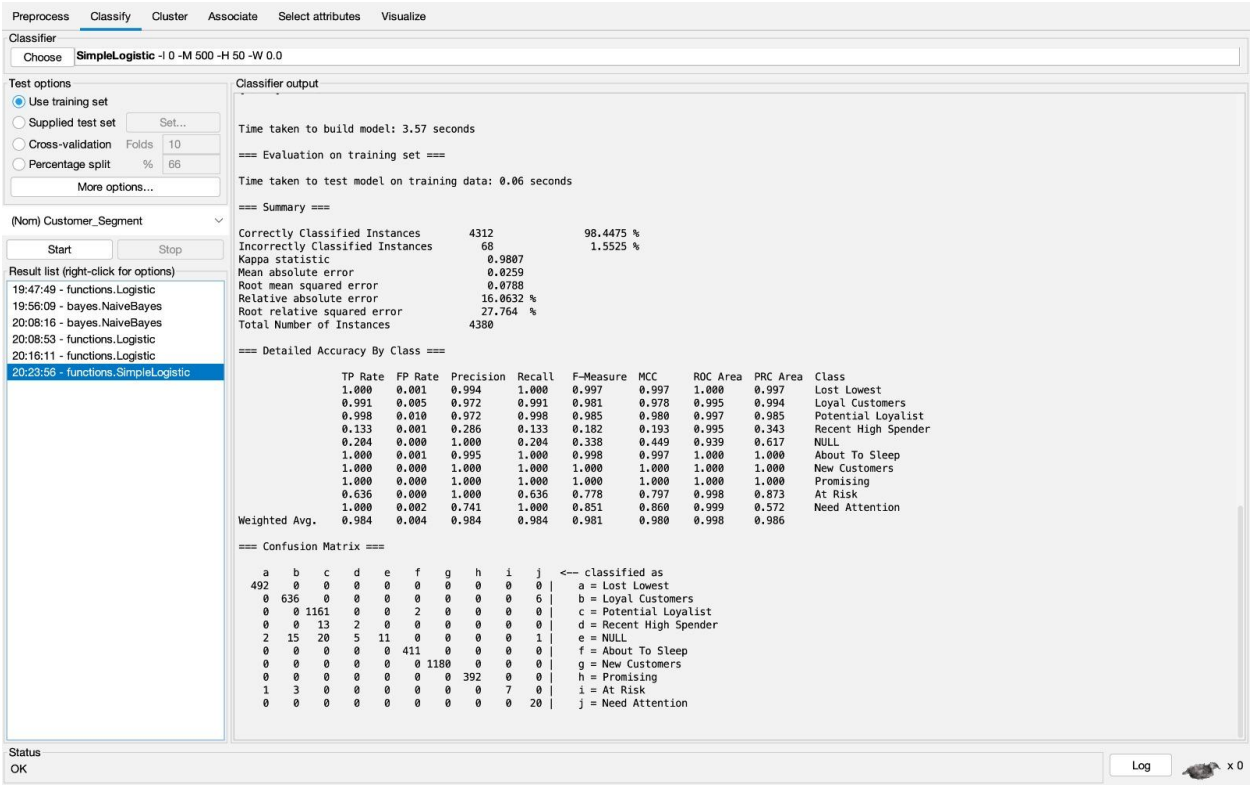
Classifier output

```
Time taken to build model: 0.04 seconds

=== Evaluation on training set ===

Time taken to test model on training data: 0.23 seconds

=== Summary ===

Correctly Classified Instances        4266               97.3973 %
Incorrectly Classified Instances       114                2.6027 %
Kappa statistic                          0.9678
Mean absolute error                      0.006
Root mean squared error                  0.0677
Relative absolute error                  3.7247 %
Root relative squared error             23.8644 %
Total Number of Instances             4380

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.982    0.000    1.000      0.982    0.991      0.990    1.000     1.000     Lost Lowest
                 0.944    0.002    0.989      0.944    0.966      0.960    0.987     0.980     Loyal Customers
                 0.979    0.006    0.983      0.979    0.981      0.974    0.998     0.997     Potential Loyalist
                 1.000    0.001    0.833      1.000    0.909      0.913    1.000     1.000     Recent High Spender
                 0.444    0.007    0.429      0.444    0.436      0.429    0.975     0.348     NULL
                 0.966    0.000    1.000      0.966    0.983      0.981    1.000     1.000     About To Sleep
                 1.000    0.000    0.999      1.000    1.000      0.999    1.000     1.000     New Customers
                 1.000    0.005    0.954      1.000    0.976      0.974    1.000     1.000     Promising
                 1.000    0.003    0.478      1.000    0.647      0.691    1.000     1.000     At Risk
                 1.000    0.005    0.500      1.000    0.667      0.705    1.000     0.995     Need Attention
Weighted Avg.    0.974    0.003    0.978      0.974    0.975      0.972    0.997     0.988

=== Confusion Matrix ===

   a   b   c   d   e   f   g   h   i   j   <-- classified as
 483   0   0   0   0   0   0   0   9   0 |   a = Lost Lowest
   0 606   2   0  30   0   0   0   0   4 |   b = Loyal Customers
   0   0 1138   1   2   0   1  19   0   2 |   c = Potential Loyalist
   0   0   0  15   0   0   0   0   0   0 |   d = Recent High Spender
   0   7  18   2  24   0   0   0   3   0 |   e = NULL
   0   0   0   0   0 397   0   0   0  14 |   f = About To Sleep
   0   0   0   0   0   0 1180   0   0   0 |   g = New Customers
   0   0   0   0   0   0   0 392   0   0 |   h = Promising
   0   0   0   0   0   0   0   0  11   0 |   i = At Risk
   0   0   0   0   0   0   0   0   0  20 |   j = Need Attention
```

