IOSR Journal of Computer Engineering (IOSR-JCE)
e-ISSN: 2278-0661,p-ISSN: 2278-8727, Volume 26, Issue 2, Ser. 3 (Mar. – Apr. 2024), PP 50-61
www.iosrjournals.org

Designing A Predictive Model To Uncover Consumer Spending Patterns In India

Umarfaruk Muhammed Sidat[1], Ahmad Said Moldhariya[2], Fatai Kareem[3]
[3]*(Management And Accounting, Obafemi Awolowo University, Nigeria)*

Abstract
This study examines credit card transactions in India to analyze spending behaviors influenced by factors such as gender, card type, and location. The primary objective is to elucidate the diverse shopping patterns across demographic groups and cities, thereby offering insights into consumer lifestyles and regional economic activities. This analysis provides valuable data for businesses and policymakers, enhancing decision-making in India's evolving marketplace. The research methodology involves a rigorous examination of transaction data from a dataset available on kaggle, employing advanced predictive modeling techniques to decode spending behaviors. The data is categorized based on demographic and geographical attributes, facilitating the identification of significant trends and variations in consumer spending. This approach provides a granular analysis of consumer behavior, highlighting specific preferences and spending methods prevalent among various segments of the Indian population. The findings reveal considerable disparities in spending across different groups and cities, reflecting India's heterogeneous economic landscape. Insights into consumer expenditure not only aid in tailoring products and services to diverse consumer bases but also enhance the ability to predict future market trends. Fundamentally, this study equips businesses and policymakers with strategic insights to develop more effective marketing strategies and optimize resource allocation within the economy, thereby promoting targeted marketing and economic efficiency.

Key Word *- predictive model, spending patterns,*

Date Of Submission: 08-04-2024 Date Of Acceptance: 18-04-2024

I. Introduction

Knowledge of how consumers spend money is very important in the changing marketplaces of today. It gives useful understanding for businesses and policymakers about economic trends as well as what kinds of things people like to buy. In India's diverse and quickly changing environment, studying consumer spending has become more important for adjusting products and services to different groups of people (Malik, P. (2016). This research aims to understand the details of consumer behavior by looking at information from credit card transactions. The goal is to discover the complicated elements that affect spending habits among many demographic groups and geographic places in India.

In this research, the main goals are three. First is to look at spending habits according to gender and card type. The focus is on different demographic groups and breaking down their buying behaviors for understanding subtle variations in what they buy or like to purchase. Second goal involves examining how spending gets spread out across various cities within India - it's about finding areas having diverse levels of economic activity based on where people spend money most frequently when using credit cards there. Lastly, the study wants to look into the kinds of costs brought about by using credit cards for transactions. It also hopes to understand more about what consumers like as well as their way of life.

This study has significant potential for useful understanding to benefit businesses, policymakers and researchers. The focus on the Indian market could help in making informed decisions by using advanced forecasting models. This research is hoping to reveal the secrets behind consumer behavior. The aim is not only giving suggestions that can be used for business growth and policy making but also helping understand more about complicated forces which influence how people spend money in India. It could then guide towards strategies that are better targeted and successful in business as well as economics overall.

Need for the Study

Cybercrime is a big problem in India's digital world. It includes many harmful actions like stealing data, tricking people for money, taking someone's identity and demanding ransom through software attacks. Due to fast growth of digital technology and more critical systems becoming digitized, the threat from cyber crimes has become very serious for India. For this reason, using machine learning (ML) offers an optimistic method to improve safety against cybersecurity threats and reduce danger linked with cybercrime.

Rising Cyber Threat Landscape: India's crime scene is becoming more digital, with reports of a big increase in cybercrime incidents like hacking attempts, phishing scams and malware infections. As the types of cyber threats change quickly, it becomes harder for law enforcement agencies, government bodies and businesses to handle them all effectively. This demands fresh ways to fight against cybercrime correctly (Kaloudi, N., & Li, J. 2020).
Complexity of Cyber Attacks: The actions of cybercriminals are becoming more complex, using advanced methods to avoid typical security tactics. This makes it harder to discover and stop cyber attacks as they happen in real-time. Traditional systems that depend on rules and finding signatures often find it hard to keep up with the changing danger picture, showing why there is a requirement for flexible and advanced defense tactics.
Role of Machine Learning: ML presents a new way in cybersecurity, providing automated abilities for identifying threats, finding anomalies and recognizing patterns. ML algorithms are capable of examining large amounts of data to discover hidden patterns that suggest malicious actions and provide useful information for improving strategies against cyber threats. By using methods such as supervised learning, unsupervised learning and deep learning from machine intelligence (MI), groups can improve their resistance to cyber risks and strengthen the security of their digital systems (Apruzzese, et al.2023)..
Need for Contextual Analysis: Classic cybersecurity methods might depend on inflexible rules and signatures, which could not adjust to changing cyber dangers and progressing attack paths. Algorithms of machine learning are excellent in providing context analysis, this gives them the ability to distinguish normal behavior from abnormal behavior and recognize slight deviations that imply cyber attacks. Through utilizing the strength of ML-based anomaly detection, groups can find and deactivate new threats before they cause serious harm.
Importance of National Cybersecurity: As India progresses towards a digital economy and welcomes new technologies like cloud computing, artificial intelligence, and Internet of Things (IoT), it is very important to protect crucial digital systems. A strong cybersecurity structure that uses ML for threat understanding and predictive analysis helps in safeguarding sensitive data, maintaining trustworthiness in the digital world and protecting national security benefits. Cybersecurity has become a critical concern for every country as they depend more on their digital systems. When countries improve their online abilities to support social, economic and government activities, it's necessary that they also have effective protection against cyber attacks. This is especially essential because many vital services such as electricity grids or financial networks are nowadays managed via computer networks which makes them vulnerable targets for hackers or other malicious actions from cyberspace (Biswas & Bhattacharya 2021).

