DV 2626 - Final Project Report- Customer Segmentation and Churn Prediction

Jeevanthi Panawala
jepa23@student.bth.se

Sreedharani Machunuru
srma24@student.bth.se

I. Introduction

The rapid growth of e-commerce has intensified the need for businesses to understand and retain their customers. Customer segmentation and churn prediction help companies to know their customers well and align their marketing strategies to retain existing customers. This project aims to apply traditional machine learning techniques on the Online Retail II dataset from the UCI Machine Learning Repository to perform 3 tasks.

1) Customer segmentation based on Recency, Frequency, and Monetary (RFM) metrics
2) Build a model to predict the likelihood of customer churn
3) Based on the customer segments and calculated churn probabilities, identify shoppers with high churn risk, analyze their purchase patterns and provide marketing recommendations to retain them

This report provides a detailed account of the project objectives, method, results, and conclusions.

II. Method

A. Dataset

The Online Retail II dataset includes transactional data from a UK online retailer between 2009 and 2011. It contains details such as invoice numbers, product descriptions, quantities, prices, country of purchase, and customer IDs. The number of instances of the total dataset is around 1 million, but after cleaning (handling missing values and purchase returns) it was around 800,000.

B. Task 1 - Customer Segmentation

Customer segmentation is a critical process in understanding customer behavior and tailoring marketing strategies accordingly. The dataset is unlabeled and we need to mine for customer segments in it. An unsupervised machine learning technique should be used to execute such a task. So we decided to perform descriptive clustering to identify distinct customer groups based on their purchasing patterns.

A limitation of this online retail dataset is that it only contains sales transactions of the customers and no additional attributes about them. For clustering, a value should be assigned to each customer. The RFM (Recency, Frequency, and Monetary value) model is mentioned as a commonly used model in customer segmentation in [1]. The RFM model was chosen to calculate customer values since recency, frequency, and monetary values could be derived from the existing attributes of the dataset.

Since we planned to develop a model to predict customer churn in Task 2, there was a need for a means to label customers as churned or not, to train and test the prediction model. As proposed in the methodology in [2], transaction data from January 2009 to December 2011 was split into two parts: one from January 2009 to May 31, 2011, and the other from June 2011 to December 2011. The latter was used to label the customers who purchased gifts from January 2009 to May 2011 as churned or retained during the period from June 2011 to December 2011. It was decided to use the transaction data from January 2009 to May 2011 for customer segmentation as well.

1) Data Preprocessing: Following preprocessing steps were performed before applying the clustering algorithm for customer segmentation.

- Rows with missing CustomerID and Description were removed as the segmentation of the customer, relies on the identifiers of the customer.
- Rows with negative values for 'Quantity' were removed because we wanted to get rid of the rows which represented purchase returns. However we kept the corresponding purchases as they provided information about the customer's buying patterns.
- Extracted transaction data from January 2009 to May(31st) 2011 for customer segmentation
- A new column was added to store the total value of the purchase
- Transactions were grouped by the CutomerID and the Recency, Frequency and the Monetary Value were calculated for each customer.
  Recency (R): Days since the last transaction for each customer (Reference date for this calculation was set as the The most recent Invoice Date in the Dataset + 1 Day)
  Frequency(F): Number of transactions made by each customer.
  Monetary Value (M): Total monetary value spent by each customer
- As there were varying ranges in R,F,M values (specially in Monetary Value), they were transformed to log scale and standardized using standard scaler to make their mean equal to 1 and variance equal to 0

2) Methodology:

a) Clustering Algorithm: K-Means clustering is an unsupervised machine learning algorithm that partitions data into k clusters based on feature similarity. According to [3], K-means clustering is the most commonly used unsupervised learning algorithm for identifying and grouping data sets. The performance of K-means clustering was compared to BIRCH clustering (a hierarchy-based clustering technique) in RFM-based customer segmentation in [4], and it was revealed that K-means outperforms BIRCH techniques based on the performance measure "Silhouette Score." In addition to these references, we considered the simplicity and widespread usage of the algorithm in customer segmentation in the e-commerce domain when choosing it for the clustering task in this project.

b) Cluster Validation: For validation of clusters, there is no ground truth (already labeled data as each customer belong to which cluster) exists. So an internal validation technique was employed to determine the goodness of the clusters generated.

For clusters to be good, they need to be compact groups(data points in one cluster occur close to each other) and at the same time, they should be well separated groups (Two groups should have as large distance among them as possible) [1]. Silhoette Index was selected for cluster validation as it takes both compactness and separation into account for individual clusters and also for clustering solutions.

