Next-Generation Customer Retention Analysis for E-commerce

*

Keerthi Yanala	Sreekar Thanda	SriHarsha Ponakala	Pooja Bhathaluri
studentID *11629948*	student ID *11636306*	studentID *11580529*	student ID *11673390*
keerthiyanala@my.unt.edu	*sreekarthanda@my.unt.edu*	sriharshaponakala@my.unt.edu	*poojabhathaluri@my.unt.edu*

Abstract—In the changing world of online shopping, understanding and predicting how customers behave is crucial, for the success and satisfaction of businesses. This project aims to create a machine learning model that can analyze customer retention in e commerce platforms. Our goal is to predict when customers are likely to make purchases during their online browsing sessions. By doing marketing teams can tailor their strategies effectively. We will use data sets and perform thorough data pre-processing, including cleaning and feature engineering to prepare for a detailed analysis of customer behavior. Through techniques we will uncover insights into both the qualitative and quantitative aspects of consumer behavior. Afterward we will select features. Utilize advanced machine learning algorithms like Gradient Boosting, Support Vector Machines and K Nearest Neighbors to build predictive models. These models will be rigorously evaluated using metrics such as AUC ROC, Precision, Recall F 1 Score and Confusion Matrix to ensure their accuracy and reliability. The expected outcome of this project is a machine learning tool that provides insights to enhance customer experiences while fostering brand loyalty and driving financial growth for e commerce businesses. This research not contributes to discussions on customer retention strategies but also offers practical recommendations, for e commerce platforms aiming to improve user engagement and retention.

I. INTRODUCTION

In the changing world of online shopping being able to predict how customers will make purchases is crucial, for success. Our project, called "Next Generation Customer Retention Analysis for E commerce " aims to use machine learning to forecast the likelihood of customers making purchases while browsing online. This prediction ability is not a pursuit; it has practical value for marketing teams who want to optimize their advertising campaigns pricing models, personalized techniques and promotions.

To improve customer satisfaction and drive revenue growth our project takes an approach to creating models. This involves collecting data pre-processing it analyzing it exploratively and rigorously evaluating our models. We gather data from available sources, like kaggle and Google Big Query that cover a wide range of e commerce interactions including consumer demographics, detailed purchase information and engagement metrics.

The foundation of our process lies in ensuring the integrity of our data through cleaning and integration. Through feature engineering techniques we uncover patterns. Create variables that enhance the predictive abilities of our models.

Our analysis of the data involves exploring both quantitative aspects. We use tests and correlation analyses to uncover meaningful relationships, within the data.

When it comes to selecting and training machine learning models we carefully consider the dynamics of e commerce data. We compare classifiers like Gradient Boosting, Support Vector Machines and K Nearest Neighbors to find the suitable algorithms that address the challenges of customer retention. Additionally we tune our models using hyper-parameter tuning techniques to ensure they perform well with real world data.

The main goal of this project is to advance the infrastructure in the e commerce industry. By improving customer retention we not aim to enhance outcomes and brand loyalty for e commerce platforms but also strive to elevate the overall customer experience in this digital centric era. Our project holds significance as it emphasizes using cutting edge techniques to develop an effective strategy for retaining customers.

By combining insights, from our research with knowledge from a range of literature sources we aim to provide practical strategies that can be implemented by e commerce businesses to improve their customer engagement and retention approaches. This project aims to provide businesses with a tool that offers valuable insights. It will help them successfully navigate the challenges of retaining customers, in the changing marketplace.

II. GOALS AND OBJECTIVES

A. Motivation

I was inspired by observing the transformations happening in the e commerce industry. The need, for infrastructure is growing as businesses expand globally. The online market is paced customer focused and highly competitive. This motivated me to start a project that aims to enhance customer retention through development. While retaining customers has always been important the digital era offers opportunities to personalize and improve customer experiences using data and various touchpoints. This in turn can lead to growth in both aspects and brand loyalty for e commerce organizations. In the landscape of e commerce maintaining customer loyalty and reducing churn is crucial due, to constantly changing consumer

behaviors and preferences. Ultimately my motivation stems from a desire to enhance customer satisfaction optimize business performance and stay ahead in this evolving e commerce environment.

B. Significance

The projects importance lies in its potential to improve the long term viability of businesses and enhance customer experiences resulting in growth and increased brand loyalty. From a perspective it aims to generate understandings of how machine learning can be applied in the e commerce industry providing valuable knowledge to the field. In terms the project is expected to offer strategies, for e commerce platforms that will optimize user engagement and encourage customer retention.

C. Objectives

Our goals are threefold; Firstly we aim to create a model that accurately predicts the likelihood of customer purchases. Secondly we want to analyze and identify the factors that influence customer decisions and purchasing behaviors. Lastly based on these insights our objective is to come up with strategies, for retaining customers on e commerce platforms.

D. Features

Collecting Customer Data; Gather information such, as age, gender, location and other demographic details. Keep track of customer interactions with our customer service team, their comments and their reactions to our marketing materials.

Creating Metrics; Calculate the Recency, Frequency and Monetary values for each customer to gain an understanding of their engagement and purchasing behavior. Develop metrics like session duration, click through rates and conversion rates to assess the level of customer involvement.

Utilizing Predictive Models; To forecast the likelihood of customers churning based on data employ models, like logistic regression, decision trees and neural networks. Consider using techniques or combining models to enhance prediction accuracy.

Assessing Churn Risks; Categorize clients into risk levels based on their probability of churning. This will help in tailoring retention strategies. Implement mechanisms that automatically update churn probabilities when new data becomes available.

E. Reporting

We plan to provide an explanation of our approach including how we handle the data choose the models and measure their effectiveness. We will carefully examine the results discussing how well the models perform, which features are most important and how accurate our predictions are. We'll make sure to share our findings through journals, conferences and industry seminars to reach an audience.

F. Improvement

The model will go through a process of improvement by taking into account feedback on its performance and recognizing patterns, in the data. The strategies derived from the model, which are based on data driven approaches will be put into action within e commerce environments. Their effects will be closely observed. Moreover the project will consistently explore sources of data well as innovative machine learning methods and analytical approaches to remain at the forefront of research, in e commerce analytics.