Considering these elements, a solid reason for carrying out research on using machine learning to lessen cybercrime in India is present. By investigating fresh ML-focused methods for discovering, forecasting and lessening cyber threats, researchers could aid in creating powerful cybersecurity answers designed to tackle the distinct hurdles and intricacies of India's cyber scenario. This type of research could possibly help in making India more resistant to cyber issues, promoting digital creativity and securing the country's digital future as our world gets more connected through technology.

II. Literature Review

Cybercrime is a major danger to digital systems, and its tactics are always changing as technology progresses. The use of machine learning (ML) methods offers an encouraging way to fight against cyber threats by improving detection, prediction and lessening actions. Work by Mathew et al (2021) looks into using ML algorithms for diminishing cybercrime in India, concentrating on the investigation of consumer behavior data from e-commerce platforms. The authors emphasize the power of ML, especially in areas like sentiment analysis, information extraction and user influence analysis to discover patterns that could show signs of cybercriminal activities.

Analysis of consumer behavior is very important for knowing transaction activities and finding strange patterns that might indicate fraud (Mohanty et al 2014). In this study, by using ML algorithms like SeqLearn, the goal is to understand the interests and habits of consumers based on the e-commerce data so as to predict how they will pay in future. This matches with earlier research that highlighted the need to analyze online consumer behavior in order to detect cyber dangers and protect digital transactions (Bollen et al., 2011; Asur et al., 2011). This study shows the importance of predictive modeling, which can help improve cybersecurity by allowing interventions before problems happen.

Also, the research makes clear how important it is to use network data analysis for finding cyber threats and discovering relationships between nodes in digital groups. Taking inspiration from methods used in social network analysis, the study investigates group behavior studies and information spreading dynamics (De, & Chattopadhyay, 2020).This gives understanding about how online activities are connected and what this means for keeping things safe (Gonzalez-Bailon et al., 2011). By using analytics on network data, organizations can understand better changing cyber threats. This helps them make strong defense systems and get ready ahead of time.

In conclusion, the studies show that the use of predictive modeling and network data analysis helps to improve cybersecurity and protect digital transactions from changing dangers. This has important implications for policy makers, businesses and those working in cybersecurity. The studies suggest the need to take active steps that use advanced ML methods in order to reduce cyber dangers and maintain trust in our digital world during this time of widespread connection.

III. Research Methodology

Data Analysis Techniques

The first step of the research was Exploratory Data Analysis (EDA) and Data Visualization. This is very important to understand all aspects of the dataset. EDA helped in summarizing main characteristics, identifying patterns, anomalies and distributions without any prior assumptions. A close look at data's inherent properties was possible through statistical summaries as well as visual methods.

Techniques of Data Visualization were used a lot to explain more clearly the connections in data. These helped to understand complex patterns and trends by using methods such as histograms, scatter plots, and box plots. These methods provide a better understanding about how the data is spread and how variables interact with each other.

Predictive Modeling

The primary aim of the study was to carry out a binary classification task, which is foreseeing customer churn (Nickerson, R. S. (1972). For this purpose,the Random Forest model was used. It's an algorithm that has proven its strength and effectiveness in dealing with two possible outcomes like yes or no responses. Random Forest is a group or ensemble of decision trees known for having high accuracy in classification problems. It works by constructing numerous decision trees in the training phase and giving out the class that represents the mode of the classes (classification) for individual trees. In this dataset, Random Forest model displayed an impressive accuracy of 93%, showing its reliable strength in forecasting if a customer would leave or stay on.

Customer Segmentation Analysis

In order to divide customers into groups, the K-means clustering was used. This method is a kind of unsupervised learning that helps in finding clusters within the data. The main advantage of K-means clustering was its straightforwardness and effectiveness in forming clusters. The method of k-means clustering that divided the data into k distinct groups, with each group represented by the center point of its members, made it simple to interpret how the data was segmented. This helped in comprehending what behavioral and demographic patterns are unique to separate customer segments.The use of K-means clustering gave a good view of the separate parts in the customer group. This is important for focused marketing plans and tailored customer involvement ideas. (Kodinariya, T. M., & Makwana, P. R.2013).

V. Results And Discussion

Customer Segmentation and Attrition Analysis

The Bank Chunners dataset originally contains 10,127 rows and 23 columns.

