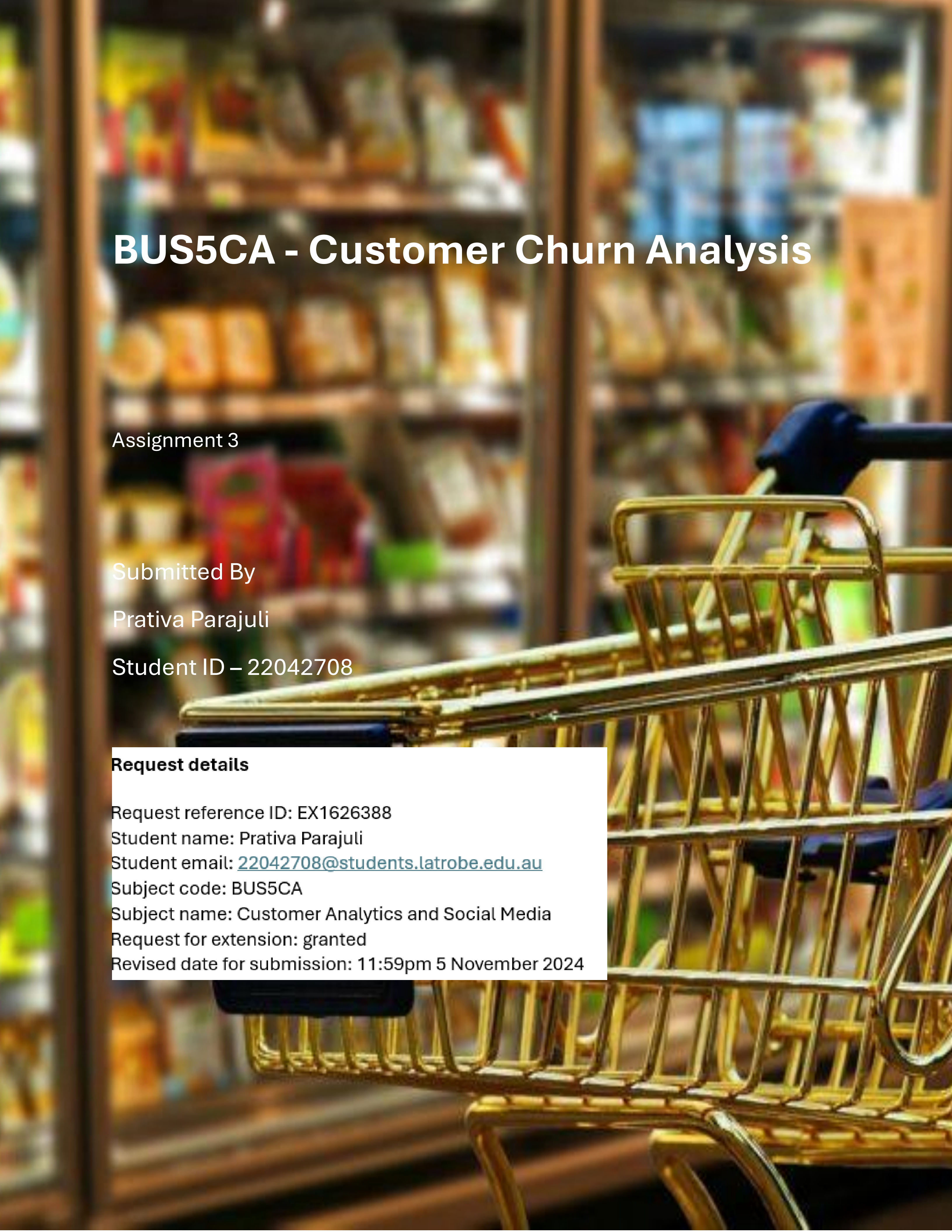


BUS5CA - Customer Churn Analysis

Assignment 3

Submitted By

Prativa Parajuli

Student ID – 22042708

Request details

Request reference ID: EX1626388
Student name: Prativa Parajuli
Student email: 22042708@students.latrobe.edu.au
Subject code: BUS5CA
Subject name: Customer Analytics and Social Media
Request for extension: granted
Revised date for submission: 11:59pm 5 November 2024

Table of Contents

Introduction

Customer Retention has become a crucial part of every organization aiming for sustainability and profit maximization, especially when the businesses mature beyond their initial expansion phases. In today's competitive markets, customer churn can significantly impact the revenue streams and customer lifetime value. Effective churn management not only minimizes these risks but also enhances customer relationship management (CRM) by fostering long-term loyalty and engagement.

ABC E-commerce, a leading online retailer with a diverse global customer base, has had significant growth in recent years. Nonetheless, with expansion, the challenges of maintaining customer loyalty and retaining them has be a challenge leading to threat to company's revenue and long-term sustainability.

To embark on this issue, e-commerce has initiated a comprehensive churn analysis project leveraging vast customer data collected over time- including demographics, purchase history, feedback and engagement metrics. The report aims to analyze the data for insights into churn behavior, develop and evaluate predictive models and recommend some data-driven strategies to improve customer retention. Also, the report suggests some actionable recommendations for targeted retention campaigns to strengthen customer loyalty and reduce churn.

Task 1: Understanding Characteristics of Churned, Non-Churned and High-Value Customers

To explore the given dataset, I have chosen Excel and Tableau.

Excel

Since the dataset does not comprise of any indication of who the high-value customer is, two columns are generated.

kAmou	Total Charge	High-Value Indicator
271.81	33	1
167.84	24	0
155.67	15	0

Initially, a column named “Total Charge” is created using the sum of Tenure, Days since last order, Order count, Coupons Used and Satisfaction Score for each of the customers.

Afterwards, using an “IF” function where a customer with total charge above 25 is considered of High Value, a column named “High Value Customer” is created where 1 denotes that the customer is of High-Value to the company and 0 denoting that the customer is not of High-Value to the company.

Total Charge	High-Value Indicator
33	=IF(V2>25,1,0)

Tableau

The final prepared dataset is imported in the Tableau to further explore the features of churned, non-churned and high-value customers.

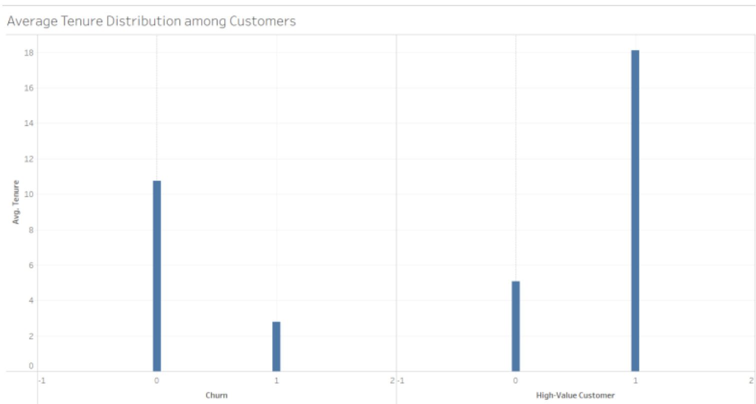
Let’s have a look at the visualizations created.