III. BACKGROUND WORK

[1]The primary aim of the study was to investigate how psychographic factors impact customer satisfaction and loyalty in the field of e commerce. Additionally the researchers aimed to identify customer segments based on these factors. They used a research approach and distributed a questionnaire among 411 participants. The data was analyzed using Path Analysis, Cluster Analysis and Cross tabulation as the methods. The results showed that factors such, as Perceived Usefulness and Trust directly influence customer loyalty while elements like Brand Image, Information Quality, Promotion, Value and Service Quality indirectly affect loyalty through customer satisfaction. This research identified three market segments in e commerce; The Functional Shopper, The Credibility Matters Shopper and The Money Dietary Shopper. Each segment has its characteristics and priorities.

Path Analysis was employed to determine how variables (such as Perceived Usefulness, Trust, Brand Image) impact customer retention through customer satisfaction. This method provided insights into both indirect influences of these variables on retaining customers.

Cluster Analysis played a role, in identifying market segments based on factors. It helped group e commerce customers into three clusters with characteristics and preferences regarding e commerce platforms. Cross tabulation; When applied to factors this technique allowed the researchers to use variables such, as age, occupation and spending habits as components in market segmentation. This provided a understanding of different customer groups.

Key Findings; Direct Influencers; The study discovered that Perceived Usefulness and Trust have an impact on customer retention. This emphasizes the significance of these factors in building loyalty and encouraging repeat purchases.

Indirect Influencers; Other elements like Brand Image, Information Quality, Promotion, Value and Service Quality were found to influence customer retention through their effect on customer satisfaction.

[2]Market Segments; By categorizing customers into three segments with unique characteristics this study offers valuable insights, for targeted marketing strategies.

The research paper, titled "Measuring Customer Retention, in the B2C Electronic Business; An Empirical Study," effectively tackles the challenge of quantifying customer loyalty in a online marketplace. By analyzing data from than 45,000 customers of a web based beauty retailer over a span of

six years (2005 2010) the study proposes an approach to measuring customer retention. It suggests using the percentile of time between purchases as a benchmark for evaluating customer loyalty. This metric offers an understanding of customer loyalty as a state and enables longitudinal analysis of repeat patronage behavior. The studys thorough data analysis, which categorized customers based on their purchase year revealed fluctuations in retention rates over time. Notably it observed a contrast in retention levels; approximately 3 of customers who were acquired in 2005 remained loyal by 2010 compared to 59 of those acquired in 2010 indicating an escalating attrition rate, with the passage of time since their initial purchase.

[3]The research paper titled "The Importance of Customer Segmentation, in E Commerce" emphasizes the role that customer segmentation plays in e commerce. It highlights how customer segmentation helps businesses identify customers and meet their needs. The paper explores algorithms used for segmentation such as the K Means Clustering Algorithm, which's well known for its simplicity and effectiveness in reducing cluster performance errors. It also discusses Clustering, which creates a structure of data points within clusters using agglomerative and divisive methods. Another algorithm explored is Density Based Clustering (DBSCAN) which excels at forming clusters with shapes based on density making it suitable for datasets. Additionally the paper examines Affinity Propagation Clustering, where all data points are considered as cluster centers utilizing a similarity matrix to improve effectiveness. Through an analysis the paper highlights the efficiency of the K Means algorithm for customer segmentation in e commerce. In conclusion the paper asserts that effective customer segmentation is critical, in the e commerce sector and recommends using K Means clustering to enhance marketing focus and improve business dynamics.

IV. DATA SET

- Name: Ecommerce customer churn Analysis and Prediction.
- Number of features: 19
- Number of Rows: 5630

https://www.kaggle.com/datasets/ankitverma2010/ecommerce-customer-churn-analysis-and-prediction/

The dataset is sourced from a known E commerce company. The company aims to identify customers who're likely to churn so they can reach out to them. Offer special promotions.

The dataset, called "Ecommerce " contains customer information and behaviors. It includes columns such, as Customer ID, Tenure (representing the duration of their association with the company) Preferred Login Device (indicating their used device for accessing the platform) City Tier, Warehouse To Home (referring to the distance, between the warehouse and customers location) Preferred Payment Mode, Gender, Hour Spend On App, Number Of Device Registered, Preferred Order Category, Satisfaction Score, Marital Status, Number Of Address Complain (if any) Order Amount Hike From last Year Coupon Used, Order Count, Day Since Order and Cashback Amount.

Customer ID serves as an identifier that links attributes and interactions with each specific customer. This helps in conducting analysis.

Tenure represents how long a customer has been associated with or using the service or platform.

Preferred Login Device indicates the device that a customer prefers when logging into the system. It could be a phone or any other preferred device. City Tier. This column categorizes cities or geographic regions based on their level of development, population size, infrastructure and other relevant factors.

Warehouse, to Home. This indicates the estimated time it takes for a product to be delivered from the distribution center or warehouse to the customers specified delivery address. It also considers the distance and efficiency of the delivery system.

Preferred Payment Method. The "Preferred Payment Mode" column typically captures customers favored method of payment when conducting transactions or making purchases, on a platform or service.

Gender. The "Gender" column is used to record an individuals gender identity within a dataset. It classifies individuals based on their gender, including male, female and other gender identities.

Hours Spent on App. The "Hours Spend On App" column indicates the amount of time a user actively engages with or utilizes an application or platform over a given period usually measured in hours.

Number of Registered Devices. This specifies how many devices (e.g., smartphones, tablets, computers) are registered or authorized to access a service or application through one user account.

V. DETAILED DESIGN OF THE FEATURES

A. Analysis

Data Collection and Preprocessing Data collection: We gather a detailed e-commerce information and customer public data from the Kaggle.

Data cleaning: As a process of data cleaning, we deal with missing numbers, outliers, and discrepancies by cleaning the obtained data. combine data from numerous sources to produce a single dataset for analysis. Now displayed the top 5 rows in the dataframe.

Fig. 1. top 5 rows

Found the number of missing values in Tenure, WarehouseToHome, HourSpendonApp, Order Amount Hike From last Year, CouponUsed, OrderCount, DaySinceLastOrder to

be 264, 251, 255, 265, 256, 258, 307 respectively. Now, In our dataset, Being Null Values means we considering them as Zero, So filled/replaced all those null values with value 0. For this process, I used fillna(0) is used. After filling up null values with zero value, again calculated if there are still any missing values and the result is none.

As a process of cross verification, We used info() to display the information of data, where it resulted to showing that the number of to total entries are 5630 rows and 20 columns, Then printed each column with number of entries in rows, whether it is a null count or non-null count, then the datatypes.