Figure 1: Column Check

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 10127 entries, 0 to 10126
Data columns (total 23 columns):
 #   Column                    Non-Null Count  Dtype
---  ------                    --------------  -----
 0   CLIENTNUM                 10127 non-null  int64
 1   Attrition_Flag            10127 non-null  object
 2   Customer_Age              10127 non-null  int64
 3   Gender                    10127 non-null  object
 4   Dependent_count           10127 non-null  int64
 5   Education_Level           10127 non-null  object
 6   Marital_Status            10127 non-null  object
 7   Income_Category           10127 non-null  object
 8   Card_Category             10127 non-null  object
 9   Months_on_book            10127 non-null  int64
 10  Total_Relationship_Count  10127 non-null  int64
 11  Months_Inactive_12_mon    10127 non-null  int64
 12  Contacts_Count_12_mon     10127 non-null  int64
 13  Credit_Limit              10127 non-null  float64
 14  Total_Revolving_Bal       10127 non-null  int64
 15  Avg_Open_To_Buy           10127 non-null  float64
 16  Total_Amt_Chng_Q4_Q1      10127 non-null  float64
 17  Total_Trans_Amt           10127 non-null  int64
 18  Total_Trans_Ct            10127 non-null  int64
 19  Total_Ct_Chng_Q4_Q1       10127 non-null  float64
 20  Avg_Utilization_Ratio     10127 non-null  float64
 21  Naive_Bayes_mon_1         10127 non-null  float64
 22  Naive_Bayes_mon_2         10127 non-null  float64
dtypes: float64(7), int64(10), object(6)
memory usage: 1.8+ MB
```

Source: Author's computation (2024)

There was no column with null values.

Figure 2: Missing values check

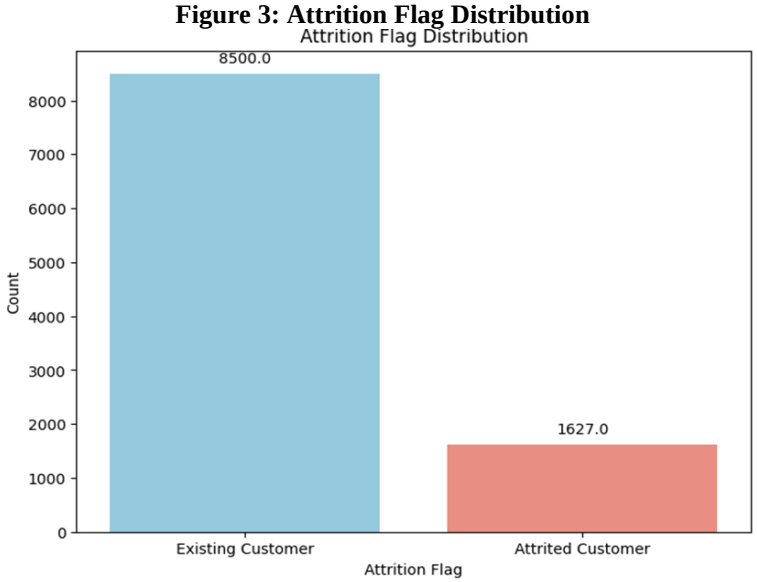
```
# Check for missing values
df.isnull().sum()
```

