Data Mining Approach to Target Customers for Marketing Campaigns Based on Customer Segmentation

Venkat Vijay Gudapati, Mahidhar Reddy Narala, Mauricio Espinoza
CAP 5571, Department of Computer Science, Florida State University

1. Introduction

With rapid growth in technology and increase in the data stored by the enterprises, it has changed how they manage their marketing strategies. In the recent times we have observed how highly competitive Ecommerce industry has become and how difficult it is to retain an existing customer as well as to acquire a new customer. To design marketing strategies and design advertisements company should be able to understand characteristics of their customers. Enterprises are investing in CRM enabled with Big Data to store customer related data which can be used in identifying needs, preferences and behavior of their customers. Customer segmentation is a process of identifying customers with similar characteristics and club them in the same group. These characteristics can be obtained from customer's behavior such as transaction frequency and monetary value. Clustering is the most common method in customer segmentation literatures. They are developed with similarity and distance so that the data with similar characteristics is grouped near to each other.

The main goal of this project is to develop a comprehensive algorithm to group the customers into well-defined groups and build a real time predictive machine learning model to predict which group a customer belongs to. Customer segmentation is done through behavioral characteristics such as recency, frequency, and monetary value.

2. Literature Survey

Initially proposed by Arthur Hughes, the RFM (Recency Frequency Monetary) model is widely used to understand customer behavior. The model refers to the "recency of last purchase, purchase frequency, and monetary value of purchase." Previous research has shown that the greater the *R* and *F* value, the more likely a customer will create a new transaction with the company or business. The *M* value corresponds to the likelihood of a customer conducting repeated interactions with a company again. Previous research suggests each variable could have differing importance based on industries conflicting with Hughes's idea of equality (Wu, et al., 2020).

Customer segmentation is used to provide more "personalized" marketing strategies to customers. Clustering, or grouping a set of objects (customers) in such a way that objects in the same group, is used to achieve customer segmentation. However, there are some issues with using clustering algorithms (Maulina, Surjandari, & Rus). In the case of Maulina's research, CLARA AND PAM algorithms were used to find the optimum number of clusters ranging from k=2 to k=6. By comparing the Dunn and Silhouette score, it was found that k = 3 was the optimum number of clusters.

In the paper by Li, the authors noted that the widely used k-means algorithm is sensitive to the initial set of selected centroids. Thus, the quality of the final clusters can vary widely depending on the centroids' selection. The authors propose an improved version of the k-means algorithm to mitigate the drawbacks of the original k-means algorithm. Their research suggests the improved algorithm improves customer segmentation accuracy (Li, 2013).

Customer Segmentation using Centroid Based and Density Based Clustering Algorithms,

compares k-means and density-based clustering algorithms (DBSCAN) for customer segmentation. In the article, it is shown that DBSCAN provides more meaningful customer segmentation than k-means (Hossain, 2017).

However, in our chosen paper, the authors compare centroid-based algorithms with Fuzzy C Means to find the optimum cluster number.

3. Methodology

Data is in Semi Structured format where we do not have labels for the data points. In order to create the labels, we have used unsupervised learning algorithms such as K-Means, K-Medoids. Another method we have experimented with is RFM (Recency, Frequency, Monetary) model.

Before creating the segments, we have to clean the data to correct any abnormalities and null values. After creating segments based on the unsupervised learning models and RFM model. We have compared the results of the models to see which model is able to cluster the customers appropriately.

Once the labels have been created, we have trained an ML model using Random Forest, XGBoost and Stack models to predict the segment which a customer might belong to. We have experimented with these three models as it is a multi-class classification problem.

4. Implementation

Preprocessing

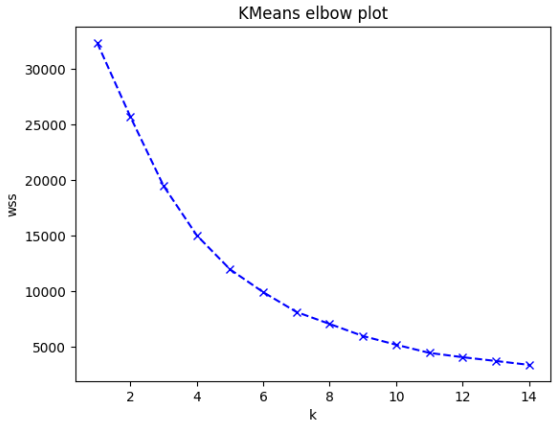
The dataset we used had some missing values for some of the features. In the instance of Customer ID, there were around 180k nulls upon exploration and domain knowledge we found that guest customers do not have an CustomerID assigned to them, so in order to not lose them we have used the Invoice number as their CustomerID. In order get more information from the dataset, we decided to do some calculations between the invoice date and the last day of the study (12/09/2011). These calculations can be found in the new features (previous_visit, days_bw_visit) we created for the data. Based on these features we have created new features such as Recency, Frequency and Monetary value. Finally, we have used sklearn's StandardScalar to standardize our data so that whole data is on the same scale as it effects the unsupervised algorithms as they used distance metric in defining the clusters.