Now, concentrating on the Cash Back Amount variable, tried to find the outliers in the dataset by visualizing the trend in the variable. Found that there are some outliers, could be seen in below screenshot.
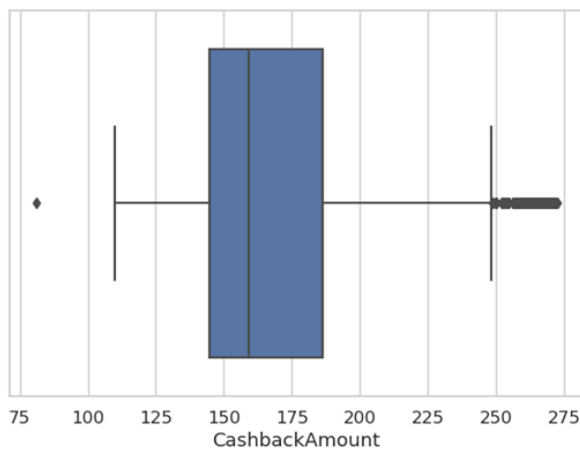

Fig. 2. outliers

From the above screenshot, could understand that most of the customers received a cashback in the range of 140 to 180. There are less customers who received from 1 to 52 cashback and 260 to above.

Below screenshots shows the exact potential outliers, Count see that customer ID 10 received 295.45 cashback, customer ID 5561 received 321.36 ans so on. Finally there are around 438 outliers fig[3].

As a process of cross verification, We used info() to display the information of data, where it resulted to showing that the number of to total entries are 5630 rows and 20 columns, Then printed each column with number of entries in rows, whether it is a null count or non-null count, then the datatypes fig[4].

After Removing the outliers, the number of rows for cleaned dataset has above variables with 5192 rows. So now finally our cleaned dataset has 5192 rows with 20 columns.

We found the statistical summary of each variable in the dataset. The count gives the number of rows counted in the dataset, The mean of Customers spending their hours in the App is 2.79, which means on an average 2.79 hours is the time that customers spending on the App. The minimum time recorded is 0 while the maximum hours recorded is 5hrs. On an average all customers registered in 3 devices.
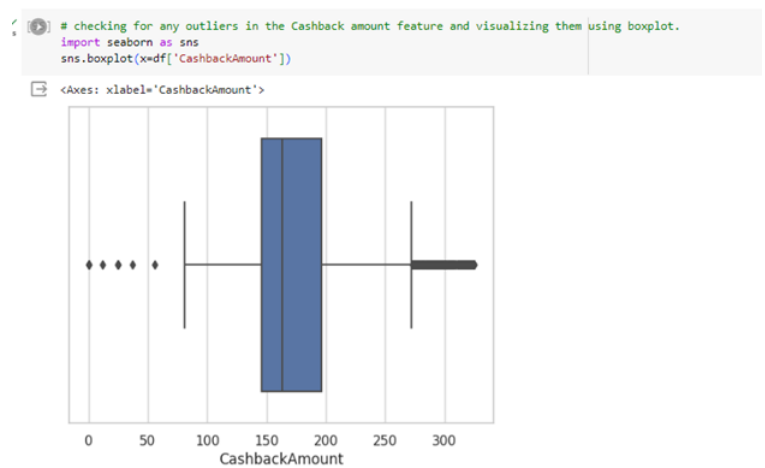
Fig. 3. after removing outliers

Fig. 4. Datatypes of Each Variable

While maximum devices registered is 6. The satisfaction score records say that customers feedback rated on average 3 while minim is 1 and highest rating is obviously 5. The statistics from the Order Amount Hike From Year variable are as follows, the average hike increased is USD 15.7 while the minimum is USD 11 and maximum is USD 26. The average Order Counts placed by customers is 2.66 while the minimum is 1 and maximum is 16 orders. The number of days that a customer lastly ordered recorded in the database is 0 to 46 in range. Mean gives the mean of the respective variable, for example, Mean of Cashback received to customers is USD 167.57, The minimum is USD 81/- while the maximum is USD 272.32.

fig[9] The WareHousetohome graph clearly shows that customers are located close to the companys warehouses, with the majority residing within a 20 mile radius. This proximity likely has an impact on delivery efficiency and customer satisfaction.

```
         CustomerID       Churn      Tenure    CityTier  WarehouseToHome \
count   5192.000000 5192.000000 5192.000000 5192.000000      5192.000000
mean   52793.843413    0.178737    8.823190    1.664869        15.047573
std     1625.966652    0.383168    8.144189    0.920545         9.036360
min    50001.000000    0.000000    0.000000    1.000000         0.000000
25%    51385.750000    0.000000    1.000000    1.000000         9.000000
50%    52767.500000    0.000000    7.000000    1.000000        13.000000
75%    54199.250000    0.000000   14.000000    3.000000        20.000000
max    55630.000000    1.000000   51.000000    3.000000       127.000000

        HourSpendOnApp  NumberOfDeviceRegistered  SatisfactionScore \
count      5192.000000               5192.000000        5192.000000
mean          2.786402                  3.680663           3.064908
std           0.947893                  1.031253           1.382886
min           0.000000                  1.000000           1.000000
25%           2.000000                  3.000000           2.000000
50%           3.000000                  4.000000           3.000000
75%           3.000000                  4.000000           4.000000
max           5.000000                  6.000000           5.000000

        NumberOfAddress   Complain  OrderAmountHikeFromlastYear  CouponUsed \
count       5192.000000 5192.000000                  5192.000000 5192.000000
mean           4.160632    0.288328                    15.709553    1.636171
std            2.584575    0.453028                     3.668870    1.773960
min            1.000000    0.000000                    11.000000    0.000000
25%            2.000000    0.000000                    13.000000    1.000000
50%            3.000000    0.000000                    15.000000    1.000000
75%            6.000000    1.000000                    18.000000    2.000000
max           22.000000    1.000000                    26.000000   16.000000

          OrderCount  DaySinceLastOrder  CashbackAmount
count    5192.000000        5192.000000     5192.000000
mean        2.661787           4.076079      167.570089
std         2.636236           3.543519       35.331390
min         0.000000           0.000000       81.000000
25%         1.000000           1.000000      144.832500
50%         2.000000           3.000000      159.235000
75%         3.000000           7.000000      186.202500
max        16.000000          46.000000      272.320000
```
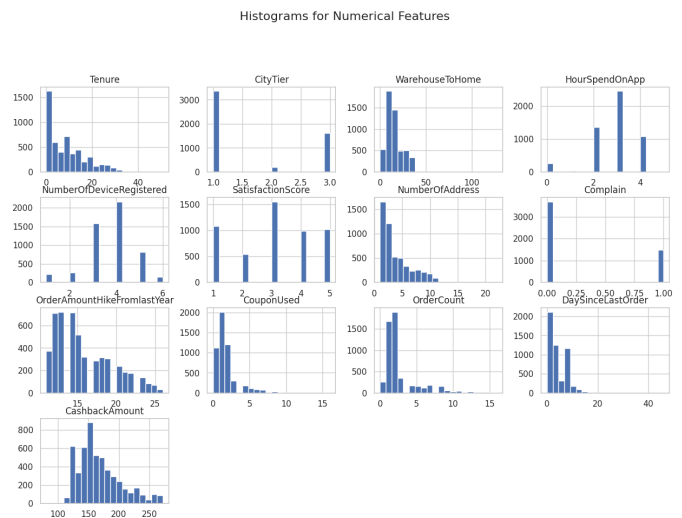
Fig. 5. statistical summary of cleaned data