```
CLIENTNUM                 0
Attrition_Flag            0
Customer_Age              0
Gender                    0
Dependent_count           0
Education_Level           0
Marital_Status            0
Income_Category           0
Card_Category             0
Months_on_book            0
Total_Relationship_Count  0
Months_Inactive_12_mon    0
Contacts_Count_12_mon     0
Credit_Limit              0
Total_Revolving_Bal       0
Avg_Open_To_Buy           0
Total_Amt_Chng_Q4_Q1      0
Total_Trans_Amt           0
Total_Trans_Ct            0
Total_Ct_Chng_Q4_Q1       0
Avg_Utilization_Ratio     0
Naive_Bayes_mon_1         0
Naive_Bayes_mon_2         0
dtype: int64
```

Source: Author's computation (2024)

From Our Dataset there were 8,500 existing customers and 1,627 attrited customers

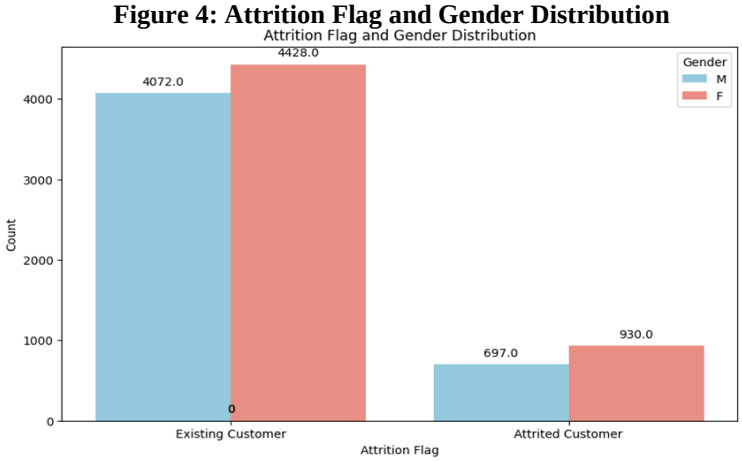
Figure 3: Attrition Flag Distribution



Source: Author's computation (2024)

Total percentage of male customers that attrited are 42.8% and the total percentage of female customers that attrited are 5716%. More female customers attrition than male customers.

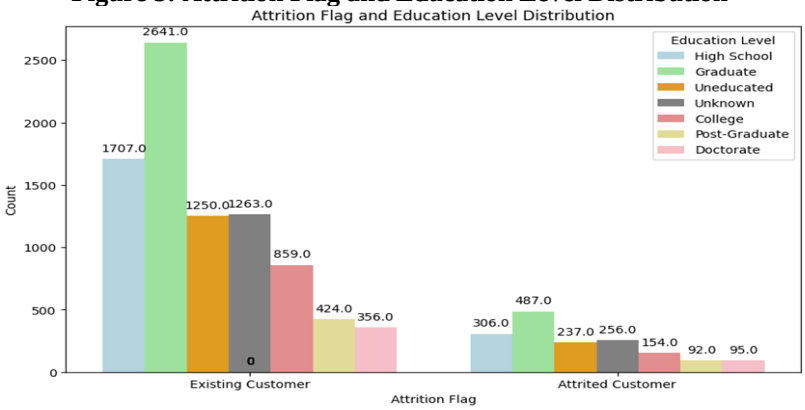
Figure 4: Attrition Flag and Gender Distribution



Source: Author's computation (2024)

Customers with graduate education level attrited more than all other customers by 29.93%

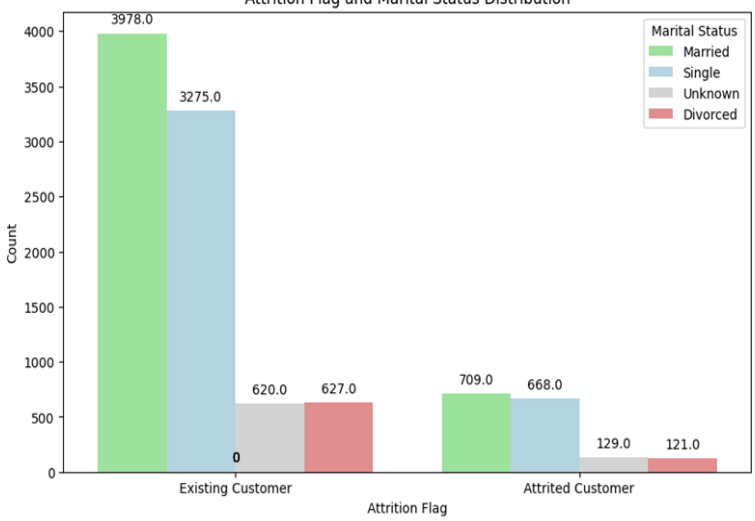
Figure 5: Attrition Flag and Education Level Distribution



Source: Author's computation (2024)

Married customers attrited more by 43.58% with single customers closely behind with 41.1% attrition rate

Figure 6: Attrition Flag and Marital Status Distribution



Source: Author's computation (2024)

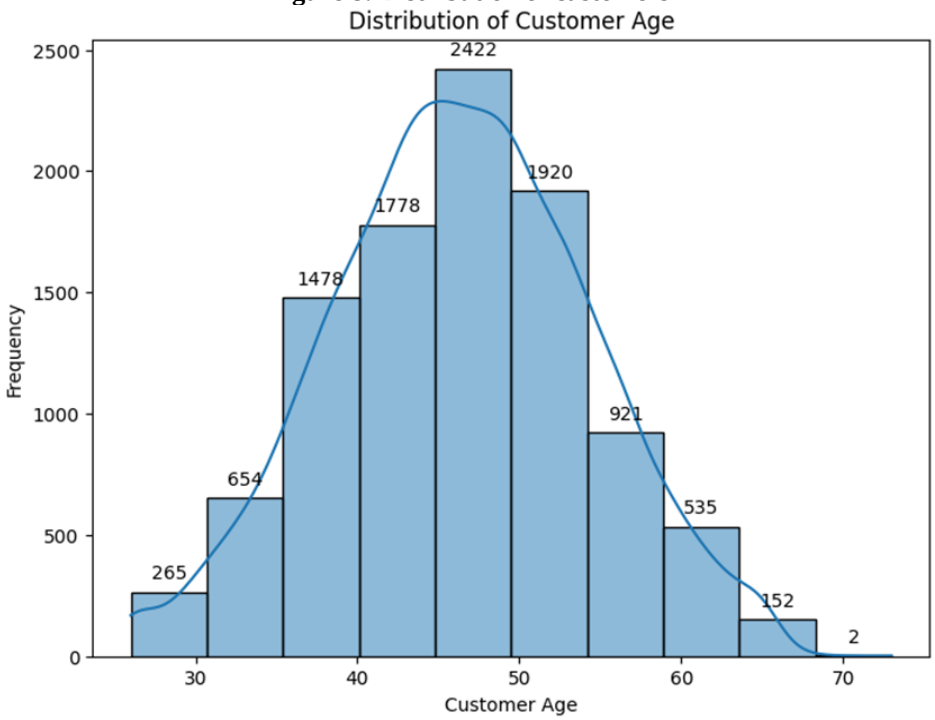
Customers earning less than $40k attrited more than other customers by 37.62%. Making Income a significant factor as to why customers attrited.

Figure 7: Attrition Flag and Income Distribution

Source: Author's computation (2024)

Many customers are between age 45 to 50

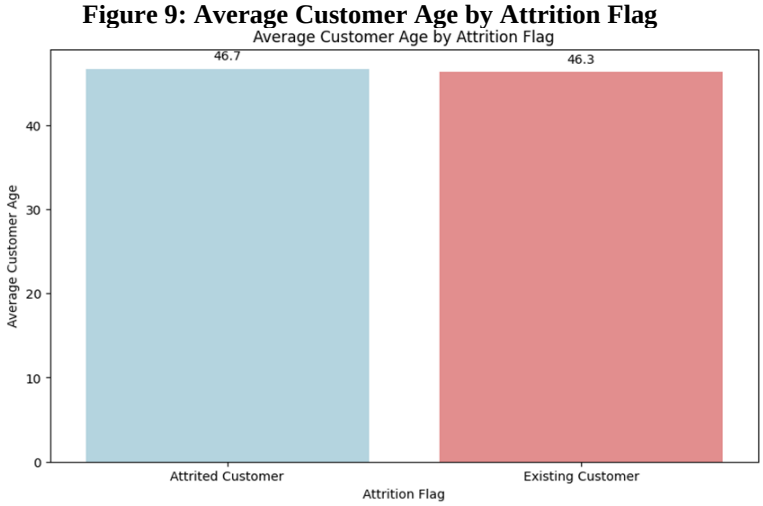
Figure 8: Distribution of customers



Source: Author's computation (2024)

There is no significant difference in the average age of customers who stayed and who attrited both are age 46.

Figure 9: Average Customer Age by Attrition Flag

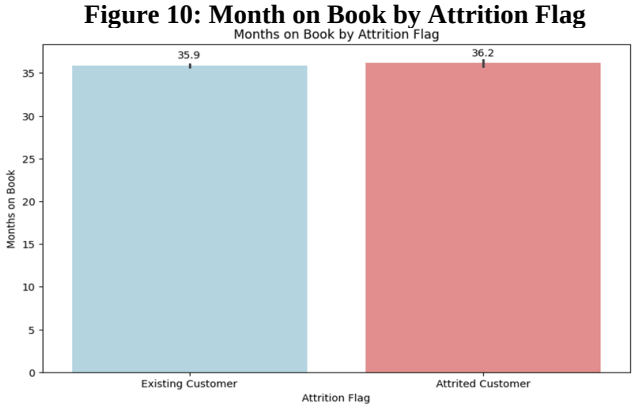


Source: Author's computation (2024)

Customers who are still with the company (existing customers) have been customers for approximately 35.9 months. Which is roughly 2 years and 11 months.

While customers who have churned or left the company (attrition customers) had been customers for around 36.2 months. Which is about 3 years.

Figure 10: Month on Book by Attrition Flag

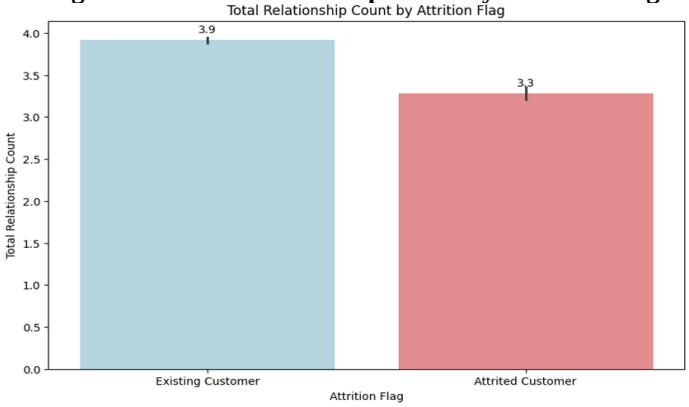


Source: Author's computation (2024)