Tenure

In the bar chart as follows, we can see the average tenure of customers across different customer segments based on churn and high-value status.

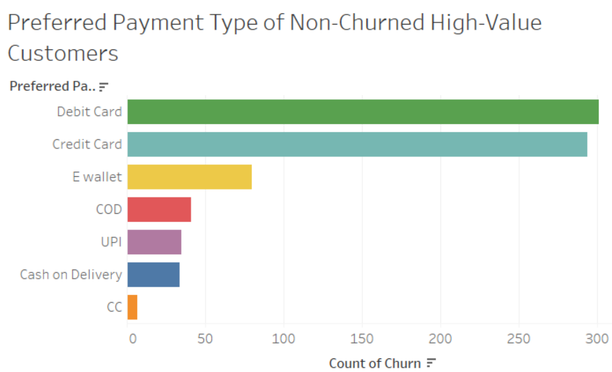
The non-churned customers (0) seems to have significantly higher average tenure than the churned customers suggesting that the longer the tenure, the higher the customers are likely to stay which correlates with the loyalty of customers.

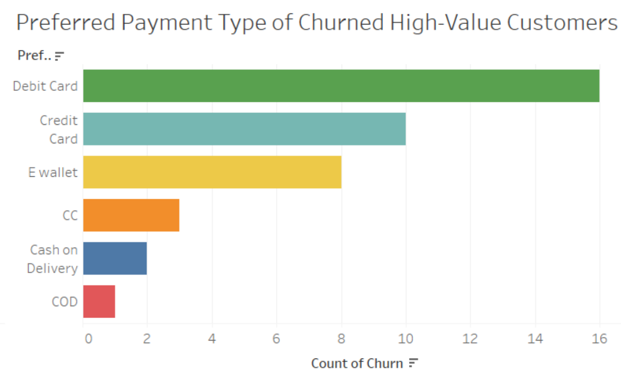
On the other hand, high-value customers have a higher average than the other, depicting that tenure could also be an indicator of customer value.



Preferred Payment Type

Here, the preferred payment method of both churned and non-churned high value customers are evaluated.

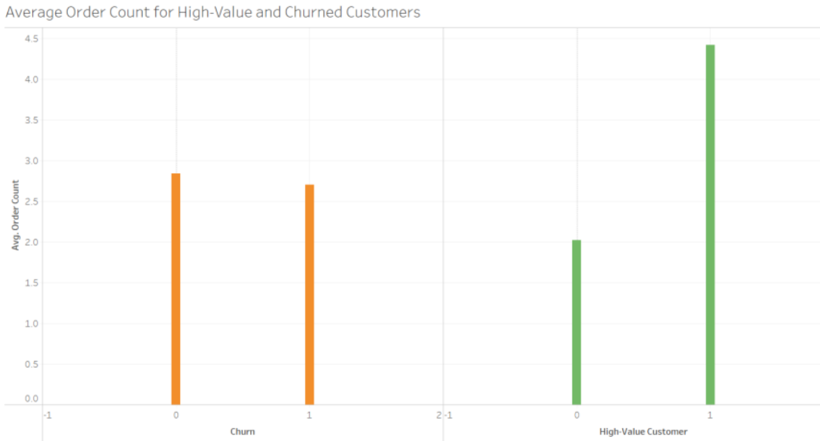




The use of debit card and credit card is prominent in both the customer segments followed by a ewallet. Nonetheless, the use of cards is more prevalent among the non-churned customers indicating that high-value customers retaining in the company may be more comfortable with these conventional payment methods. This might also be the result of the point collection and rewards system that the debit and credit cards usage provide.

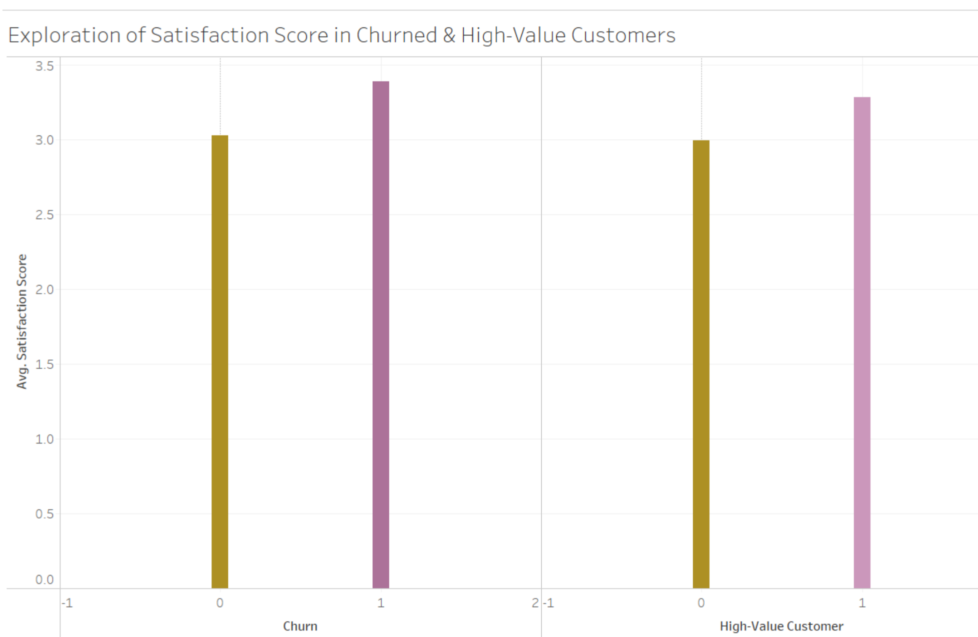
However, more of the churned customers prefer ewallet and might churn as a result of lacking digital incentives. Such customers might rely on providing targeted retention efforts providing exclusive offers or loyalty rewards to increase their engagement.

Average Order Count



The above bar chart compares average order count among different customer segments. High value customers have higher average aligning with their status as frequent buyers. However, the slight lower average of the order count of churned customers could indicate that a drop in frequency of purchase could be an early sign of churn among customers and penetrating appropriate retention strategies at this stage might help to reduce the churn.