	CustomerID	Customer Name	Tenure
3743	53744	Logan Baker	51.0
928	50929	Lucas Edwards	50.0
4166	54167	Ethan Turner	31.0
4520	54521	Jackson Cooper	31.0
4323	54324	Liam Nelson	31.0
4385	54386	Ethan Murphy	31.0
4392	54393	Elijah Jordan	31.0
3448	53449	Noah Bell	31.0
4918	54919	Harper Cooper	31.0
4903	54904	Noah Murphy	31.0

Fig. 6. Top 10 Customers whose Tenure with Company is high

	CustomerID	Customer Name	Tenure
2297	52298	Olivia Roberts	0.0
2040	52041	Mason Murphy	0.0
1815	51816	Addison Peterson	0.0
2684	52685	Ethan Russell	0.0
568	50569	Zoey Allen	0.0
2683	52684	Zoey Thompson	0.0
570	50571	Carter Simmons	0.0
1388	51389	Sophia Murphy	0.0
572	50573	Mia Jordan	0.0
1387	51388	Wyatt Richardson	0.0

Fig. 7. Least 10 Customers with Less Tenure

AverageOrderValue
11.0
15.0
14.0
23.0
11.0

Fig. 8. Top 5 datapoints of newly derived variable- AverageOrderValue

Histograms for Numerical Features

Fig. 9. Histogram Distribution of Numerical variables

Another important metric is app engagement as than 2000 customers spend over three hours on the app highlighting its user nature and engaging features.

The device registration data reveals that 2000 customers have registered four devices each followed by around 1500 customers with three registered devices. This suggests a tech customer base. Underscores the significance of implementing a multi platform strategy. Looking at customer satisfaction scores most customers rate their satisfaction at a level of three out of five indicating room for improvement in enhancing their experience.

Address data indicates that 1500 customers have saved than three addresses with the company. This could be seen as an indication of trust and commitment, to the platform. The data related to the increase in order amounts from year suggests that there has been a rise, in both the volume of orders and the amount spent. It appears that around 600 to 650 customers have increased their order amounts from $11 to 15$. This could be attributed to either an increase in product prices or an improvement in the purchasing power of customers.

When we look at coupon usage and order placement patterns it becomes evident that there is customer engagement. 2000 customers have used at 2 coupons and have placed a maximum of 3 orders. Furthermore analyzing the DaysSinceLastOrder data reveals that, over 2000 customers have recently placed orders with their recent order being very recent. On the hand some customers have a gap of 18 days since their order indicating varying purchasing frequencies.

Lastly when we examine the cashback analysis we find that most customers receive cashback ranging from $140 to 200$. This incentive likely boosts customer loyalty. Encourages them to place orders frequently.

From the fig[10] graphs, The Distribution of Gender and marital status, city tier is very clear. Most customers are male who are around 3000+, 2500+ customers are married, and 1600 customers are single while 750 customers are having
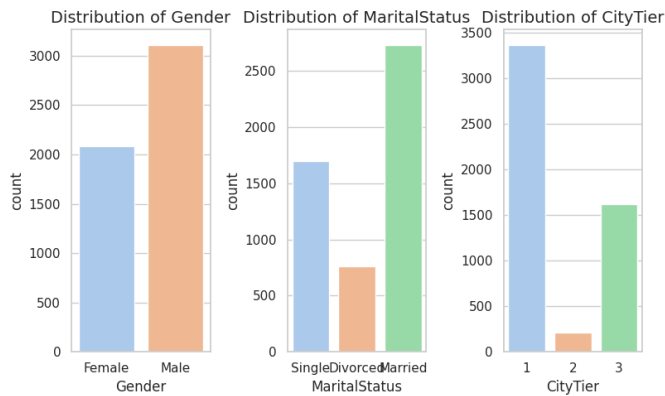
Fig. 10. Bar Graph distributions of Gender, Marital Status, Resident City Tiers of Customers
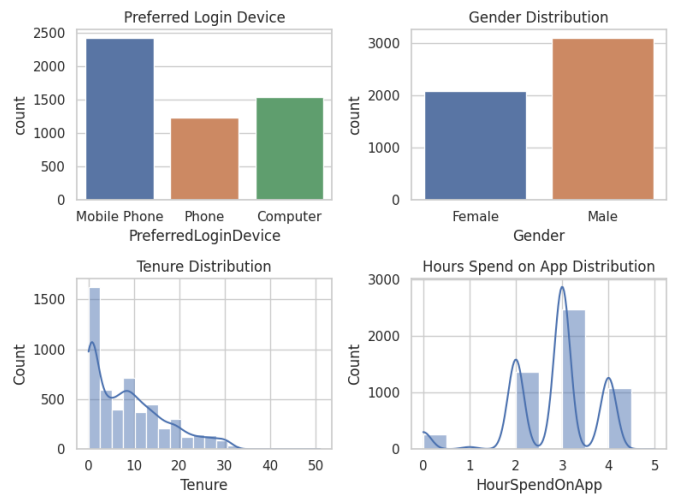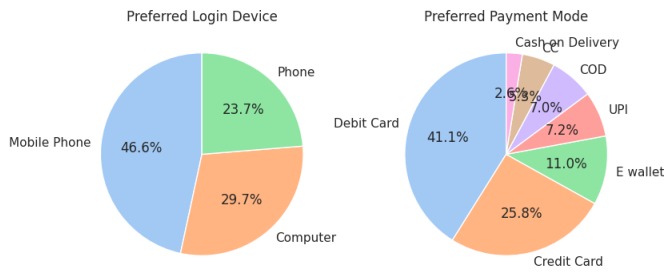

Fig. 11. Customer's preferred Login Devices and Payment methods using Pie-Plots

Fig. 12. Distribution of Customers

Fig. 13. Correlation Analysis

divorced status, 3300 customers are from Tier 1 city,1550 customers are from tier 3 cities and rest from tier 2.