Status OK — Log — x 0

Fig. Classification


Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Classifier
Choose
Logistic -R 1.0E-8 -M -1 -num-decimal-places 4
Test options
Use training set
Supplied test set
Set...
Cross-validation
Folds
10
Percentage split
%
66
More options...
(Nom) Customer_Segment
Start
Stop
Result list (right-click for options)
19:47:49 - functions.Logistic
19:56:09 - bayes.NaiveBayes
20:08:16 - bayes.NaiveBayes
20:08:53 - functions.Logistic
20:16:11 - functions.Logistic
20:23:56 - functions.SimpleLogistic
20:30:07 - functions.Logistic from file 'logistic.model'
20:31:04 - functions.Logistic
Classifier output
Country=Denmark 2.8811911483604395E96 1.4584892658248308E71 0 0 0
Country=Australia 2.523898376632027 3E84 8.625994776314849E137 0 0 0 5
Country=France 2.2346437933927768E42 0 0 1.4665668718781 82E43 1459418893.9419 3
Country=Germany 0 0 0 3.0160838994321947E35 261465753.5848
Country=RSA 5.919498788012716E160 1.0138408819720147E124 2.940647268869943E17 0 0 2
Country=Greece 2.4526133213817817E87 1.5083303780457336E114 0 0 0
Country=Sweden 0 1.0164782740662209E117 8.1931 0 0
Country=Israel 2.6888253168642814E107 2.610780853965498E163 0 0 0
Country=USA 0 5.847030631686322E186 0 0 0
Country=Saudi Arabia 1.0248463913088816E278 Infinity 2.8001281183110634E112 0 3.9304337202561764E88
Country=Poland 3.275016271285 1954E26 4.025367251282326 5E171 25051.9843 0 0
Country=United Arab Emirates 7.583408824671267E118 2.7297779058827474E223 0 0 0
Country=Singapore Infinity Infinity 0 0 0
Country=Japan 0 Infinity 4.493613511722509E67 0 0 1.
Country=Netherlands 1.1603734801303238E73 2.138379230417949E128 0 0 0
Country=Lebanon 1.2655775433762895E263 Infinity 3.1603843648015223E39 0 1.0512674315610127E103 1
Country=Brazil 1.0375184943774848E111 1.1778445028426 45E84 0 0 0 2.
Country=Czech Republic 0 Infinity 0 0 0 8
Country=EIRE 2.7739986057005805E35 1.6809786049131866E246 0 0 0 7
Country=Channel Islands 2.201484412889477E78 1.285272346977141E50 0 0 0
Country=European Community 2.6360046071560315E83 0 0 0 0
Country=Lithuania 2.3670030934832415E154 0 0 0 0
Country=Canada 1.9857970926681747E103 1.1848903851088475E217 0 0 0 1
Country=Malta 6.836149220086101E97 5.68947277718737E138 3516.2272 0 0
=== Re-evaluation on test set ===
User supplied test set
Relation: ecom_data_rfm
Instances: unknown (yet). Reading incrementally
Attributes: 11
=== Predictions on user test set ===
inst# actual predicted error prediction
1 1:? 1:Lost Lowest 1
2 1:? 2:Loyal Customers 1
3 1:? 3:Potential Loyalist 0.955
4 1:? 4:Recent High Spender 0.817
5 1:? 1:Lost Lowest 1
=== Summary ===
Total Number of Instances 0
Ignored Class Unknown Instances 5
Status
OK
Log
x 0


Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Classifier
Choose
SimpleLogistic -I 0 -M 500 -H 50 -W 0.0
Test options
Use training set
Supplied test set
Set...
Cross-validation
Folds
10
Percentage split
%
66
More options...
(Nom) Customer_Segment
Start
Stop
Result list (right-click for options)
19:47:49 - functions.Logistic
19:56:09 - bayes.NaiveBayes
20:08:16 - bayes.NaiveBayes
20:08:53 - functions.Logistic
20:16:11 - functions.Logistic
20:23:56 - functions.SimpleLogistic
Classifier output
Time taken to build model: 3.57 seconds
=== Evaluation on training set ===
Time taken to test model on training data: 0.06 seconds
=== Summary ===
Correctly Classified Instances 4312 98.4475 %
Incorrectly Classified Instances 68 1.5525 %
Kappa statistic 0.9807
Mean absolute error 0.0259
Root mean squared error 0.0788
Relative absolute error 16.0632 %
Root relative squared error 27.764 %
Total Number of Instances 4380
=== Detailed Accuracy By Class ===
TP Rate FP Rate Precision Recall F-Measure MCC ROC Area PRC Area Class
1.000 0.001 0.994 1.000 0.997 0.997 1.000 0.997 Lost Lowest
0.991 0.005 0.972 0.991 0.981 0.978 0.995 0.994 Loyal Customers
0.998 0.010 0.972 0.998 0.985 0.980 0.997 0.985 Potential Loyalist
0.133 0.001 0.286 0.133 0.182 0.193 0.995 0.343 Recent High Spender
0.204 0.000 1.000 0.204 0.338 0.449 0.939 0.617 NULL
1.000 0.001 0.995 1.000 0.998 0.997 1.000 1.000 About To Sleep
1.000 0.000 1.000 1.000 1.000 1.000 1.000 1.000 New Customers
1.000 0.000 1.000 1.000 1.000 1.000 1.000 1.000 Promising
0.636 0.000 1.000 0.636 0.778 0.797 0.998 0.873 At Risk
1.000 0.002 0.741 1.000 0.851 0.860 0.999 0.572 Need Attention
Weighted Avg. 0.984 0.004 0.984 0.984 0.981 0.980 0.998 0.986
=== Confusion Matrix ===
a b c d e f g h i j <-- classified as
492 0 0 0 0 0 0 0 0 0 | a = Lost Lowest
0 636 0 0 0 0 0 0 0 6 | b = Loyal Customers
0 0 1161 0 0 2 0 0 0 0 | c = Potential Loyalist
0 0 13 2 0 0 0 0 0 0 | d = Recent High Spender
2 15 20 5 11 0 0 0 0 1 | e = NULL
0 0 0 0 0 411 0 0 0 0 | f = About To Sleep
0 0 0 0 0 0 1180 0 0 0 | g = New Customers
0 0 0 0 0 0 0 392 0 0 | h = Promising
1 3 0 0 0 0 0 0 7 0 | i = At Risk
0 0 0 0 0 0 0 0 0 20 | j = Need Attention
Status
OK
Log
x 0