After preprocessing we have used K-means, K-medoids and RFM models to create labels for the data. The observations for these models are described as below

Unsupervised learning Models

K-means

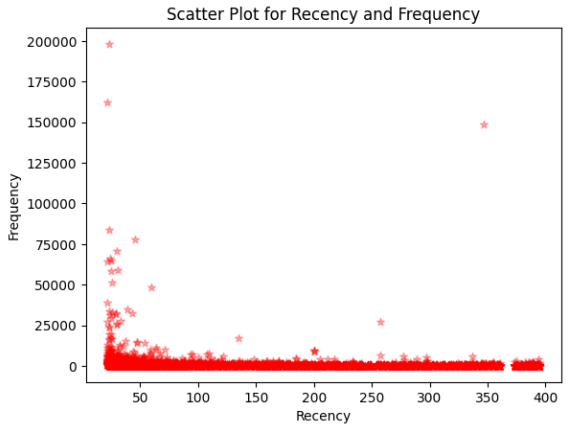
To maximize the performance of our k-means model, we needed to determine the optimal number of nearest neighbors. To find the optimal k value, we analyzed the loss over range of clusters. To do this, we captured the inertia measure of each k-means model and plotted it to create an elbow plot. The "elbow" method allows us to visually see the performance of a k-means cluster by showing us the point of inflection on the graph. The inflection point lets us know the optimum k. In this case, our optimum number of clusters is 5.



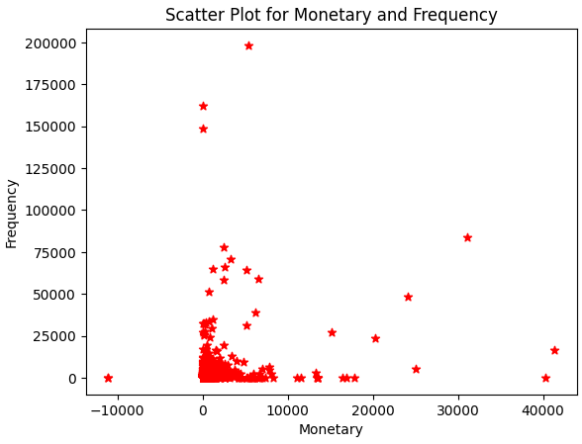
RFM Analysis

As discussed in the previous section, the Recency Frequency Monetary (RFM) models is used to understand customer behavior. RFM uses 3

measures or categories to rank a customer's habits by analyzing their spending history. By using this information, businesses can predict which customers are likely to purchase products, how often they purchase products, etc. Our RFM analysis shows that customers who purchase products less than 50 days have a high frequency (they are buying more). We also investigated the relationship between Monetary and Frequency



and found that frequency decreases when monetary value is low.

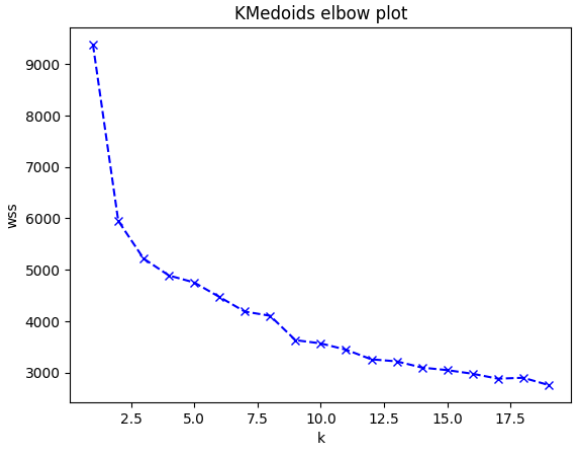


Using the RFM score and by looking at their descriptive statistics, we were able to segment customers into 5 categories:

- Best Customers
- Potential Loyal Customers
- New Customers
- At Risk Customers
- Churned Customers

K-Medoids

Like k-means, k-medoids is another clustering classifier. Instead of using the mean of points to choose a new centroid, k-medoids chooses an actual data point from the dataset to serve as the centroid. To find the optimum number of clusters for k-medoids, we decided to use the elbow method. For our case, the point of inflection occurs when k = 5.