Satisfaction Score

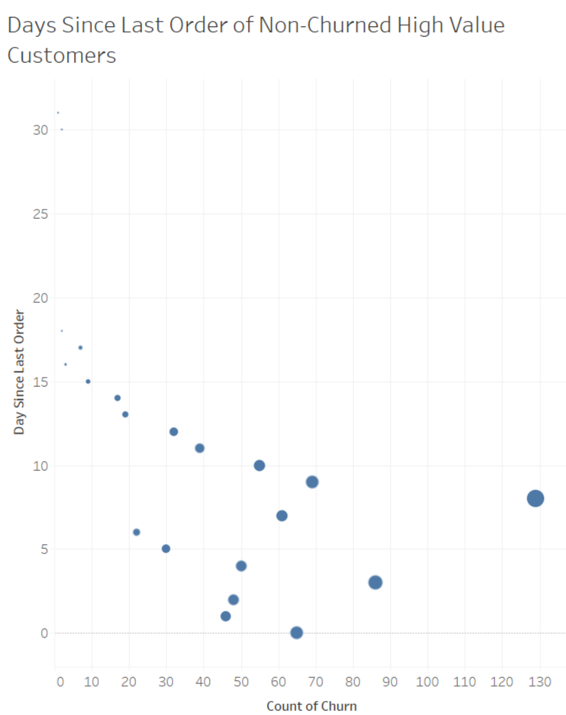


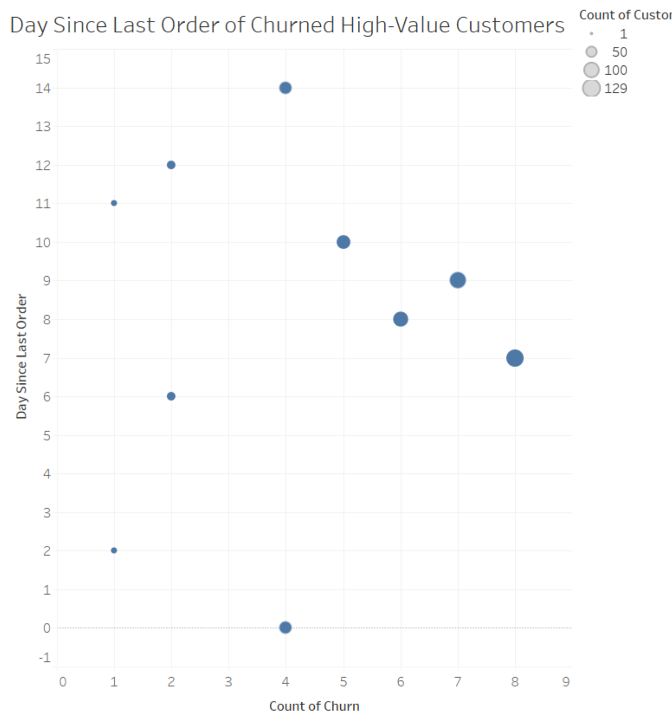
Here, the satisfaction score of the customers in different segments are evaluated. It is evident that the findings from this chart is counterinitiative, as the customers who churned have higher satisfaction score, indicating that satisfaction score might not be the perfect factor to evaluate whether a customer will churn. In this case, we might opt to other factors for the assessment of customer churn rather than relying on the satisfaction score.

On the other hand, high-value customers have slightly more average satisfaction score compared to the other. Nonetheless, the insignificant difference also indicate that satisfaction score solely cannot explain high-value status of the customers.

Days Since Last Order | Churned vs non-churned customers

Here, the days since last order, which might not solely indicate the duration since last engagement with the ABC E-comm, of churned and non-churned customers are evaluated side by side.





Here, the size of the bubbles indicate the number of customers. In the left chart, we can see that the concentration of non-churned high-value customers is more towards the bottom of the chart i.e., below 10 days including larger bubble size i.e., larger number of customers. This suggests that recent engagement with the company is a strong retention indicator among the high-value customers. Though there are few customers with over 20 days of engagement, these might be at lower risk of churning due to their purchasing behavior (such as bulk purchase, seasonal purchase, etc.).

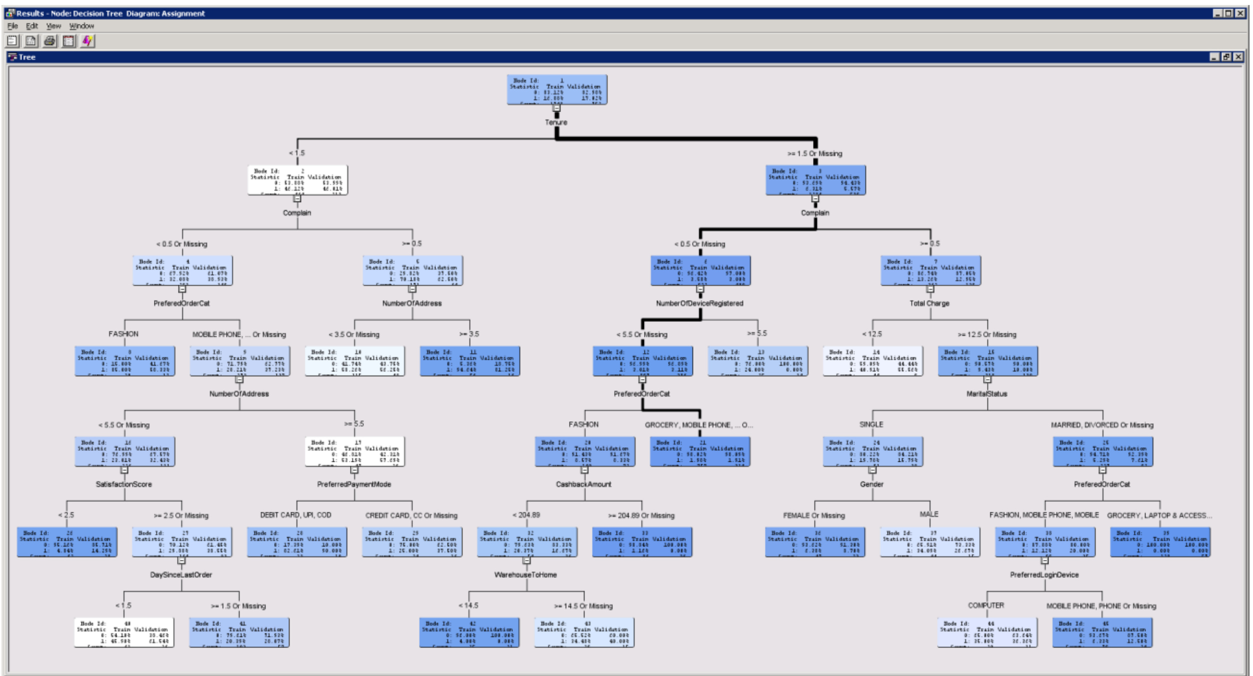
However, the chart on the right indicating the days since last order of high value churned customers illustrates a higher gap with many customers having a gap of 5 days or more since last order before churning. This also indicates that customers who haven’t ordered recently (within last 5+ days) are at the higher risk of churning. ABC can make retention efforts to nudge and re-engage these customers with personalized offers and reminders.