To determine the optimal number of clusters, we calculated the Silhouette coefficient for k values ranging from 2 to 10. The k values with the highest Silhouette scores (closest to 1) were then selected for further analysis to gain insights.

C. Task 2 - Churn Prediction

When customers are dissatisfied with the products, they leave the company and start buying products from the competitors and this phenomena is called churn [5]. As the retailing industry is highly competitive, gaining new customers is challenging. As mentioned in [5], it is easier to retain existing customers than acquiring new customers. The importance of not allowing customers to churn, made us working on a churn prediction model as the Task 2 of our project.

1) Data Preprocessing: The studies [2], [5] propose recency, frequency and the monetary value of customers as good candidates for the feature set of a churn prediction model. [2] uses RFM score in addition to R,F,M values and states that the accuracy, precision and f1-socre were improved with the additional parameter. In task 1, we derived the customer segments and assigned archetypes, using the R,F,M values of each customer. So it was decided to add the customer archetype along with R,F,M to the feature set of the churn prediction model. The dataset used for this task was the resulting customer segmentation dataset from Task 1 and to calculate the churn (dependent variable) transaction data (cleaned) from June 2011 to December 2011 was joined with the customer segmentation dataset.

Data Preprocessing Steps:

- Removed the purchase returns from the transaction dataset and extracted transactions within the period of June 2011 to December 2011
- Calculated customer churn using the condition "If the CustomerID in the customer segmentation dataset, is present in the transaction dataset (he has purchased products in the last 6 months of 2011), the customer is not churned (0). Otherwise he is churned (1)"
- Converted Archetype column (norminal) to an ordinal column to create the prediction model
- Split the resulted dataset into train (80%) and test(20%) sets

2) Methodology:

a) Used Algorithm: Studies [2] and [6] recommend two class boosted decision tree algorithms for higher accuracy in churn prediction using RFM values. These two studies provided the motivation to select GradientBoostingClassifier to build the churn prediction model. GradientBoostingClassifier is an ensemble machine learning technique which combines the predictions of several weaker models, typically decision trees, to produce a stronger, more accurate model.

We created Gradient Boosting Classifier with these parameter values to optimize model performance: nestimators=100 provides sufficient boosting stages for accuracy, learningrate=0.1 ensures gradual learning , maxdepth=3 prevents overfitting with simpler trees, subsample=0.8 introduces randomness for better generalization.

b) Model Evaluation: Confusion matrix, classification report and ROC-AUC curve were generated to evaluate the churn prediction model. The confusion matrix provides a detailed breakdown of true positives, true negatives, false positives, and false negatives, offering insights into the classifier's accuracy. The classification report further elaborates with precision, recall, F1-score, and support for each class, giving a comprehensive view of the model's performance across different metrics. Lastly, the ROC-AUC score is calculated, measuring the model's ability to distinguish between classes, with a higher score indicating better performance. This combination of metrics ensures a robust evaluation of the model's effectiveness.

c) Hyper Parameter Tuning: For improving the model performance, we decided to try hyper parameter tuning of Gradient Boost Classifier. A new model was built using suggested best parameters in hyper parameter tuning. The new model was again evaluated using Confusion matrix, classification report and ROC-AUC score.

D. Task 3 - Further Analysis for Marketing Recommendations

The final task of this project involved offering recommendations for the marketing team. The prediction model developed in Task 2 was applied to the dataset containing the customer segments from Task 1 and churn probability of each customer was calculated. The resulting dataset was then analyzed to generate marketing insights and recommendations.

III. Results and Analysis

A. Customer Segmentation

Silhouette Scores were calculated and plotted for k=2 to k=10 to determine the best number of clusters (k value) for k-means clustering.According to the plot, 2 and 4 are the best number of clusters (with Silhouette Scores near to 1), which the retail shoppers can be grouped into.

To further study the separation distance between the resulting clusters, a silhouette plot was generated for the number of clusters from 2 to 4. The silhouette plot displays a measure of how close each point in one cluster is to points in neighboring clusters, and thus provides a way to visually assess parameters such as the number of clusters. Its thickness also provides an idea of the cluster sizes [7]. Figure 2 depicts the Silhouette score of each cluster and the cluster centers of the discovered clusters. This visualization proved that 2 clusters have the best separation between the clusters. However, when the cluster sizes were considered, 3 and 4 clusters provide a clearer insight about the customer segments. When the frequency, recency, and monetary value of the cluster centers were analyzed, it was confirmed that, even if the Silhouette score is better for two number of clusters, four number of clusters give us more clear insights about the customers and those insights provided by four clusters seem more useful to make data-driven decisions.