Fig[11] In the above graph, the preferred login device and preferred payment mode is visualized, around 46.6 customers prefer mobile phone for login and 29.7 customers prefer computers while the rest use phones. 41.1 customers prefer debit cards for their payment while 25.8 customers prefer credit cards for payment, lest number of customers (2.6) prefer the cash on delivery payments.

Fig[12] The examination of the e commerce data-set uncovers findings about customer preferences and behaviors. To start with a majority of users prefer logging into the platform using their phones. This preference highlights the significance of having a mobile experience. Reflects the increasing trend of mobile usage, in e commerce activities. When it comes to gender distribution there is an representation of male and female customers indicating that the platform appeals to a wide range of genders.

Regarding customer tenure there is a range in how customers have been associated with the platform. However there is also a concentration of customers who have recently joined or have been associated with it for a period. This suggests either growth in popularity or successful acquisition strategies that attract customers. Moreover analyzing app usage patterns reveals that most customers spend between 2 to 4 hours on the app with 3 hours being the duration. This high level of engagement indicates that the app plays a role, in customers
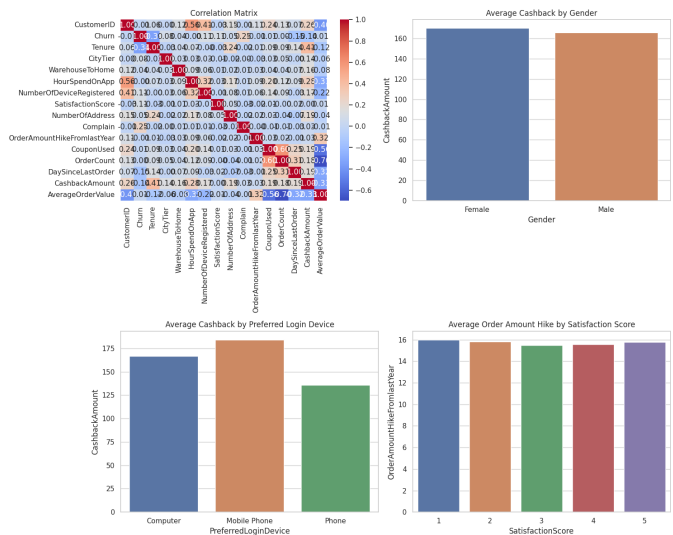
shopping experience and interaction with the service.

Fig[13] The correlation matrix heatmap shows how different numerical features are related to each other. For example we might notice correlations, between features like Tenure and OrderCount suggesting that customers who have been with the platform for a time tend to place orders.

When analyzing the cashback amount we find that on average female customers receive a cashback amount of USD 170.11 compared to male customers who receive USD 165.87. Moreover customers who prefer logging in using a phone receive the average cashback of USD 184.23 followed by those using a computer (USD 166.80) and lastly those who prefer using a phone (USD 135.68).

Regarding satisfaction levels and order amount hike it is interesting to note that customers with a satisfaction score of 1 experience the highest average order amount hike from the year at 15.99. Conversely those with a satisfaction score of 3

Fig. 14. Distribution of Customers order counts

Fig. 15. Distribution of Customers Registered devices, Marital Status, Tenure, Satisfaction Score

Fig. 16. Pair-plot distribution of variables

witness the hike at, around 15.00.

Fig[14] Based on the order count distribution analysis it can be observed that the majority of customers place two orders with 1895 occurrences. This is followed by one order, which occurred 1675 times and three orders, with 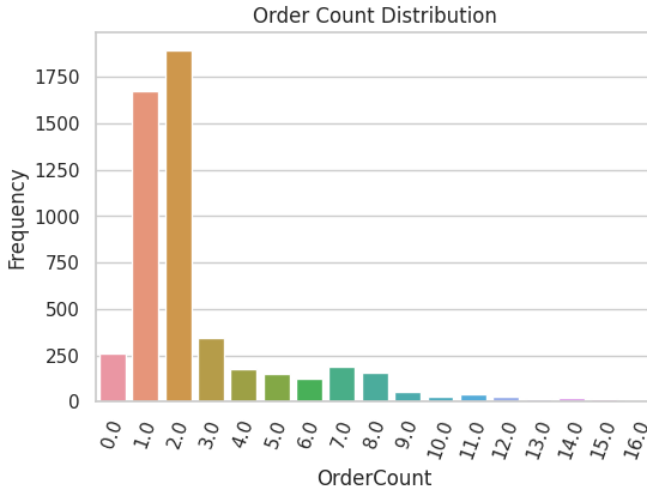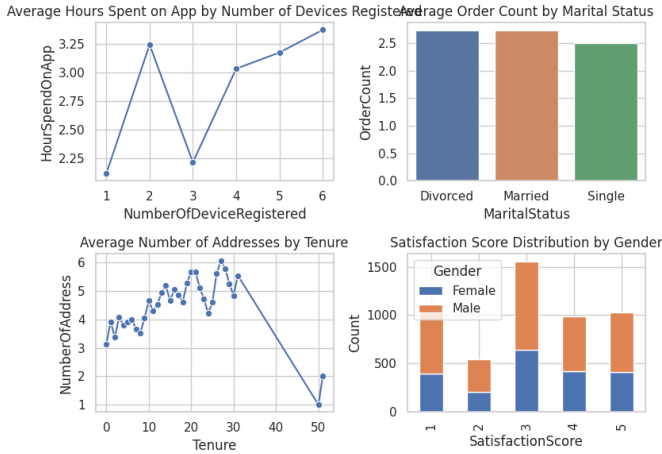345 occurrences. As the order count increases the frequency of orders decreases. These findings can provide insights for enhancing customer satisfaction and engagement. For example studying the connection between cashback amounts and customer login devices can help determine where to concentrate marketing efforts or make improvements to the platform. Likewise exploring how satisfaction levels correlate with an increase in order amounts can enable tailored strategies, for retaining customers based on their satisfaction levels.