Task 2: Developing and Evaluating Models to Predict Propensity to Churn

Decision Tree Model

As the first predictive model to evaluate the customer's propensity to churn, decision tree model is created using necessary steps in SAS EN.

Here's the result of the decision tree model.



Most Important Variables

Variable Name	Label	Number of Splitting Rules	Importance	Validation Importance	Validation to Training Importance
Tenure		1	1.0000	1.0000	1.0000
Complain		2	0.5789	0.3134	0.5413
NumberOfAddress		2	0.3974	0.2998	0.7544
PreferedOrderCat		3	0.3687	0.1942	0.5268
PreferredPaymentMode		1	0.2686	0.2406	0.8958
Total_Charge	Total Charge	1	0.2662	0.2494	0.9366
SatisfactionScore		1	0.2285	0.2321	1.0159
DaySinceLastOrder		1	0.2149	0.2574	1.1981
Gender		1	0.1797	0.1019	0.5668
MaritalStatus		1	0.1589	0.0545	0.3427
PreferredLoginDevice		1	0.1558	0.1489	0.9558
WarehouseToHome		1	0.1520	0.2344	1.5423
CashbackAmount		1	0.1505	0.1337	0.8886
NumberOfDeviceRegistered		1	0.1408	0.0000	0.0000

On evaluation of the decision tree, we can see that the tree starts with splitting on Tenure and Complain, indicating that these characteristics are the most important ones in predicting customer churn.

So, based on the position in the tree and frequency of use, Tenure is the most important variable with a strong relation to churn of customers.

Afterwards, also at a high level, Complain is another important variable, indicating complaints of the customers have high correlation with churn.

Besides these, the other significant variables to predict churn are Number of Device Registered, Preferred Order Category, and Total Charge.

Customer Groups with High Churn Prediction

To identify the customer groups that are most likely to churn, the average for churn is evaluated. The average is closer to 1, more likely the customers will churn.

Here, customer groups that are most likely to churn are:

i. Customers with low tenure (Tenure <5) who have made complaints (Complain >= 1.5).
ii. Customers with low tenure, who have not complained but prefer different categories such as Fashion or Mobile Phone in their order.

These groups represent relatively new customers or the ones with specific preferences, making them more susceptible to churn.

Group of Customers with Difficulty in Predicting Churn

The customers with churn probability close to 50/50 are the most challenging groups to predict the churn. Here, their churn and non-churn rates are similar, and it is often found in the nodes with average churn value of around 0.5.

For example, in this case, the following nodes represent the most difficult groups to predict churn.

Tenure >= 5, Complain < 0.5 and Total Charge < 12.5.

Tenure < 1.5, Complain < 0.5 or Missing, PreferedOrderCat – MOBILE PHONE, ... Or Missing, and Number of Address >= 5.5

These groups may indicate mixed engagement, raising difficulties to predict the churn behavior accurately.

Decision Tree vs Random Guess

A random guess might have a baseline accuracy approximately equal to the overall churn rate in the dataset.

However, using decision tree in SAS for prediction of churn rate could be more effective as it can leverage the structure of data, resulting in lower misclassification rate and higher accuracy for both churned and non-churned customers. By emphasizing on the variables with higher importance such as Tenure and Complain in this case, the decision tree can make data driven prediction that cannot be achieved with random guessing.

Moreover, the decision tree model captures real-world patterns to deliver actionable insights, marking itself as a valuable tool for predictive analytics.

Additional Predictive Models

Regression Model

The regression model is generated with following properties.

Model Selection	
Selection Model	Backward
Selection Criterion	Validation Error
Use Selection Defaults	Yes

On evaluating the results of the regression model, we can find the following :

Results - Node: Regression Diagram: Assignment

File Edit View Window

Output

Type 3 Analysis of Effects

Effect	DF	Wald Chi-Square	Pr > ChiSq
CashbackAmount	1	16.8869	<.0001
CityTier	1	8.4060	0.0037
Complain	1	107.8673	<.0001
CouponUsed	1	1.8105	0.1784
DaySinceLastOrder	1	10.6509	0.0011
Gender	1	7.1930	0.0073
High_Value_Indicator	1	0.1448	0.7036
HourSpendOnApp	1	0.5355	0.4643
MaritalStatus	2	15.8945	0.0004
NumberOfAddress	1	56.3160	<.0001
NumberOfDeviceRegistered	1	34.3773	<.0001
OrderAmountHikeFromlastYear	1	3.5023	0.0613
OrderCount	1	1.8010	0.1796
PreferedOrderCat	5	64.7773	<.0001
PreferredLoginDevice	2	2.1944	0.3338
PreferredPaymentMode	6	10.7320	0.0970
SatisfactionScore	1	17.0972	<.0001
Tenure	1	130.0537	<.0001
Total_Charge	0	0.0000	.
WarehouseToHome	1	7.9362	0.0048

Most Important Variables in Regression

On evaluating the output of Analysis of Effects in the above results, the variables with highest Wald Chi-Square values indicates higher significance. They are

Complain: Chi-Square = 107.8675, $p < 0.0001$

NumberOfDeviceRegistered: Chi-Square = 34.3773, $p < 0.0001$

MaritalStatus (Married): Chi-Square = 56.3182, $p < 0.0001$

CashbackAmount: Chi-Square = 16.8869, $p < 0.0001$

CityTier: Chi-Square = 8.4600, $p = 0.0037$

These variables have high statistical significance in predicting churn in the regression model.

Comparison with the Decision Tree Model

The decision tree model and the regression model has some common variables in both the models.

Complain and CityTier were significant variables in both the models, suggesting the consistency of the variables, also indicating the robustness of its prediction of churn across different models.

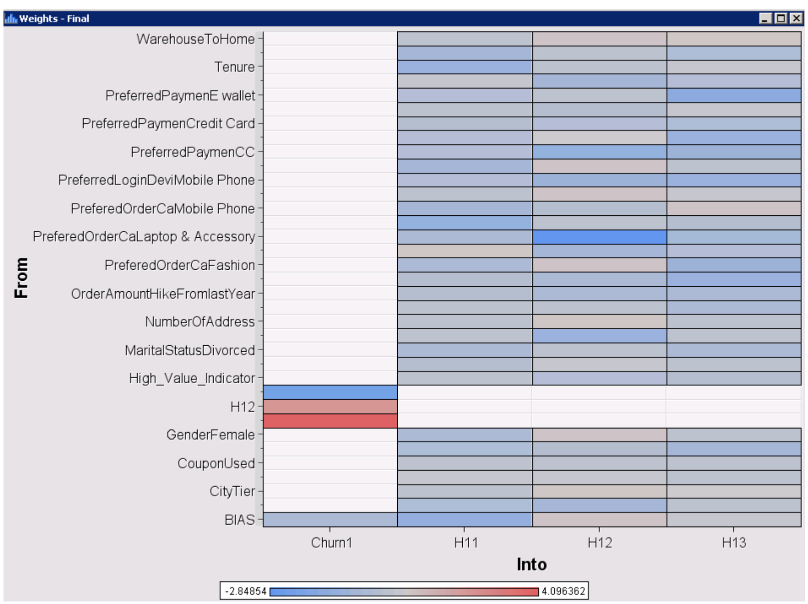
NumberofDeviceRegistered has also appeared as one of the significant predictors in both the models.