In 4 clusters, we could identify 4 types customers: High end, frequent buyers and recent shoppers, High end, moderately frequent buyers but they haven't shopped recently, Medium, recent and moderately frequent shoppers and Low end, less frequent and old buyers.Differences of each of these segment's sales amount (monetary value of spending), frequency and recency were visualized and 4 archetypes (Loyal Luxury Shoppers, Long lost Rare Shoppers , Lost Luxury Shoppers and Moderate Recent Buyers) were assigned to each segment. Figure 1 depicts the monetary value differences of the customer archetypes.

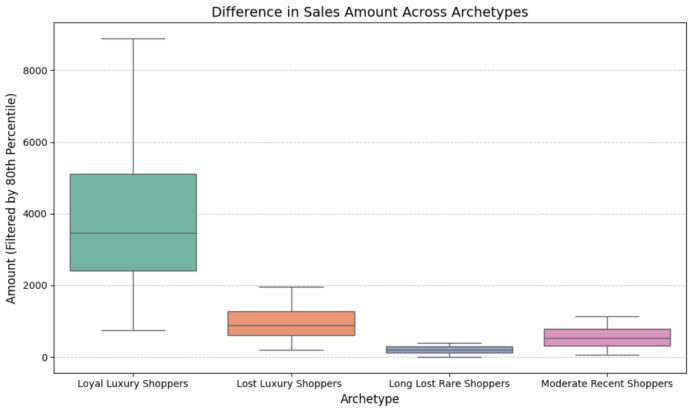

Fig. 1. Monetary Value Differences of Customer Archetypes

B. Churn Prediction

The churn prediction model was developed using segmented customer data derived from RFM (Recency, Frequency, Monetary) analysis. The key features used in the model included absolute RFM values and cluster categorization (from task 1). Our dataset did not contain any other customer demographics such as customer type, age,gender etc. which have been used in other customer churn prediction studies ([2], [5]) and have improved the performance of their models.

The Gradient Boosting Classifier was employed to predict customer churn. The model was trained and tested and performance metrics, including precision, recall, and F1-score, were evaluated before and after hyper-parameter tuning. The performance metrics before tuning the model (with model parameters:nestimators=100,learningrate=0.1, maxdepth=3, subsample=0.8) are displayed in the table I. ROC-AUC value of the baseline model was calculated as 0.824. Confusion matrix

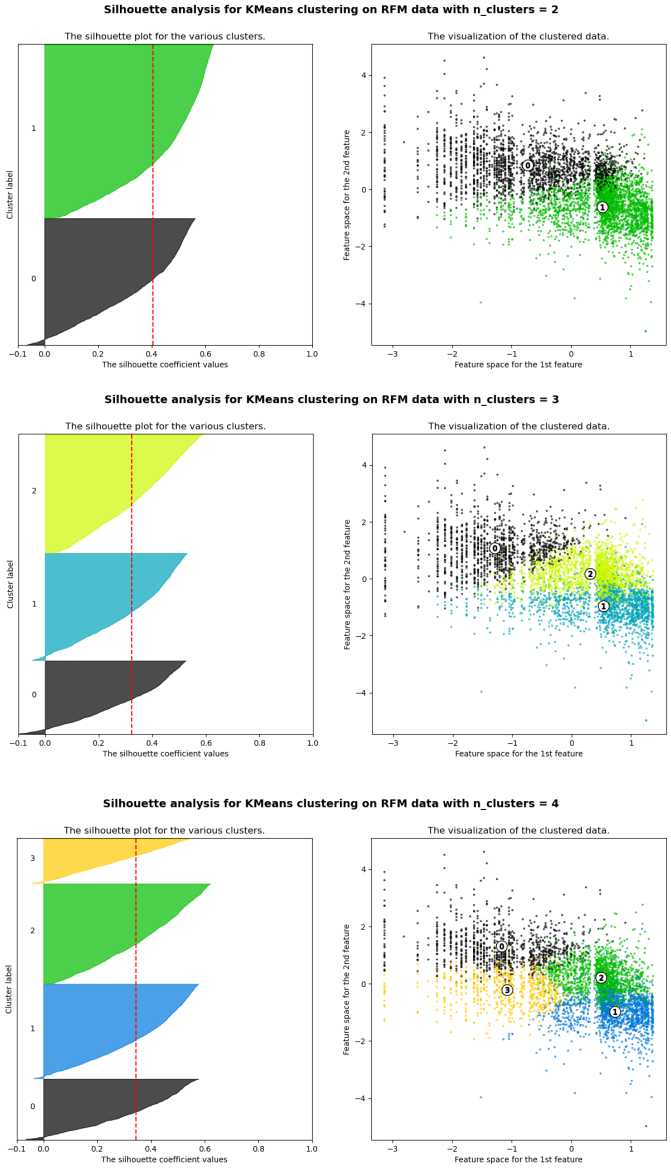

Fig. 2. Silhouette Analysis with 2,3 and 4 clusters

of the model is depicted in Table II.