Device Usage vs. App Time: There is a varied relationship between the number of devices registered and the average hours spent on the app. Interestingly, customers with 6 registered devices spend the most time on the app, followed by those with 2 devices. However, the relationship is not strictly linear, indicating other factors may influence app usage time. Marital Status and Order Count: Married customers place slightly more orders on average than Divorced or single customers. This suggests marital status might influence shopping behavior, potentially due to varying needs or time availability. Tenure and Addresses: Customers with a longer tenure tend to have more addresses on average, which could be a sign of customer loyalty or a more extensive use of the e-commerce platform over time. Notably, there are peaks at certain tenure lengths, such as 14, 21, and 28 years, where the average number of addresses is notably higher. Satisfaction Score by Gender: The distribution of satisfaction scores between genders shows that both female and male customers report a range of satisfaction levels, with the majority falling at score 3. However, females seem to have a slightly higher representation at the extreme satisfaction scores of 1 and 5 compared to males. Additionally, Fig[16] the pair plot of key numerical features provides a visual overview of pairwise relationships and distributions. This can help identify patterns and areas for deeper investigation, such as clusters of users with similar behaviors or outliers that may warrant further scrutiny.

VI.

VII. IMPLEMENTATION AND PRELIMINARY RESULTS

A. Linear Regression

In the study of customer churn, our primary focus is on the dependent variable, churn, which represents whether a customer is likely to discontinue their service. The model developed to predict churn uses several features, each with

Fig. 17. Customers Churn Imbalance

a coefficient indicating its impact on the likelihood of churn. The significance of these features is determined by their p-values, where a value lower than 0.05 typically suggests statistical significance in predicting churn. These coefficients are instrumental in understanding how changes in each feature influence the probability of churn, assuming all other variables remain constant.
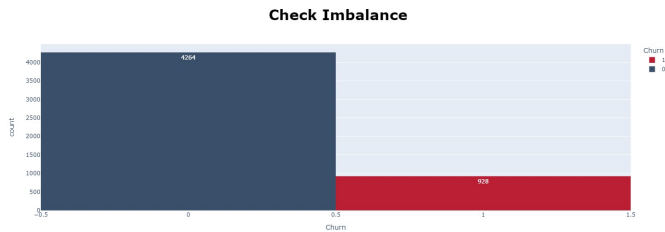
To evaluate the model's performance, various metrics are employed. The Recall Score, sitting at 53.8, measures the proportion of actual churn cases that the model successfully identifies. This implies that the model correctly predicts churn in slightly over half of the actual cases. The Precision Score, at approximately 72.5, indicates the accuracy of the model's positive predictions. This means that around three-quarters of the predictions made by the model are accurate. Additionally, the model's Accuracy Score reveals an overall prediction correctness of about 88.3, showcasing its efficacy in churn prediction.

The influence of different features on churn prediction varies. Features like Tenure, City Tier, Number Of Device Registered, Satisfaction Score, Marital Status, Number Of Address, and Complain have significant impacts on churn likelihood due to their notable coefficients. Conversely, features such as Hour Spend On App, Order Amount Hike From last Year, Coupon Used, Cashback Amount, and Average Order Value appear less influential, as their coefficients are less significant or closer to zero.

Furthermore, a Decision Tree Model was also employed, showing a Recall Score of approximately 62.3. This indicates that the model correctly identified about 62.3 of actual churn cases. The Precision Score of the model is 72, suggesting that around 72 of the churn predictions made were accurate. With an Accuracy Score of about 89.1, the model demonstrates high overall correctness in distinguishing between churn and non-churn cases. These metrics collectively provide a comprehensive understanding of the model's capability in predicting customer churn, thereby enabling more informed decision-making in customer retention strategies.

	Feature	PC1	PC2
1	Customer Name	0.999970	0.001273
20	AverageOrderValue	0.000702	-0.052758
18	DaySinceLastOrder	0.000228	0.015677
13	NumberOfAddress	0.000105	0.012669
4	CityTier	0.000062	0.004846
7	Gender	0.000057	-0.000761
3	PreferredLoginDevice	0.000011	-0.006640
14	Complain	-0.000008	0.000119
6	PreferredPaymentMode	-0.000012	0.008442
12	MaritalStatus	-0.000045	-0.001876

Fig. 18. Dimensionality Reduction using PCA

Fig. 19. Enter Caption

```
                Generalized Linear Model Regression Results
==============================================================================
Dep. Variable:              Churn   No. Observations:                 3453
Model:                        GLM   Df Residuals:                     3431
Model Family:            Gaussian   Df Model:                           21
Link Function:           Identity   Scale:                         0.11320
Method:                      IRLS   Log-Likelihood:                 -1127.3
Date:            Mon, 20 Nov 2023   Deviance:                        388.40
Time:                    05:48:02   Pearson chi2:                     388.
No. Iterations:                 3   Pseudo R-squ. (CS):             0.2997
Covariance Type:        nonrobust
==============================================================================
                          coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
const                   0.9799      0.255      3.843      0.000       0.480       1.480
CustomerID          -2.292e-05   5.24e-06     -4.376      0.000   -3.32e-05   -1.27e-05
Customer Name       -3.123e-05   2.54e-05     -1.232      0.218   -8.09e-05    1.85e-05
Tenure                 -0.0158      0.001    -19.295      0.000      -0.017      -0.014
PreferredLoginDevice   -0.0154      0.008     -1.881      0.060      -0.031       0.001
CityTier                0.0329      0.007      4.970      0.000       0.020       0.046
WarehouseToHome         0.0027      0.001      4.066      0.000       0.001       0.004
PreferredPaymentMode   -0.0072      0.004     -1.737      0.082      -0.015       0.001
Gender                  0.0200      0.012      1.689      0.091      -0.003       0.043
HourSpendOnApp          0.0018      0.008      0.233      0.816      -0.013       0.017
NumberOfDeviceRegistered 0.0460     0.006      7.337      0.000       0.034       0.058
PreferedOrderCat        0.0261      0.007      3.562      0.000       0.012       0.041
SatisfactionScore       0.0293      0.004      7.031      0.000       0.021       0.038
MaritalStatus           0.0738      0.009      8.251      0.000       0.056       0.091
NumberOfAddress         0.0197      0.002      8.325      0.000       0.015       0.024
Complain                0.2136      0.013     16.749      0.000       0.189       0.239
OrderAmountHikeFromlastYear -0.0019  0.002    -0.950      0.342      -0.006       0.002
CouponUsed              0.0041      0.005      0.856      0.392      -0.005       0.014
OrderCount              0.0150      0.004      3.696      0.000       0.007       0.023
DaySinceLastOrder      -0.0135      0.002     -7.284      0.000      -0.017      -0.010
CashbackAmount         -0.0002      0.000     -0.734      0.463      -0.001       0.000
AverageOrderValue       0.0013      0.002      0.607      0.544      -0.003       0.005
==============================================================================
```
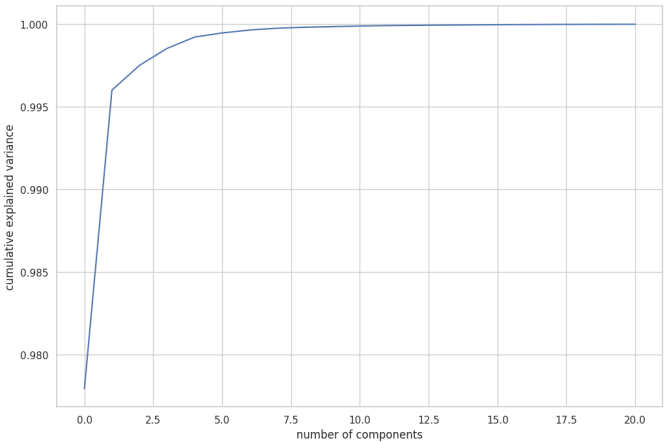
Fig. 20. Summary of Logistic Regression

```
array([[1168,   53],
       [ 120,  140]])
```