Nonetheless, other variables such as CashbackAmount and MaritalStatus are highlighted in the regression model, but these variables did not appear as primary splits in the decision tree model. This difference is normal as the decision tree focuses on hierarchical splits while regression captures the additive effects of variables rather conveniently.

Neural Network

Neural Network is yet another effective predictive model which is exercised in this case as well. Here, a number of layers of networks are created to make predictions by reading the data and understanding the hidden patterns.

Here's the final weight of the neural networks generated in SAS EN.



Here, the color intensity and weight values show the strength and direction of influence of the variables (predictors) on churn through the hidden layers i.e., H11, H12, and H13.

Based on the color intensity, the following are the most influential variables to predict churn.

Gender (Female) - It is the most strongest positive predictor on the churn (indicated by the dark red).

CityTier – It also has a significant influence on the churn prediction that is indicated by a darker shade.

Tenure – With a darker shade in the first hidden layer, we can see that it is also significant to the model.

PreferredOrderCat (Mobile phone and laptop & accessory categories) – This variable also has a notable influence aligning with the previous models.

PreferredPaymentMode (Ewallet) – The H13 layer shows a significant highlight for this category as well.

Comparison with Decision Tree

Tenure, Gender (Female), and PreferredOrderCat are prominent in both Neural Network model as well as decision tree model.

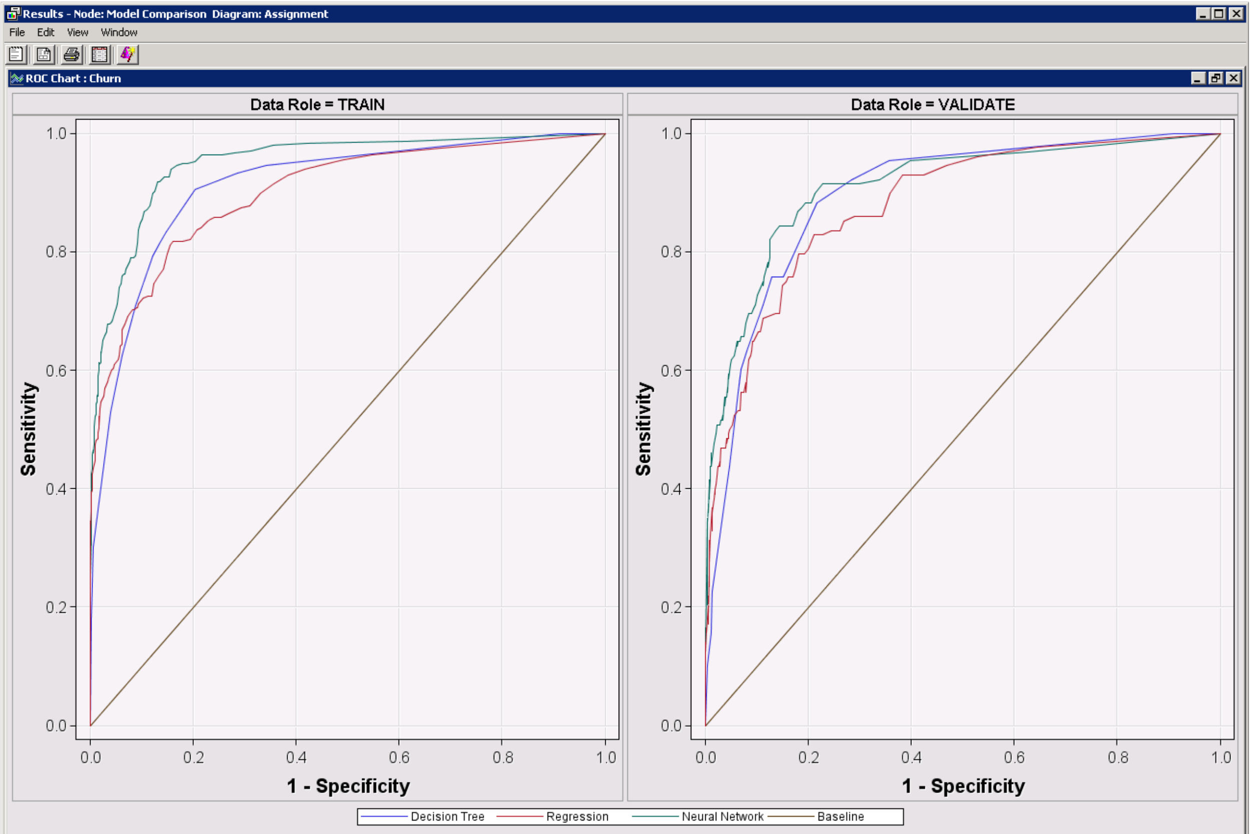
PreferredPaymentMode (E wallet) is also consistent in the previous models indicating the strong association with churn.

Predictive Performance Analysis

The 3 different predictive models are then further analyzed and examined based on different metrics to rank based on their suitability.

ROC (AUC)

The Receiver Operating Characteristic (ROC) Curve is a plot of sensitivity that provides area under the curve (AUC) value. It helps to identify the ability of model to segregate between churned and non-churned customers.



Neural Network

Indicated by the green curve, it has the highest AUC in training as well as validation datasets, indicating that it is the best amongst all to segregate the churned and non-churned dataset.

Regression

Similar to Neural Network, it also has a high AUC with significantly higher than 0.5 ROC with only little area of the chart uncovered by the curve in both training as well as validation sets.

Decision Tree

It has the lowest AUC among the three, indicating it is the least accurate in classifying churned customers. Nonetheless, the AUC of this model is also significant.