The only parameter change suggested during hyper-parameter tuning was,learningrate (0,05). The new model resulted in a marginal increase in ROC-AUC value (0.83) while the other performance metrics remained same/ slightly changed. Classification report of the tuned model is shown in the Table III and the confusion matrix is shown in Table IV.

The choice of the model after hyper-parameter tuning is based on the slight increase in the ROC-AUC score from 0.824 to 0.830, which justifies selecting the tuned model for several reasons. The ROC-AUC score measures the model's ability to distinguish between churn and non-churn classes across all threshold levels, and the improvement indicates better overall performance in differentiating between these classes. In predictive modeling, particularly for churn prediction, even small gains can translate into significant business benefits, such as more effectively targeted retention efforts. Additionally, a higher ROC-AUC score suggests a better balance between the True Positive Rate and the False Positive Rate, which is essential in churn prediction to minimize unnecessary retention costs and missed opportunities to retain valuable customers.

	Precision	**Recall**	**F1-Score**
Non-churn	0.77	0.74	0.76
Churn	0.73	0.76	0.74

TABLE I
CLASSIFICATION REPORT BEFORE HYPER-PARAMETER TUNING

	Predicted Not Churn	**Predicted Churn**
Actual Not-Churn	383	132
Actual Churn	115	357

TABLE II
CONFUSION MATRIX BEFORE HYPER-PARAMETER TUNING

	Precision	**Recall**	**F1-Score**
Non-churn	0.77	0.73	0.75
Churn	0.72	0.76	0.74

TABLE III
CLASSIFICATION REPORT AFTER HYPER-PARAMETER TUNING

	Predicted Not Churn	**Predicted Churn**
Actual Not-Churn	377	138
Actual Churn	113	359

TABLE IV
CONFUSION MATRIX AFTER HYPER-PARAMETER TUNING

C. Marketing Recommendations

To provide marketing recommendations, we decided to identify the set of customers who are highly valuable to the selected online retail company and, at the same time, are at high risk of churn. The distribution of the customer archetypes is shown in Figure 3 and the distribution of the customer archetypes for the risk of churn is shown in Figure 4. These graphs highlighted "Lost Luxury Shoppers" as an important customer segment to the company containing around 30% of the total customer base and it has a comparatively high risk of churning.

Fig. 3. Distribution of Customer Archetypes

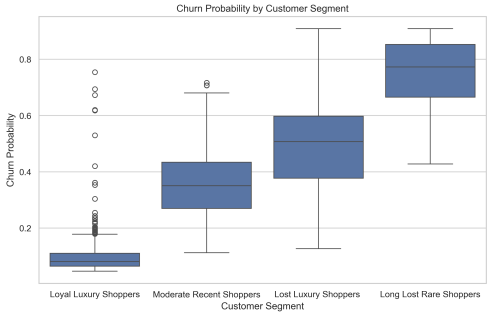

Fig. 4. Distribution of Customer Archetypes in Churn Probability

The set of Lost luxury shoppers with the highest churn probability (more than 50%), high frequency, high purchase value and shopped no later than 6 months were selected for seasonal purchase pattern analysis and market basket analysis for providing insightful marketing recommendations. Seasonal buying patterns of this customer segment (depicted in figure 5) revealed that the peak month of purchase is December (Start of the Winter) and within autumn months it shows an upward trend in purchasing amounts.

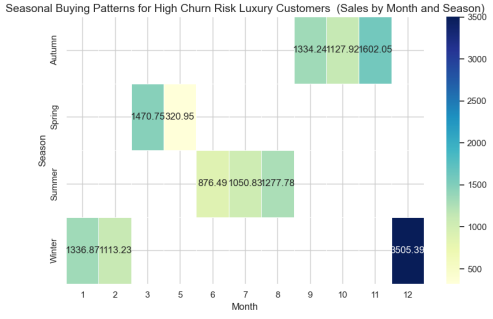

Fig. 5. Seasonal Buying Patterns of Lost Luxury Shoppers

The analysis of products purchased by these luxury shoppers with a high churn risk during the peak months revealed interesting insights. The top-selling items (based on quantity sold) related to the segment of interest, items included in most transactions, and items frequently purchased together were identified. "BATHROOM METAL SIGN" was the most frequently appearing product in transactions, featuring in 25% of them. The item sets (HEART OF WICKER SMALL, BATHROOM METAL SIGN) and (HANGING METAL HEART LANTERN, BATHROOM METAL SIGN) were purchased together in more than 10% of transactions. Based on the results and analysis, the following recommendations can be provided.