Existing customers have approximately 3.9 relationship ratings with the credit card provider.

On the other hand, customers that have left the company or churned, have an average of 3.3 relationship ratings with the credit card provider. This indicates that, on average, attrition customers had fewer relationships with the provider compared to existing customers.

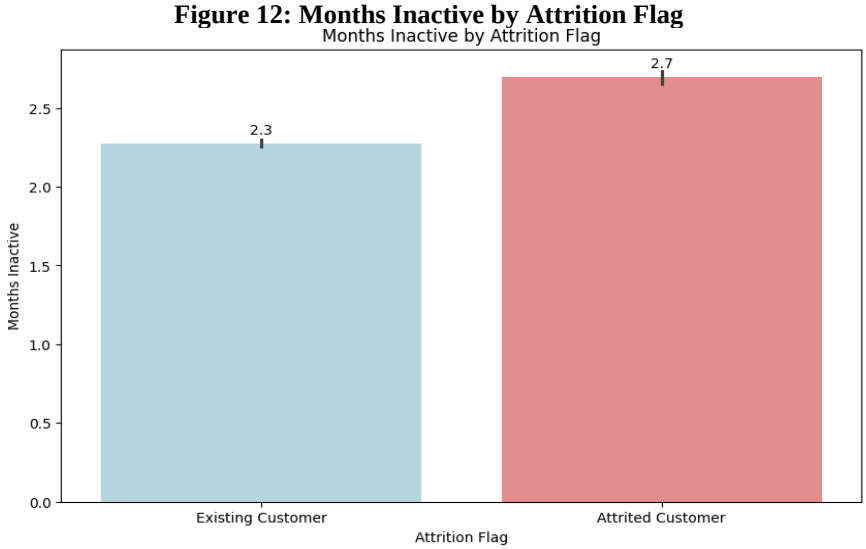
Figure 11: Total Relationship Count by Attrition Flag



Source: Author's computation (2024)

Existing customers were inactive for approximately 2.3 months within the last twelve months. On the other hand, attrition customers, who are customers that have left the company or churned, were inactive for an average of 2.7 months within the last twelve months. This indicates that, on average, customers who eventually churned were inactive for slightly longer periods compared to existing customers.

Figure 12: Months Inactive by Attrition Flag

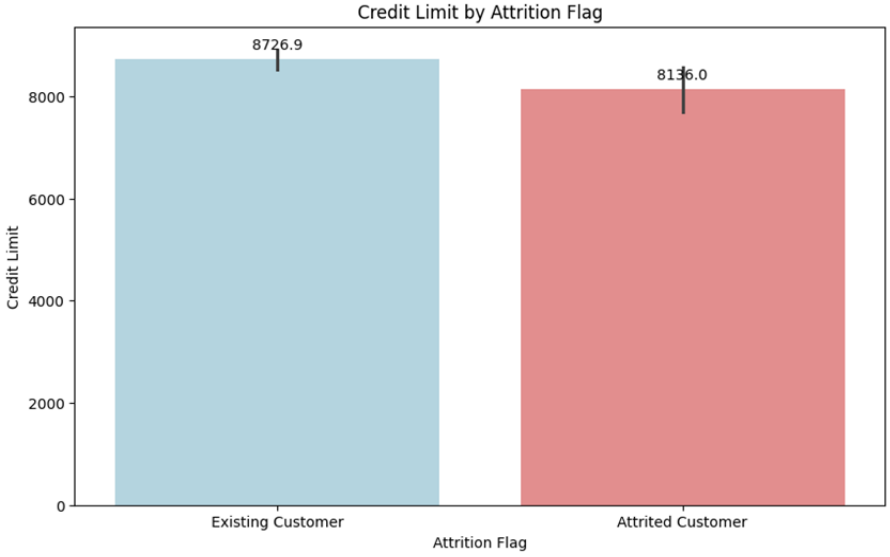


Source: Author's computation (2024)

Existing customers have a credit limit of approximately $8,726.90. The credit limit refers to the maximum amount of credit that a customer can access or borrow from their credit card provider.

On the other hand, attrition customers, who are customers that have left the company or churned, have an average credit limit of approximately $8,136.00. This indicates that, on average, customers who eventually churned had a slightly lower credit limit compared to existing customers.

Figure 13: Credit Limit by Attrition Flag

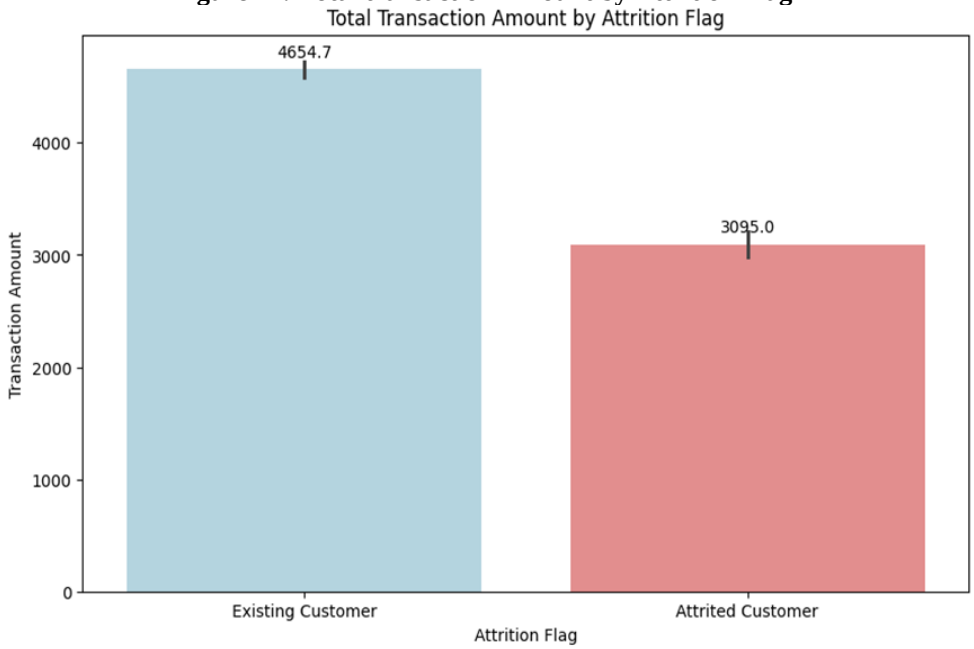


Source: Author's computation (2024)

Existing customers have a total transaction amount of approximately $4,654.70.

On the other hand, attrition customers, who are customers that have left the company or churned, have an average total transaction amount of approximately $3,095.00. This indicates that, on average, customers who eventually churned had a lower total transaction amount compared to existing customers.

Figure 14: Total transaction Amount by Attrition Flag



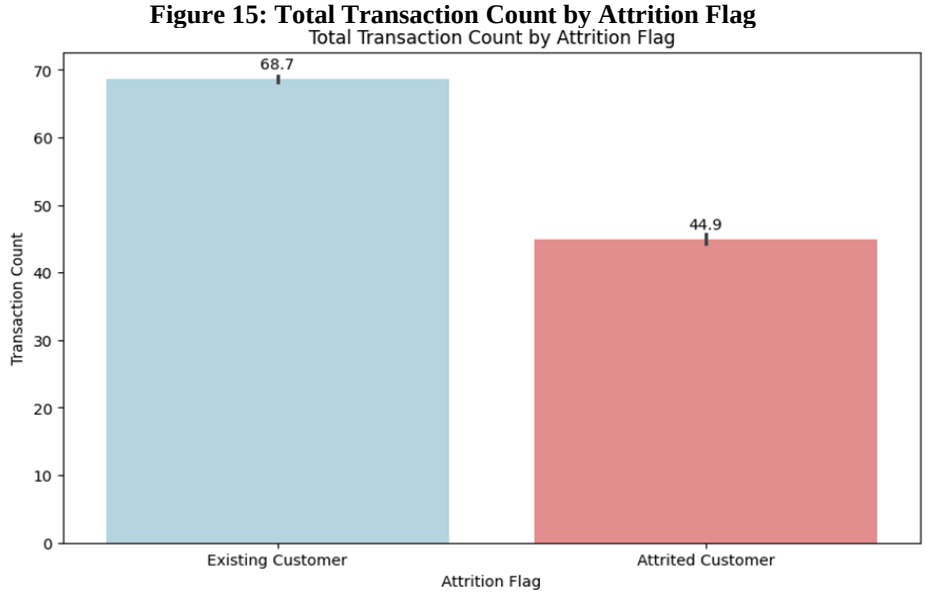
Source: Author's computation (2024)

Existing customers have a total transaction count of approximately 68.7 transactions.

On the other hand, attrition customers, who are customers that have left the company or churned, have an average total transaction count of approximately 44.9 transactions.

This indicates that, on average, customers who eventually churned had a lower total transaction count compared to existing customers.

Figure 15: Total Transaction Count by Attrition Flag



Source: Author's computation (2024)

Machine Learning Models Prediction Outcome