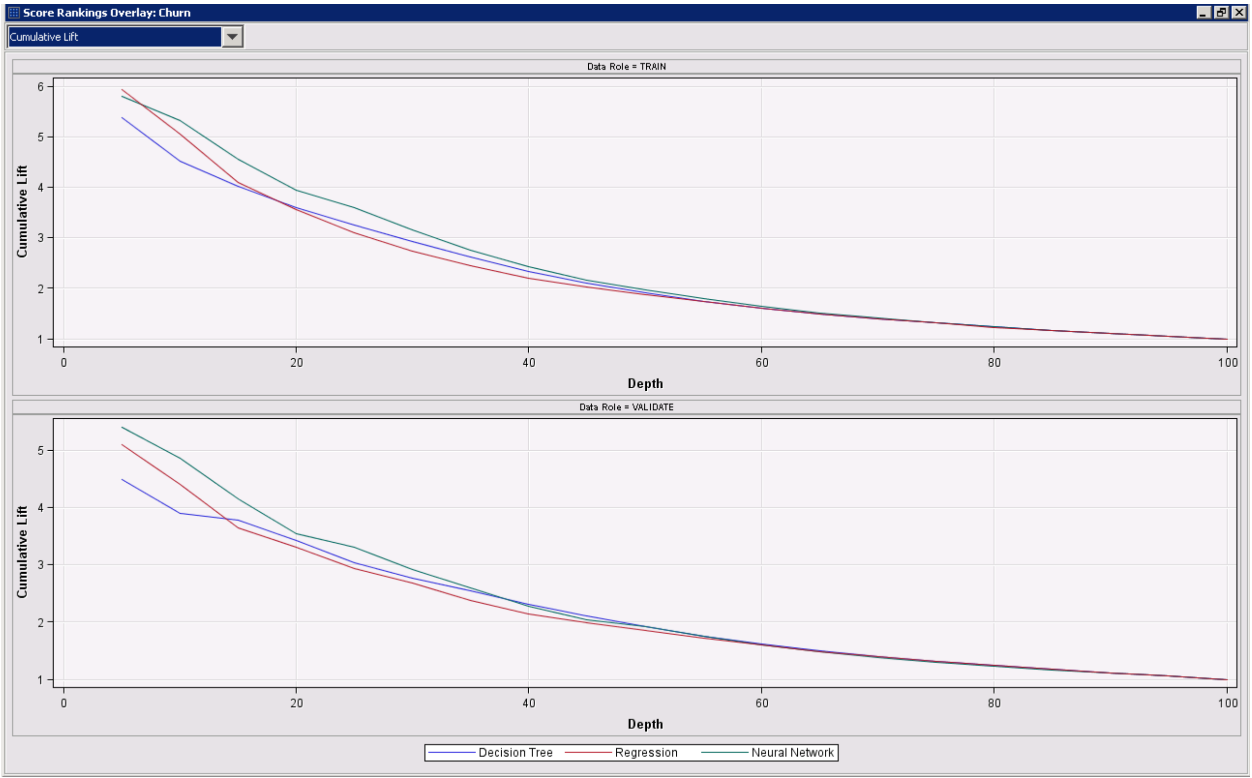
Lift

Lift chart shows the improvement of a model above a random guess. Higher lift indicates the ability of a model to capture true positives (churned customers) leading to better performance.

Here, the green curve, i.e. neural networks has the highest lift in both training as well as validation datasets indicating that its effectiveness to identify churned customers.

Slightly lower than neural network curve, the regression model (red curve) has slightly low but similar ability to capture the churned customers.

Lastly, Decision tree has the lowest lift of all, showing its least effectiveness to distinguish between churned and non-churned customers.



Misclassification Rate

It represents the proportion of incorrect predictions in the dataset performed by each models.

Fit Statistics

Selected Model	Predecessor Node	Model Node	Model Description	Target Variable	Target Label	Selection Criterion: Valid: Average Squared Error	Train: Misclassification Rate	Valid: Misclassification Rate
Y	Neural	Neural	Neural Net...	Churn		0.077278	0.080092	0.107713
	Reg	Reg	Regression	Churn		0.090175	0.096682	0.12367
	Tree	Tree	Decision Tr...	Churn		0.091095	0.1127	0.135638

Neural Network

It has the lowest misclassification rate on both training (0.08) and validation (0.107) datasets, suggesting it to be the most accurate model.

Regression

With a slightly higher rate than the Neural Network on both training (0.0967) and validation (0.1237) datasets, it is still reasonably accurate model to predict the customer churn.

Decision Tree

With the highest misclassification rates on both training (0.0911) and validation (0.1356) datasets, decision tree can be considered the model that is likely to make more incorrect predictions than the other models.

Average Squared Error

The Average Squared Error measures the difference between predicted and actual values of the customer churn, lower the ASE better the prediction.

Fit Statistics

Selected Model	Predecessor Node	Model Node	Model Description	Target Variable	Selection Criterion: Valid: Average Squared Error	Train: Average Squared Error	Valid: Average Squared Error
Y	Neural	Neural	Neural Net...	Churn	0.077278	0.059689	0.077278
	Reg	Reg	Regression	Churn	0.090175	0.07658	0.090175
	Tree	Tree	Decision Tr...	Churn	0.091095	0.077922	0.091095

Neural Network

With the lowest ASE on the validation set (0.0777), it has the highest accuracy in terms of error.

Regression

Higher ASE on validation (0.13527) compared to the Neural Network.

Decision Tree

Highest ASE on validation (0.13564), indicating the lowest accuracy.

Model Ranking

Based on the metrics discussed above, the models can be ranked as follows:

Neural Network

Neural Network performs the best in all metrics (AUC, Lift, Misclassification Rate, and ASE). It has the highest discriminatory power, lowest error, and best overall accuracy.

Regression

This model performs slightly below the Neural Network, with good accuracy but higher error rates and slightly lower lift.

Decision Tree

Among all the selected models, this model ranks the lowest, with the highest misclassification rate, lowest lift, and highest ASE. However, it remains useful for interpretability.

Best Model: Decision Tree Model

Though Neural Network has the best numbers amongst the three models, the results of decision tree is not bad either, but only is slightly less over them. Nonetheless, despite the numbers, other factors such as interpretability of the model to businesses and others also count as crucial for the best model. Let's have a look at some more reasons.

High Interpretability

To help the ecomm evaluate its customers and predict churn, decision tree can be the best suitable as it helps to suffice the primary issue of the company i.e., to find patterns and then predict churn. Stakeholders can easily identify the customer attributes that lead to higher churn risk, which is invaluable to plan targeted interventions.

Actionable Insights

The Decision Tree model highlights specific segments of customers and the factors that most impact their likelihood of churning, such as PreferredPaymentMode, Tenure, Complain, and Order Category. These insights can directly guide marketing, customer service, and retention teams on where to focus their efforts (e.g., prioritizing retention strategies for customers with short tenure who have lodged complaints).

Sufficient Predictive Power

Although the Decision Tree does not achieve the same accuracy as the Neural Network, it still performs reasonably well, especially if paired with additional business rules or used in conjunction with other simpler models (like Regression). The slight trade-off might be outweighed by the clear understanding the model provides.

Interpreting Decision Tree Model for Churn Prediction

Identifying Key Factors of Churn

Decision Tree helps to identify the key factors or predictors of churn. For instance, in this case, the top split in the tree is based on Tenure, indicating that Tenure is the most important predictor. This shows that customers with shorter tenure are at the higher risk of churning.