- if marketing campaigns are carried out for re-engaging "Lost Luxury Shoppers" with "High Churn Risk", start of Autumn (September) is a good choice to kick off.
- PAPER CHAIN KIT 50'S CHRISTMAS,LUNCH BAG CARS BLUE,WHITE HANGING HEART T-LIGHT HOLDER,BATHROOM METAL SIGN,HEART OF WICKER SMALL,CHOCOLATE HOT WATER BOTTLE and HANGING METAL HEART LANTERN can be focused in preparing sales promotions to re-engage "Lost Luxury Customers"
- (HEART OF WICKER SMALL, BATHROOM METAL SIGN) and (HANGING METAL HEART LANTERN, BATHROOM METAL SIGN) can be better candidates for bundling together if the marketing team plans for product bundling in their promotions.

IV. Conclusion

This project has provided valuable insights into customer behavior by leveraging customer segmentation, churn prediction, and marketing recommendations. Using machine learning techniques such as clustering for segmentation, Gradient Boosting for churn prediction, and collaborative filtering for marketing recommendations, we successfully identified distinct customer groups, accurately predicted churn, and proposed targeted marketing strategies.

The customer segmentation analysis identified four distinct groups. Silhouette analysis proved to be an effective method for determining the optimal number of clusters in k-means clustering. Additionally, the cluster center values of recency, frequency, and monetary (RFM) metrics facilitated a clear identification of the most meaningful customer segments. The churn prediction model achieved an accuracy of 83% in distinguishing between churn and non-churn customers. Furthermore, the marketing recommendations offered can help the marketing team focus on retaining high-value customers and enhance customer engagement.

One challenge faced during this project was the limited availability of features in the dataset, which impacted the accuracy of the churn prediction model. Expanding the dataset with additional customer demographic information would likely improve the model's performance and provide even deeper insights into customer behavior.

V. Contribution

Design and implementation of this project was done together. Brainstorming and coding of all three tasks, were done as a pair programming project, both group members sitting together at the university.

a) Jeevanthi: After deciding on the structure of the report, I wrote the Method section, Results and analysis of Task 2 and 3.

b) Sreedharani: Introduction, results and analysis of Task 1 was written by me.

Conclusion was built up and written together by both the members.

REFERENCES

[1] D. Sarkar, R. Bali, and T. Sharma, Practical Machine Learning with Python. 2017. doi: 10.1007/978-1-4842-3207-1.

[2] Y. Aleksandrova, "APPLICATION OF MACHINE LEARNING FOR CHURN PREDICTION BASED ON TRANSACTIONAL DATA (RFM ANALYSIS)," International Multidisciplinary Scientific GeoConference SGEM ..., Jun. 2018, doi: 10.5593/sgem2018/2.1/s07.016.

[3] L. Rajput and S. N. Singh, "Customer Segmentation of E-commerce data using K-means Clustering Algorithm," 2022 12th International Conference on Cloud Computing, Data Science and Engineering (Confluence), Jan. 2023, doi: 10.1109/confluence56041.2023.10048834.

[4] M. V. Rajesh, S. R. Chintalapudi, and M. H. M. K. Prasad, "A Comparative Analysis of RFM-based Customer Segmentation with K-Means and BIRCH Clustering Techniques," in Advances in computer science research, 2024, pp. 977–989. doi: 10.2991/978-94-6463-471-694.

[5] D. Sweidan, U. Johansson, A. Gidenstam, and B. Alenljung, "Predicting customer churn in retailing," pp. 635–640, Dec. 2022, doi: 10.1109/icmla55696.2022.00105.

[6] P. Lalwani, M. K. Mishra, J. S. Chadha, and P. Sethi, "Customer churn prediction system: a machine learning approach," Computing, vol. 104, no. 2, pp. 271–294, Feb. 2021, doi: 10.1007/s00607-021-00908-y.

[7] "Selecting the number of clusters with silhouette analysis on KMeans clustering,
" Scikit-learn. https://scikit-learn.org/stable/auto$_e$$xamples/cluster/plotkmeanssilhouetteanalysis.htmlsphx-glr-auto-examples-cluster-plot-kmeans-silhouette-analysis-py$