We have finalized on the RFM labels and have converted them into numeric for model training purpose. After creating the labels, we have split the data into train and test with a 70, 30 split respectively and have trained the models using the below algorithms.

Supervised Learning Models

Random Forest

Random Forest consists of smaller decision trees that act as an ensemble. Each label should have equal weight while training the model. So, we have set the class_weight parameter to balanced. Also, we have used a range of parameter in hyper parameter tuning using RandomizedSearchCV for cross validation using random selection of data points

XGBoost

XGBoost is called as Extreme Gradient Boosting. It is a scalable, distributed gradient-boosting decision tree. It provides parallel tree boosting which results in high performance of the models.

We have trained the xgboost model using hyper parameter tuning of max_depth, learning_rate, n_estimators and gamma parameters.

Stacking

Stacking uses meta learner instead of voting to combine predictions of base learners. Prediction of base learners are used as input for meta learner.

5. Results

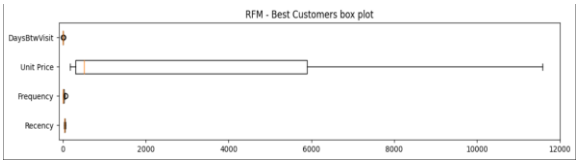
For each model we used, we generated boxplots to show the relationship between customers in a specific category. The figure below shows a summary of the number of customers in each category.

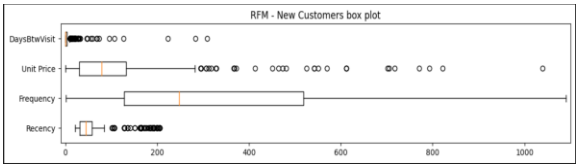
RFM	
Segment	Count
Churned Customers	2674
At Risk Customers	2614
Potential Loyal Customer	1750
New Customers	1028
Best Customers	16

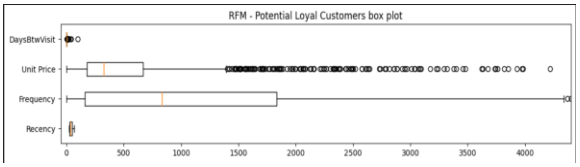
Kmeans	
Segment	Count
0	4856
1	3182
2	17
3	16
4	11

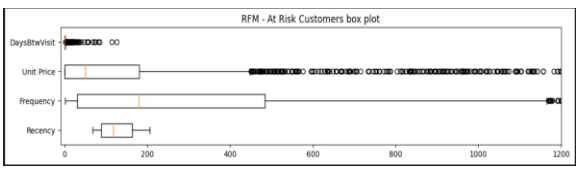
Kmedoids	
Segment	Count
0	1513
1	3115
2	1034
3	944
4	1476

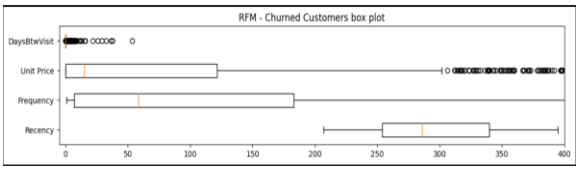
RFM









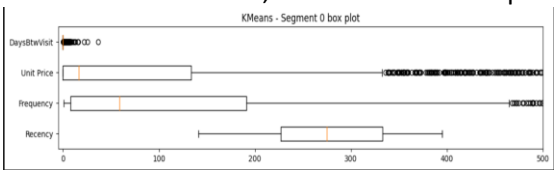


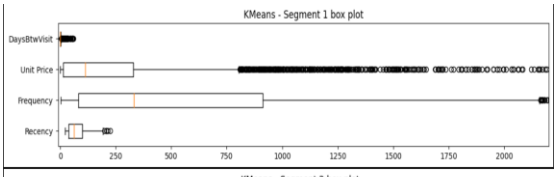
For the RFM model, we saw that the Best Customers group, had low recency and frequency but high monetary value. Although, customers in this group are not spending as frequently as the other groups, they are spending a good amount of money. Like Best Customers, Potential Loyal customers are spending higher than the other categories. To ensure high spending in these categories, businesses can use targeted advertising of high value items to these groups.