Fig. 21. Confusion matrix of Logistic Regression

```
Recall Score: 0.5384615384615384
Precision Score: 0.7253886010362695
Accuracy Score: 0.8831870357866306
```

Fig. 22. Logistic Regression Evaluation results

B. Decision Tree

With an accuracy score of 89.13pecent, the model proven proficiency in predicting both client retention and churn. The confusion matrix revealed 1158 genuine positives, effectively identifying clients probable to be retained, and 162 genuine negatives, as it should be identifying those possibly to churn. However, 63 false positives indicated times wherein the version expected retention incorrectly, and 98 fake negatives highlighted ignored possibilities to pick out customers who honestly stayed. The don't forget rating of 62.31percent implies that the version captured a great part of actual retentions.

Moreover, the precision rating of 72percent shows that, of the expected retentions, a large majority certainly stayed. While the general accuracy is outstanding, the commercial enterprise context need to manual the significance attributed to false positives and fake negatives, as precision and don't forget considerations might also range primarily based on precise purchaser retention techniques. Fine-tuning the version parameters or exploring ensemble techniques ought to in addition decorate its overall performance in figuring out and retaining treasured customers in the dynamic panorama of e-commerce.

```
array([[1158,   63],
       [  98,  162]])
```

```
Recall Score: 0.6230769230769231
Precision Score: 0.72
Accuracy Score: 0.8912896691424713
```

Fig. 23. Decision tree algorithm results

```
array([[1196,   25],
       [  80,  180]])
```

```
Recall Score: 0.6923076923076923
Precision Score: 0.8780487804878049
Accuracy Score: 0.9291019581363943
```

Fig. 24. Randomforest results

```
array([[1199,   22],
       [  52,  208]])
```

```
Recall Score: 0.8
Precision Score: 0.9043478260869565
Accuracy Score: 0.950033760972316
```

Fig. 25. Randomforest with hyperparameter tuning

C. Random Forest

When analyzing customer churn, in a Random Forest model the main objective is usually to predict whether a customer will stop using the service based on factors. To achieve this the model constructs decision trees. Combines their predictions.

Using a Random Forest model in this scenario has advantages. It can effectively handle a number of features. Is less prone to overfitting compared to individual decision trees. Additionally the models output can offer insights into the factors for predicting customer churn, which can be extremely valuable, for shaping business strategies and improving customer retention efforts.

We start by building a Random Forest model called model rfc pca with parameters. The trees, in this model have a depth of 10. We set the random state to 100. These choices aim to find a balance between model complexity and its ability to generalize well. The inclusion of "PCA" in the models name suggests that Principal Component Analysis might be employed during data preprocessing to reduce dimensionality.

Next we move on to tuning another model called model tuned rf using a state (42). The mention of KFold with 5 splits and shuffling enabled indicates the utilization of cross validation, for hyperparameter tuning or evaluating the performance of the model.

Comparing Random Forest business metric with and without hyper parameter tuning,

Random Forest without hyper parameter tuning accuracy Score: 0.9291019581363943 Random Forest with hyper parameter tuning accuracy Score: 0.950033760972316

VIII. Case Studies

A. Amazons Product Recommendation Engine Functions

Algorithmic Foundation: Amazons engine primarily relies on filtering and deep learning algorithms. Collaborative filtering analyzes the purchasing and browsing behaviors of customers to suggest products while deep learning algorithms delve into patterns within individual customer behaviors.

Data Utilization: The engine taps into a collection of data including customer purchase history items browsed, products in shopping carts and search history. This comprehensive data forms the foundation for the engines capabilities.

Dynamic Personalization: The recommendation system dynamically tailors the shopping experience for each customer by suggesting products they're likely to be interested in based on their behaviors and similar customers actions.

Personalization Enhances Customer Experience: Amazons model highlights that personalization plays a role, in enhancing customer experience. By offering recommendations customers feel. Valued, resulting in increased satisfaction and loyalty.

Utilizing Customer Data, for Personalized Suggestions: It's essential to utilize customer data. Amazons ability to use this data and provide product suggestions demonstrates the importance of understanding customer preferences and behaviors leading to meaningful interactions.

Adaptive Learning for Continuous Improvement: The engine's continuous learning algorithms ensure its adaptability and relevance, over time. As customer preferences evolve so do the recommendations resulting in a pertinent shopping experience.

Enhanced Customer Loyalty: These targeted suggestions not enhance the customer experience. Also drive sales. When customers are presented with items that align with their interests they are more likely to make purchases fostering loyalty.

Establishing Strong Customer Relationships: Beyond sales this predictive model helps build long term customer relationships. By meeting customer needs and preferences Amazon establishes a sense of trust and loyalty.

B. Netflix's Churn Prediction Model

Analytical Approach: Netflix's churn prediction model analyzes various factors such as subscribers' viewing patterns, subscription duration, and user interactions. This analysis is underpinned by sophisticated algorithms that detect subtle changes in user behavior that may indicate a likelihood to churn.