Decision Tree Metrics: Our decision tree model is accurate by 91%

Accuracy: 0.905890095426127

AUC-ROC: 0.8325990543300396

Precision: 0.9458711971552746

Recall: 0.9414077860794338

Random Forest Metrics: Our random forest model is accurate by 93%
Accuracy: 0.933201711089174
AUC-ROC: 0.8408038987479862
Precision: 0.9441913439635535
Recall: 0.9779787652379079

Naive Bayes Metrics: Our naive bayes model is accurate by 85%
Accuracy: 0.8496215860480422
AUC-ROC: 0.7186441591719204
Precision: 0.9077404222048475
Recall: 0.913094769956744

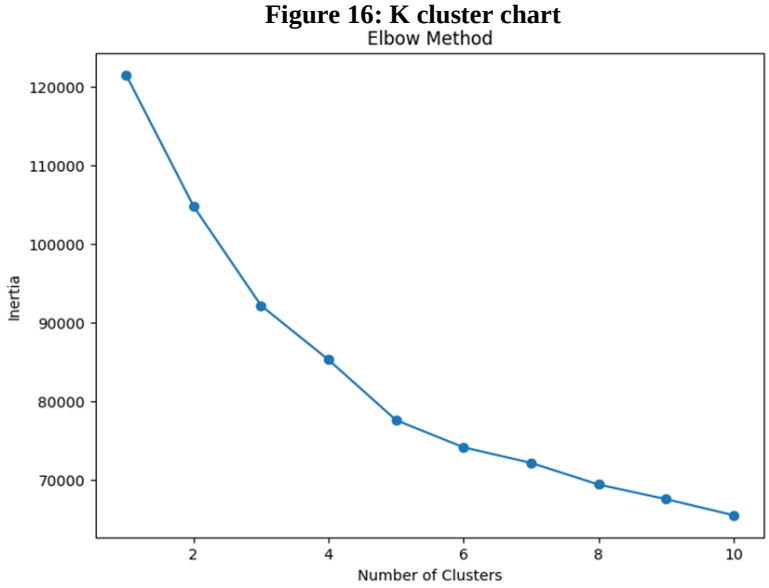
Support Vector Machine Metrics(SVM): Our SVM model is accurate by 84%
Accuracy: 0.836788417242514
AUC-ROC: 0.5
Precision: 0.836788417242514
Recall: 1.0

The most preferred machine learning model would be random forest.

Customer Segmentation with K-means Clustering

The Elbow Chart shows that there is a significant improvement in the clustering when the number of clusters is increased from 2 to 4. However, the inertia starts to level off after 4 clusters, which suggests that there is not a significant improvement in the clustering by increasing the number of clusters beyond 4.

Figure 16: K cluster chart



Source: Author's computation (2024)

Customer Age: For Cluster 0, the typical age of customers is about 46 years. Similarly, in Cluster 1 it is around 46 years and for Cluster 2 it comes out to be approximately 46 years too. These clusters seem to possess similar average ages among their clients.

Education Level: The average education level of customers in Cluster 0 is 3.12, for Cluster 1 it is 3.03 and for Cluster 2 it stands at 3.17. There is no big difference in education levels among the three clusters.

Marital Status: On average, the marital status of Cluster 0 is about 1.43, for Cluster 1 it's approximately 1.48, and in Cluster 2 it comes out to be around 1.51. There isn't a big difference shown by marital status across clusters either.

Income Category: The average income category for Cluster 0 is 3.07, for Cluster 1 it is 2.95 and for Cluster 2 it stands at 2.15. It seems that Cluster 2 has customers with higher income categories compared to the other clusters.

Months on Book: The months on book demonstrates how many months the individual has been with this company. Cluster 0 has a mean of about 35.77 months, Cluster 1 possesses an average of roughly 36.15

months while Cluster 2 stands at around 35.87 month period in total time spent by its members - all three values are very similar across clusters.

Contacts Count 12 months: The average contacts done by customers in the last 12 months is about 2.33 for Cluster 0, approximately 2.63 for Cluster 1 and around 2.41 for Cluster 2. In comparison to other clusters, it seems that Cluster number one has a bit more average contact count on an yearly basis from its customers (Brewer et al., n.d.).