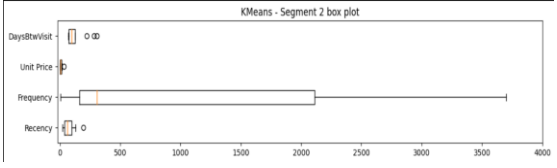
In contrast, one of the lowest performing categories in terms of monetary value is Churned Customers. These are customers that have not shopped in a while and are considered “ex” customers.

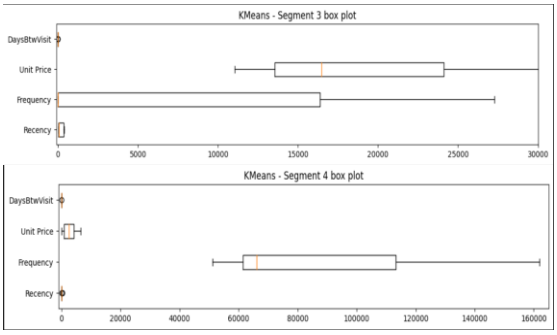
K-means

For the k-means model, we notice the Best Customer category is the inverse of the RFM model. In this case, Best Customer spends little amounts of money but has a high frequency.



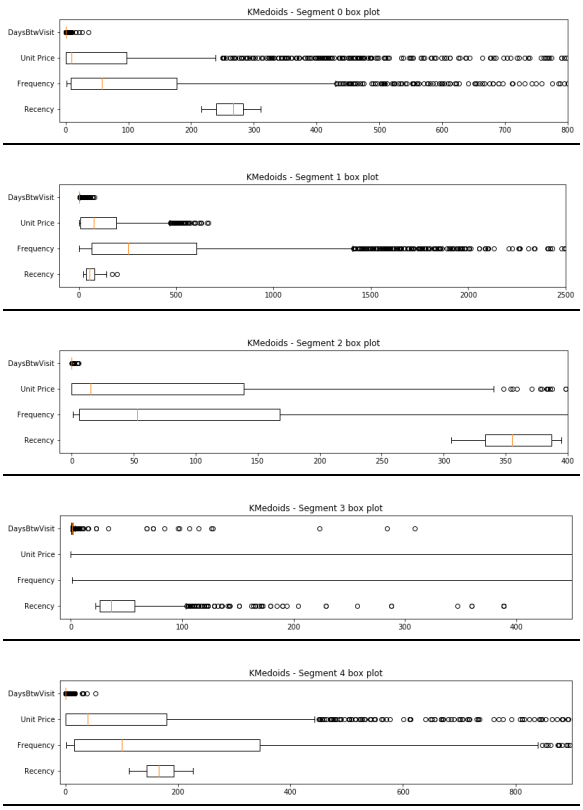






For Potentially Loyal customers, the monetary value is quite high while the frequency varies quite a bit. Since the recency is low, businesses can infer that these customers like purchasing goods frequently but have not made any recent purchases. It might be a good idea to send them a coupon or some incentive to get them to spend.

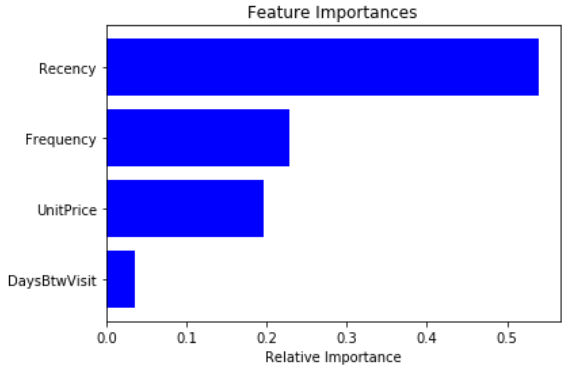
K-medoids



We have mapped 5 of categories to RFM model. Compared to the k-means Best Customers category, the k-medoids seems more representative to what a "loyal" customer would be due to the frequency and recency values. At

Customers and Churned customers categories are similar, we see that Churned customers recency value is nonexistent. The recency value explains why these customers are considered as "ex customers".