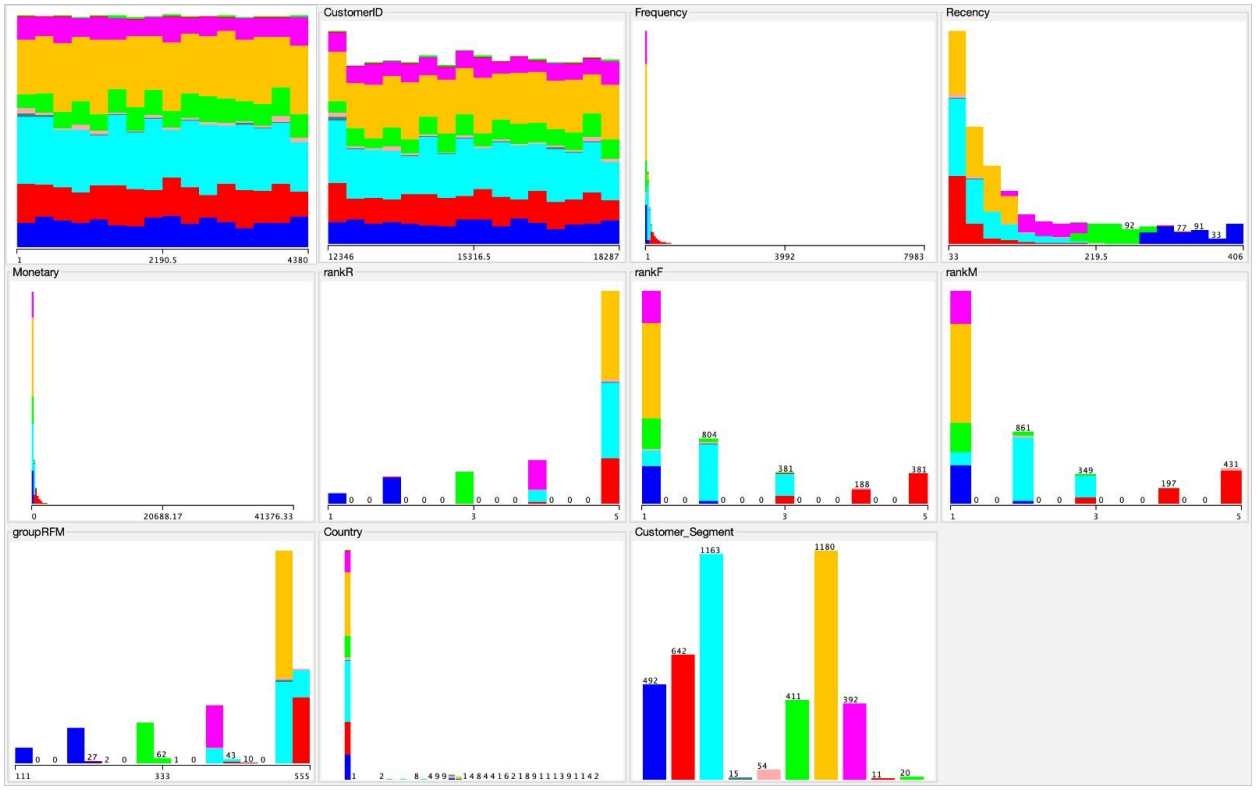
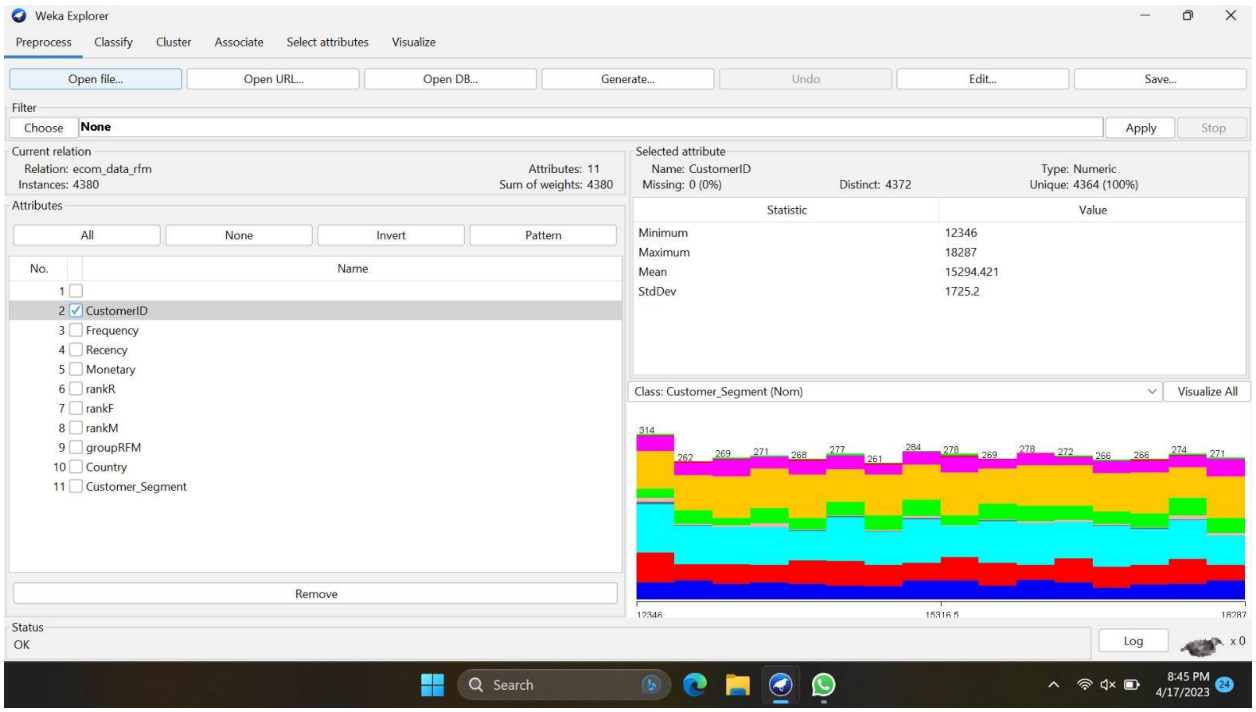