Credit Limit: Cluster 0 has a credit limit of $3,937 on average, Cluster 1 has $6,368 and Cluster 2 has $25,252. Customers in cluster 2 are having significantly higher credit limits compared to the other clusters.

Total Revolving Balance: The whole sum of money that is owed on credit cards. For Cluster 0, the average amount is $1,695; for Cluster 1 it's $450 and finally in Cluster 2 it comes out to be around $1,259. Clusters 0 and 2 show higher averages for revolving balance it is in Cluster number one.

Average Open to Buy: This shows the average credit amount that can be used. Cluster 0 has an average of $2,242, Cluster 1 has $5,917 and Cluster 2 has $23,994. Just like with credit limit, the third cluster also holds a significantly higher typical number for open to buy compared to its counterparts.

Total Transaction Amount: The average total transaction amount is $4,418 for Cluster 0, $3,390 for Cluster 1 and $6,440 for Cluster 2. This indicates that the normal transaction size in Cluster 2 is greater than in other clusters.

Total Transaction Count: Cluster 0 has an average of 67 transactions, Cluster 1 has 57 transactions and Cluster 2 has the highest with a total of 74 transactions.

Average Utilization Ratio: This shows the average proportion between credit used and credit limit. For Cluster 0, it is 0.526; for Cluster 1, this ratio stands at just 0.068 while in Cluster 2 it drops even further to merely 0.057 (A). The first cluster has a notably higher average utilization ratio than other clusters.

Figure 17: Clusters

Cluster	Customer_Age	Education_Level	Marital_Status	Income_Category
0	46.265311	3.119238	1.433672	3.071630
1	46.463349	3.032729	1.476062	2.952664
2	46.200111	3.169154	1.513543	2.152018

Cluster	Months_on_book	Contacts_Count_12_mon	Credit_Limit
0	35.773209	2.330232	3937.022982
1	36.151204	2.632134	6367.620747
2	35.869541	2.413488	25252.469873

Cluster	Total_Revolving_Bal	Avg_Open_To_Buy	Total_Trans_Amt
0	1695.397966	2241.625016	4418.252326
1	450.193400	5917.427346	3390.239383
2	1258.716971	23993.752902	6439.868988

Cluster	Total_Trans_Ct	Avg_Utilization_Ratio
0	67.198442	0.525560
1	57.453070	0.068266
2	74.016584	0.056856

Source: Author's computation (2024)#

IV. Conclusion