Data Utilization: The model leverages a vast array of data points, including the frequency and duration of viewing, types of content watched, and user engagement with the platform (such as ratings and search history). This comprehensive approach ensures a nuanced understanding of customer behavior.

Proactive Engagement: Upon identifying subscribers at a higher risk of churning, Netflix proactively engages them with personalized content and offers. This engagement is tailored to reignite their interest and enhance their overall experience with the platform.

Early Identification of At-Risk Customers: The ability to identify customers likely to churn is a significant advantage. By recognizing these signs early, Netflix can implement targeted strategies to re-engage these subscribers before they leave the platform.

Importance of Personalized Content: Personalization is a cornerstone of Netflix's strategy. The platform's success in retaining customers is largely attributed to its ability to recommend content that aligns with individual preferences, thus keeping subscribers engaged and less likely to churn.

Insight into Reasons for Churn: The predictive model provides valuable insights into why customers might be considering leaving the service. Understanding these reasons allows Netflix to address underlying issues, whether they relate to content, service quality, or user experience.

Adaptive Strategies for Retention: The churn model enables Netflix to adapt its retention strategies dynamically. As subscriber preferences and behaviors evolve, so do the efforts to keep them engaged, ensuring that the retention strategies remain effective over time.

Enhanced Customer Experience: By focusing on the needs and preferences of its subscribers, Netflix enhances the overall customer experience. This not only aids in retention but also bolsters the company's reputation for customer-centricity.

Data-Driven Decision Making: Netflix's model represents a shift towards data-driven decision-making in customer retention. By relying on analytics, the company can make informed decisions that directly impact its subscriber base.

IX. Project Management

A. Implementation Status Report

1) Work completed: Description: Raw data is cleaned and Exploratory Data Analysis is performed with various operations and visualised the data in many ways. Logistic Regression algorithm are implemented till now.

2) Responsibility: Keerthi Yanala: Performed Data Collecting, Cleaning and Preprocessing of Data.
Sreekar Thanda: Initialized Exploratory Data Analysis, Explored some data and visualized data using Histograms, Bar graphs, Pie-charts etc, Initialised Uni-variate Analysis.
SriHarsha Ponakala: Performed various Data Analysis such as Advanced Uni-variate and Bi-variate Analysis, Analysed correlation analysis and visualized data using Bargraphs, Pairplots, Heatmaps, Lineplots etc.
Pooja Bhathaluri: Performed more advancements of analysis for feature selection such as Dimensionality Reduction using PCA, built Logistic regression model, Decision Tree model, Random forest models. //

3) Contributions: Equal share is being contributed by each team member. Keerthi Yanala (25Percent), Sreekar Thanda (25Percent)) , Sri Harsha Ponakala (25Percent), Pooja Bhathaluri (25Percent).

Responsibility (Task, Person): Keerthi Yanala: Advanced boosting frameworks (XGBoost, LightGBM, CatBoost).
Sreekar Thanda: Methods for identifying unusual customer

behaviors and high-value customer segments.

SriHarsha Ponakala: Sentiment Analysis and Convolutional Neural Networks (CNNs)

Pooja Bhathaluri: Advanced time series analysis of ARIMA, SARIMA, and state space models. Issues/Concerns:

First and foremost collecting quality and relevant e commerce data is crucial, for this project. It's important to address privacy concerns while ensuring access to the data. Finding the balance between gathering data for accurate analysis without compromising user privacy can be challenging.

The accuracy and impartiality of the model are also considerations. Given the customer base in e commerce achieving an unbiased model presents a significant challenge. Careful thought must be given to the data used and the methodologies employed during model training.

Another essential aspect is designing a model that can adapt to changes in the market. E commerce is a field, with customer preferences and market trends shifting. The model should be flexible and updatable to stay relevant and effective over time. Additionally keeping up with advancements in machine learning and data analytics is vital for this projects success.

Lastly safeguarding user privacy and maintaining data security are priorities. It's crucial to strike a balance between using data for predictive analysis while ensuring robust measures are in place to protect user information. Implementing protocols, for handling data will help maintain trust while utilizing it appropriately. Each of these challenges comes with its difficulties. Effectively dealing with them will play a vital role in successfully implementing and maintaining the customer retention analysis model in the ever changing realm of e commerce.

REFERENCES

[1] B. G. Muchardie, A. Gunawan and B. Aditya, "E-Commerce Market Segmentation Based On The Antecedents Of Customer Satisfaction and Customer Retention," 2019 International Conference on Information Management and Technology (ICIMTech), Jakarta/Bali, Indonesia, 2019, pp. 103-108, doi: 10.1109/ICIMTech.2019.8843792.

[2] E. Y. Huang, C. -j. Tsui, W. K. Kuan, H. -S. Chen and M. -c. Hung, "Measuring Customer Retention in the B2C Electronic Business: An Empirical Study," 2013 46th Hawaii International Conference on System Sciences, Wailea, HI, USA, 2013, pp. 2900-2907, doi: 10.1109/HICSS.2013.396.

[3] S. Koul and T. M. Philip, "Customer Segmentation Techniques on E-Commerce," 2021 International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE), Greater Noida, India, 2021, pp. 135-138, doi: 10.1109/ICACITE51222.2021.9404659.

[4] Renjith, Shini. (2017). B2C E-Commerce Customer Churn Management: Churn Detection using Support Vector Machine and Personalized Retention using Hybrid Recommendations. International Journal on Future Revolution in Computer Science & Communication Engineering (IJFRCSCE). 3. 34 – 39. 10.6084/M9.FIGSHARE.5579482.

[5] Kuan, Huei Bock, Gee-Woo Vathanophas, Vichita. (2005). Comparing the Effects of Usability on Customer Conversion and Retention at E-Commerce Websites. 10.1109/HICSS.2005.155.

[6] Kumar, Vikas Ogunmola, Gabriel. (2020). E-retail factors for customer activation and retention: An empirical study from Indian e-commerce customers. Journal of Retailing and Consumer Services. 59. 10.1016/j.jretconser.2020.102399.

[7] Pondel, Maciej Wuczyński, Maciej Gryncewicz, Wieslawa Łysik, Łukasz Hernes, Marcin Rot, Artur Kozina, Agata. (2021). Deep Learning for Customer Churn Prediction in E-Commerce Decision Support. Business Information Systems. 3-12. 10.52825/bis.v1i.42.