Fig Classification based on Logistics Regression



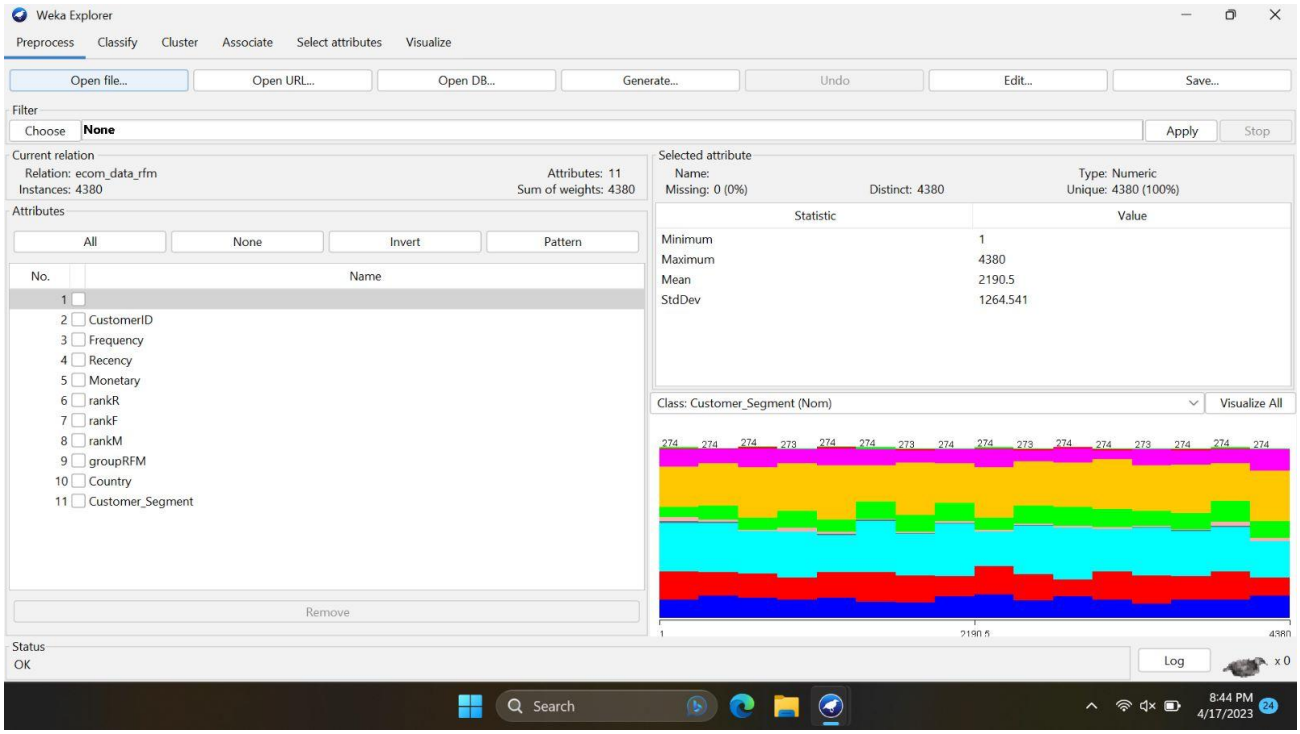


Fig. Analysis of Dataset using Weka

7. CONCLUSION

Develop targeted retention strategies for each customer segment based on the insights gained from the analysis. Use A/B testing or other techniques to test the effectiveness of the strategies and optimize them over time. Monitor the results and adjust Summarize the findings of the analysis and their implications for the business. Provide recommendations for future research or improvements to the analysis or retention strategies.

8. FUTURE WORK:

To gain a deeper understanding of customer behavior and preferences, the project can employ customer segmentation analysis using variations of the k-means clustering algorithm. This approach involves dividing customers into distinct groups or segments based on similarities in their RFM values (Recency, Frequency, Monetary Value) and other relevant features.

By experimenting with different numbers of clusters, the project can explore the optimal configuration that best captures the underlying patterns and heterogeneity within the customer base. The k-means algorithm will iteratively assign customers to clusters, with each cluster representing a distinct segment. This enables the identification of different customer groups based on their purchasing behavior and engagement with the e-commerce platform.

Once the customer segments are established, further analysis can be conducted to understand the characteristics and preferences of each segment. This includes examining the RFM values and other relevant features within each segment, such as demographic information or product preferences. By comparing and contrasting the segments, the project can uncover valuable insights regarding customer needs, preferences, and behaviors.