The study that was done using credit card transaction data to understand consumer spending in India, has given a complete view of the complicated way different groups and places spend money. The goals of this research were met as it found unique spending habits according to gender, card type and city. This provides a more detailed insight into the economic actions among different parts of the population.Through the use of complex predictive modeling methods, the study was capable of emphasizing important variances in purchasing behaviors. These differences are key for businesses who are looking to enhance their marketing tactics and product selections. This kind of analysis is not just useful for comprehending what consumers like at present, it also assists in predicting upcoming patterns - an important aspect if one wants to stay competitive and pertinent within a market that changes quickly.To sum it up, this research provides important understanding that can lead to good choices for businesses and policymakers in the Indian market. The results highlight how crucial focused methods are in marketing and making products to serve well the many requirements of customers. As India keeps

changing, growing economically and technologically, these insights based on data will be crucial for encouraging development and creativity in the commercial environment.

References

[1] Asur, S., & Huberman, B. A. (2011). Predicting The Future With Social Media. Hp Laboratories Technical Report, 2, 2011.
[2] Bollen, J., Mao, H., & Zeng, X. (2011). Twitter Mood Predicts The Stock Market. Journal Of Computational Science, 2(1), 1-8.
[3] Gonzalez-Bailon, S., Borge-Holthoefer, J., Rivero, A., & Moreno, Y. (2011). The Dynamics Of Protest Recruitment Through An Online Network. Scientific Reports, 1, 197.
[4] Mathew, A., Et Al. (2021). Leveraging Machine Learning To Reduce Cybercrime In India. Mathematical Problems In Engineering, 2021, 6688750.
[5] Malik, P. (2016). Spending Patterns Of Indian Population. Journal Of Business Management And Economics, 4(10), 1-3.
[6] Mohanty, S. K., Dubey, M., & Parida, J. K. (2014). Economic Well-Being And Spending Behaviour Of Households In India: Does Remittances Matter?. Migration And Development, 3(1), 38-53.
[7] De, K., & Chattopadhyay, P. (2020). Influence Of Gender And Financial Independence On Consumer Spending Habits In Kolkata-An Empirical Assessment. Editorial Board, 22.
[8] Nickerson, R. S. (1972). Binary-Classification Reaction Time: A Review Of Some Studies Of Human Information-Processing Capabilities. Psychonomic Society, Incorporated.
[9] Kodinariya, T. M., & Makwana, P. R. (2013). Review On Determining Number Of Cluster In K-Means Clustering. International Journal, 1(6), 90-95.
[10] Kaloudi, N., & Li, J. (2020). The Ai-Based Cyber Threat Landscape: A Survey. Acm Computing Surveys (Csur), 53(1), 1-34.
[11] Apruzzese, G., Laskov, P., Montes De Oca, E., Mallouli, W., Brdalo Rapa, L., Grammatopoulos, A. V., & Di Franco, F. (2023). The Role Of Machine Learning In Cybersecurity. Digital Threats: Research And Practice, 4(1), 1-38.