Random Forest



Upon implementing Random Forest, we have observed that the features extracted were very useful in terms of predictions and had high importance. We were able to achieve 99% accuracy using Random Forest algorithm along with hyper parameter tuning.

```
Test
0.9995876288659794
[[  5   0   0   0   0]
 [  0 525   0   0   0]
 [  0   0 308   1   0]
 [  0   0   0 784   0]
 [  0   0   0   0 802]]
              precision    recall  f1-score   support

           0       1.00      1.00      1.00         5
           1       1.00      1.00      1.00       525
           2       1.00      1.00      1.00       309
           3       1.00      1.00      1.00       784
           4       1.00      1.00      1.00       802

    accuracy                           1.00      2425
   macro avg       1.00      1.00      1.00      2425
weighted avg       1.00      1.00      1.00      2425
```

XGBoost

XGBoost model had an accuracy of 100% by using hyper parameter tuning but the execution time was long when compared to random forest model.

```
Test
1.0
[[  5   0   0   0   0]
 [  0 525   0   0   0]
 [  0   0 309   0   0]
 [  0   0   0 784   0]
 [  0   0   0   0 802]]
              precision    recall  f1-score   support

           0       1.00      1.00      1.00         5
           1       1.00      1.00      1.00       525
           2       1.00      1.00      1.00       309
           3       1.00      1.00      1.00       784
           4       1.00      1.00      1.00       802

    accuracy                           1.00      2425
   macro avg       1.00      1.00      1.00      2425
weighted avg       1.00      1.00      1.00      2425
```

Stacking

In Stacking model, we have used SVM as the metalearner and have achieved 100% accuracy on the test set.

```
Test
1.0
[[  5   0   0   0   0]
 [  0 525   0   0   0]
 [  0   0 309   0   0]
 [  0   0   0 784   0]
 [  0   0   0   0 802]]
              precision    recall  f1-score   support

           0       1.00      1.00      1.00         5
           1       1.00      1.00      1.00       525
           2       1.00      1.00      1.00       309
           3       1.00      1.00      1.00       784
           4       1.00      1.00      1.00       802

    accuracy                           1.00      2425
   macro avg       1.00      1.00      1.00      2425
weighted avg       1.00      1.00      1.00      2425
```

6. Conclusion and Future Works

We found the RFM model is performing better in customer segmentation when compared to the other models. This model gives us the data labels that businesses can use to tailor marketing campaigns to specific audiences. Also, the ML model we can finalize on is Random Forest as it is fast and easily interpretable. It had an accuracy of 99.9%. We believe that this would be extremely beneficial for small businesses since they can allocate their resources on customers who are more engaged their products or vice versa. Businesses can maximize their use of these models by continually using it with new data they recently gathered.

It is possible for transaction history data to have outliers which would affect k-means and k-medoids models. From our analysis, we see that the k-medoids clustering seems to be more representative than k-means.

Based on the limitations to the research topic, we have come up with some further research directions. The next focused direction for future research is to effectively visualize the results of the model to the businesses, so that they can understand the demographics and purchase patterns of the customer and design a marketing campaign accordingly.

References

Chen, D. (n.d.). Online Retail Data Set. London, UK: School of Engineering, London South Bank University.

Chen, D., Sain, S. L., & Guo, K. (n.d.). Data mining for the online retail industry: A case study of RFM model-based customer segmentation using data mining. *Journal of database marketing & customer strategy management, 19*(3), 197-208. doi:10.1057/dbm.2012.17

Du, X.-P. (2006). Data Mining Analysis and Modeling for Marketing Based on Attributes of Customer Relationship.

Hossain, A. S. (2017). Customer segmentation using centroid based and density based clustering algorithms. *3rd International Conference on Electrical Information and Communication Technology (EICT)*, (pp. 1-6). doi:10.1109/EICT.2017.8275249

Kohavi, R., & Parekh, R. (2004). Visualizing RFM Segmentation. *Proceedings of the Fourth SIAM International Conference on Data Mining.* doi:10.1137/1.9781611972740.36

Li, G. (2013). Application of Improved K-means Clustering Algorithm in Customer.

Applied Mechanics and Materials, (pp. 1081-1084). doi:10.4028/www.scientific.net/AMM.411-414.1081

Maulina, N. R., Surjandari, I., & Rus, A. M. (n.d.). Data Mining Approach for Customer Segmentation in B2B Settings using Centroid-Based Clustering. *2019 16TH INTERNATIONAL CONFERENCE ON SERVICE SYSTEMS AND SERVICE MANAGEMEN.*

Wu, J., Shi, L., Lin, W.-P., Tsai, S.-B., Li, Y., Yang, L., & Xu, G. (2020). An Empirical Study on Customer Segmentation by Purchase Behaviors Using a RFM Model and K-Means Algorithm. *Mathematical problems in engineering*, 1-7. doi:10.1155/2020/8884227