In addition to customer segmentation, decision tree algorithms like CART (Classification and Regression Trees) or Random Forest can be applied to determine the importance of different features in predicting customer behavior. These algorithms create a hierarchical structure of decision rules based on feature thresholds to predict outcomes such as repeat purchase or churn.

The decision tree algorithms can assess the significance of various features, including the RFM variables, in influencing customer behavior. By analyzing the decision rules and feature importance metrics generated by these algorithms, the project can identify the most influential factors that drive customer outcomes. This information can guide strategic decision-making, such as designing targeted marketing campaigns or implementing personalized retention strategies.

In summary, by combining customer segmentation analysis using k-means clustering and feature importance analysis using decision tree algorithms like CART or Random Forest, the project can gain a comprehensive understanding of customer behavior and preferences. This knowledge can inform data-driven strategies and decision-making processes to enhance customer satisfaction, retention, and overall business performance in the e-commerce domain.

9. REFERENCES:

- Murphy, C. (2011, February 14). *What is recency, frequency, monetary value (RFM) in marketing?* Investopedia.
- Doğan, O., Ayçin, E., & Bulut, Z. A. (n.d.). *Customer segmentation by using rfm model and clustering methods: A case study in retail industry*. Core.ac.uk. Retrieved April 10, 2023.
- Jain, A. K., Murty, M. N., & Flynn, P. J. (1999). *Data clustering: A review. ACM Computing Surveys* (CSUR), 31(3), 264-323.
- Han, J., Kamber, M., & Pei, J. (2011). *Data mining: concepts and techniques.* Morgan Kaufmann.
- Witten, I. H., Frank, E., Hall, M. A., & Pal, C. J. (2016). *Data mining: practical machine learning tools and techniques.* Morgan Kaufmann.
- Peacock, M., & Lapsley, D. (2019). *Marketing analytics: A practical guide to improving consumer insights using data techniques.* Kogan Page.