Similarly subsequent splits may include variables such as Complain or PreferredPaymentMode, further segmenting high-risk customers.

Actionable Segments

Each branch in the Decision Tree represents a unique customer segment with particular characteristics. For example, one branch might identify "customers with Tenure < 6 months and high complaint frequency" as a high-risk group. Another segment might include "long-tenure customers with no complaints but low engagement." With such clear and rule based segments, the model helps the organization to understand the types of customers that are likely to churn.

Visual Interpretation and Communication

The visual structure eases the understanding of how each factor contributes to churn. The splits and branches also provide a straight explanation of why certain customers are at the risk of churn without any statistical interpretation. This enables cross-functional communication as well.

Interpreting Variable Interactions and Non-Linear Relationships

The decision tree captures interactions between variables that may not be clear in other models. For instance, it results the impact of low satisfaction on churn for customers who have stayed with the company for longer term while new customers might be more influenced by payment preferences.

Transparency for Targeted Interventions

Each leaf of the tree has a segment with specific churn probability which can be used by the companies to determine which customer should be targeted for marketing and retention campaigns. The leaf with high churn probability can be prioritized for immediate intervention.

In summary, decision tree model is interpretable, actionable, and visually accessible, making it an ideal tool for an organization looking to understand and reduce customer churn through targeted, data-driven interventions.

Task 3: Campaign Recommendations Based on Insights Drawn

Lets work on building the campaigns for ABC Ecommerce based on the insights from Task 1 and Task 2.

Summary of Key Insights

From the above tasks, the following were identified:

High Risk Customer Segments

Customers with low tenure, high complaint frequency, ewallet users, and those who prefer specific order categories such as Mobile Phone or Laptop and Accessory tend to have higher churn rates. Moreover, customers with low satisfaction scores are also more likely to churn.

Influential Variables

On evaluation of the tasks, the key predictors of churn included Tenure, Complain, PreferredPaymentMode and Order Category, defining the profiles of customers that are more likely to churn providing specific characteristics to target.

Campaign Recommendations Based on Insights

Few Campaign recommendations for customer retention are:

Welcome and Onboarding Program for New Customers

Objective: Reduce churn among customers with shorter tenure by increasing their engagement and satisfaction early in their journey.

Campaign Tactics:

Onboarding Assistance: Send a series of personalized emails or app notifications explaining the features, benefits, and best practices to help new customers get the most value from the service.

Early Incentives: Offer discounts on first transactions to encourage regular use and build habits.

Follow-Up Check-Ins: After 30 and 60 days, reach out to new customers to collect feedback and ensure they're satisfied with special offers for those at risk of churn.

Studies suggest that engaging onboarding process can improve customer retention by a significant amount (Cooper & Kaplan, 2021).

Complaint Resolution and Satisfaction Improvement Program

Objective: Address customers' complaints promptly and improve satisfaction scores to reduce churn among at-risk customers.

Campaign Tactics

Proactive Complaint Resolution: Implement a priority customer service team focused on resolving complaints for high-risk segments.

Customer Feedback Loops: Introduce automated surveys after complaint resolution to gauge customer satisfaction on their complaint handling. Incentivize feedback to ensure valuable insights for improvement.

Personalized Apology and Incentive: For customers with unresolved complaints, a tailored engagement with incentives (e.g., discount or free shipping) to encourage another purchase.

Maxham & Netemeyer (2002) suggested that addressing customer complaints quickly and effectively can significantly plummet the customer churn rates.

Loyalty Program Tailored to Payment Preferences and Product Categories

Objective: Retain customers who are more likely to churn because of payment preferences (e.g., E Wallet users) and specific product categories (e.g., Mobile Phones, Laptops).

Campaign Tactics:

Customized Loyalty Rewards: For E Wallet users and high-risk order categories, the company can offer loyalty points or cashback to retain engagement within these categories.

Exclusive Discounts for Preferred Categories: Provide exclusive discounts or early access to sales for high-churn categories like Mobile Phones and Laptops to reduce churn on this category.

Payment Mode Incentives: Offer payment-specific incentives, such as bonus rewards for purchases made through E Wallet or preferred payment methods, to increase engagement.

Anderson & Sullivan (1993) suggests that addressing the specific needs of dissatisfied customers can reduce the customers churn by 20-40%.

In summary, these campaigns are designed to high-risk customer segments identified in the previous tasks, using different ideas and plans that are backed by research and best practices.

References

Anderson, E. W., & Sullivan, M. W. (1993). The antecedents and consequences of customer satisfaction for firms. *Marketing Science, 12*(2), 125-143.

Börühan, G. (2022). Churn Customer Management in Retail Industry: A Case Study. *İzmir İktisat Dergisi, 37*(4), 1094-1118.

Buttle, F., & Maklan, S. (2019). *Customer relationship management: Concepts and technologies*. Routledge.

Kumar, V., & Venkatesan, R. (2021). Transformation of metrics and analytics in retailing: The way forward. *Journal of Retailing, 97*(4), 496-506.

Matuszelański, K., & Kopczewska, K. (2022). Customer churn in retail e-commerce business: Spatial and machine learning approach. *Journal of Theoretical and Applied Electronic Commerce Research, 17*(1), 165-198.

Maxham III, J. G., & Netemeyer, R. G. (2002). A longitudinal study of complaining customers' evaluations of multiple service failures and recovery efforts. *Journal of Marketing, 66*(4), 57-71.

Morgan, R. M. (1994). The commitment-trust theory of relationship marketing. *Journal of Marketing*.

Patil, A. P., Deepshika, M. P., Mittal, S., Shetty, S., Hiremath, S. S., & Patil, Y. E. (2017, August). Customer churn prediction for retail business. In *2017 International Conference on Energy, Communication, Data Analytics and Soft Computing (ICECDS)* (pp. 845-851). IEEE.

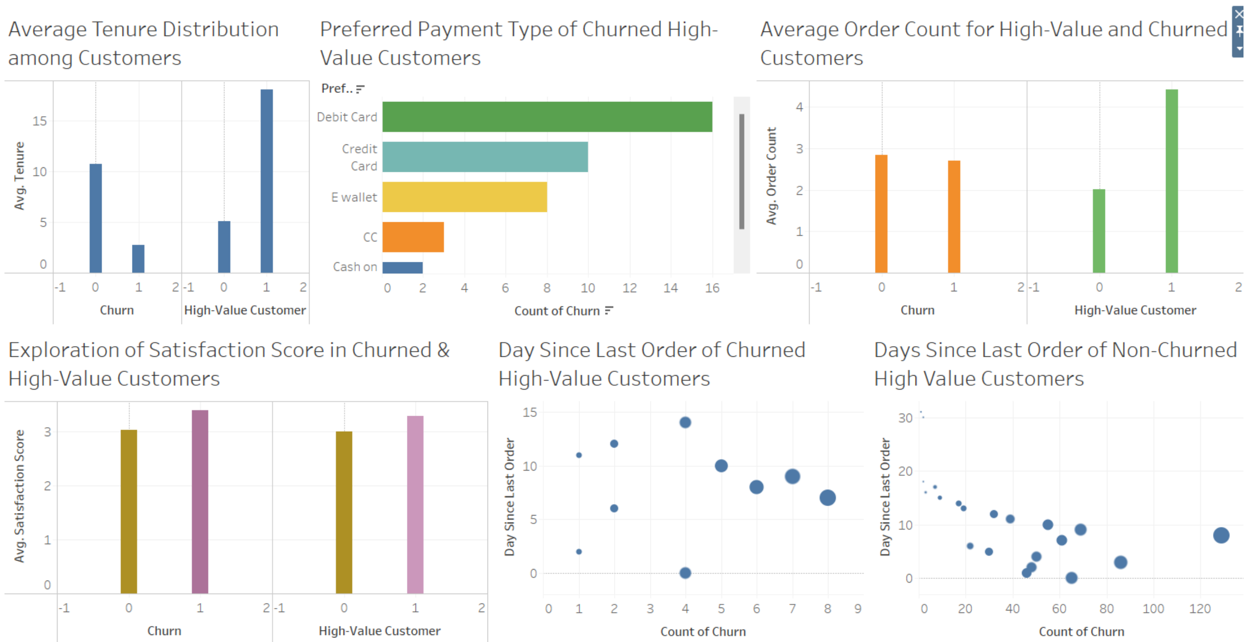
Röser, M. (2024). Customer Relationship Management in New Business. *Business Ethics: The Competitive Advantage of Trust and Reputation*, 65.

Tax, S. S., Brown, S. W., & Chandrashekaran, M. (1998). Customer evaluations of service complaint experiences: Implications for relationship marketing. *Journal of Marketing, 62*(2), 60-76.

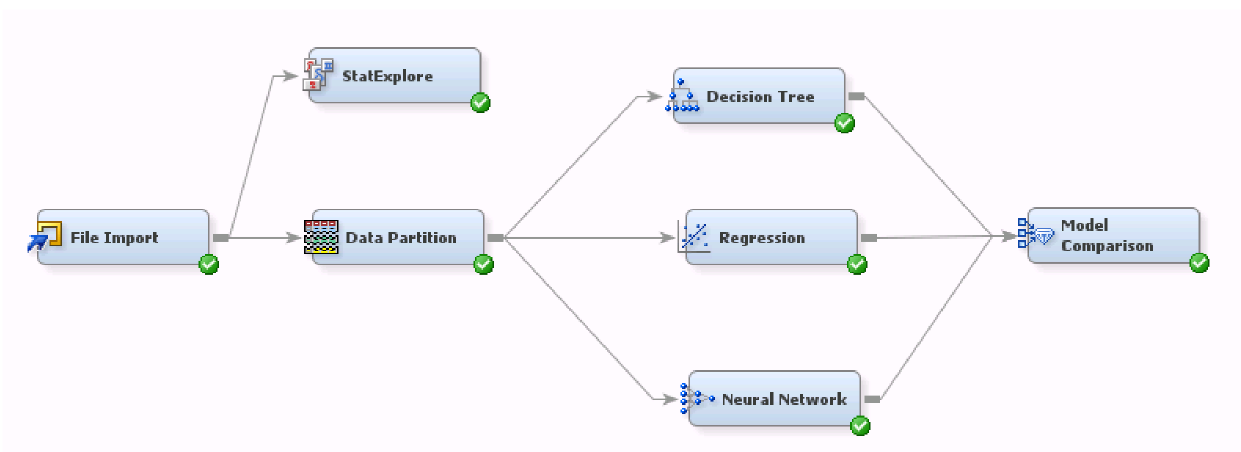
Zhang, Y. (2024). The evolution of loyalty programs in the digital age: A review of consumer engagement and retention strategies. *Business, Marketing, and Finance Open, 1*(2), 1-12.

Appendix

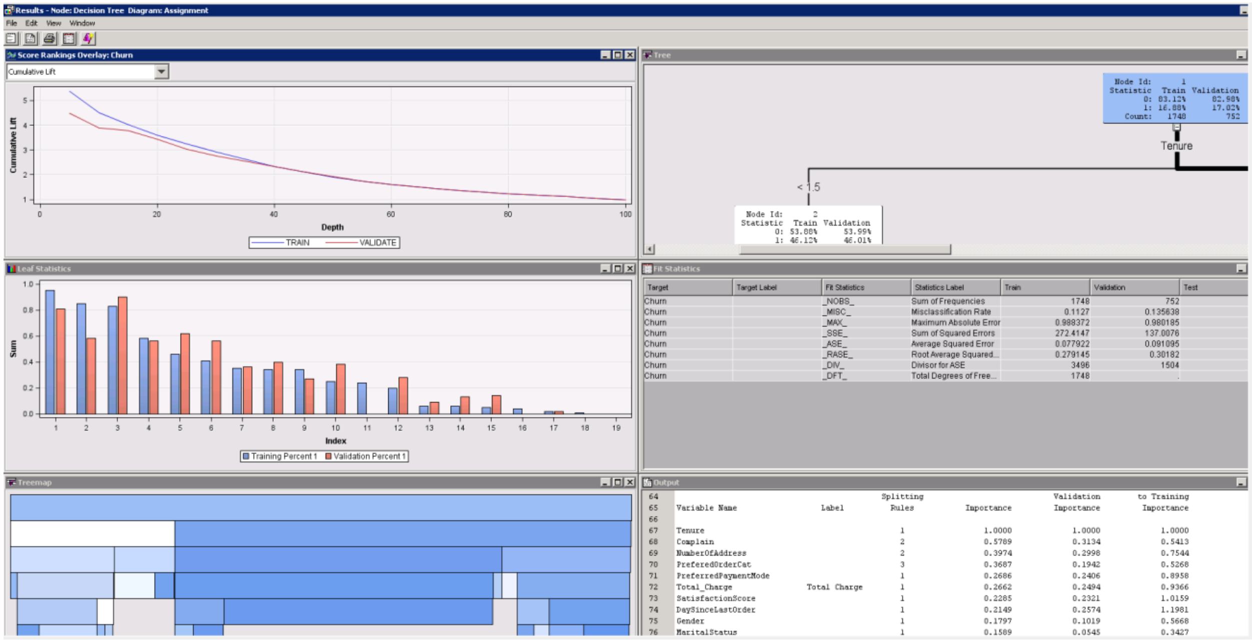
Dashboard of Tableau comprising of the visualizations created



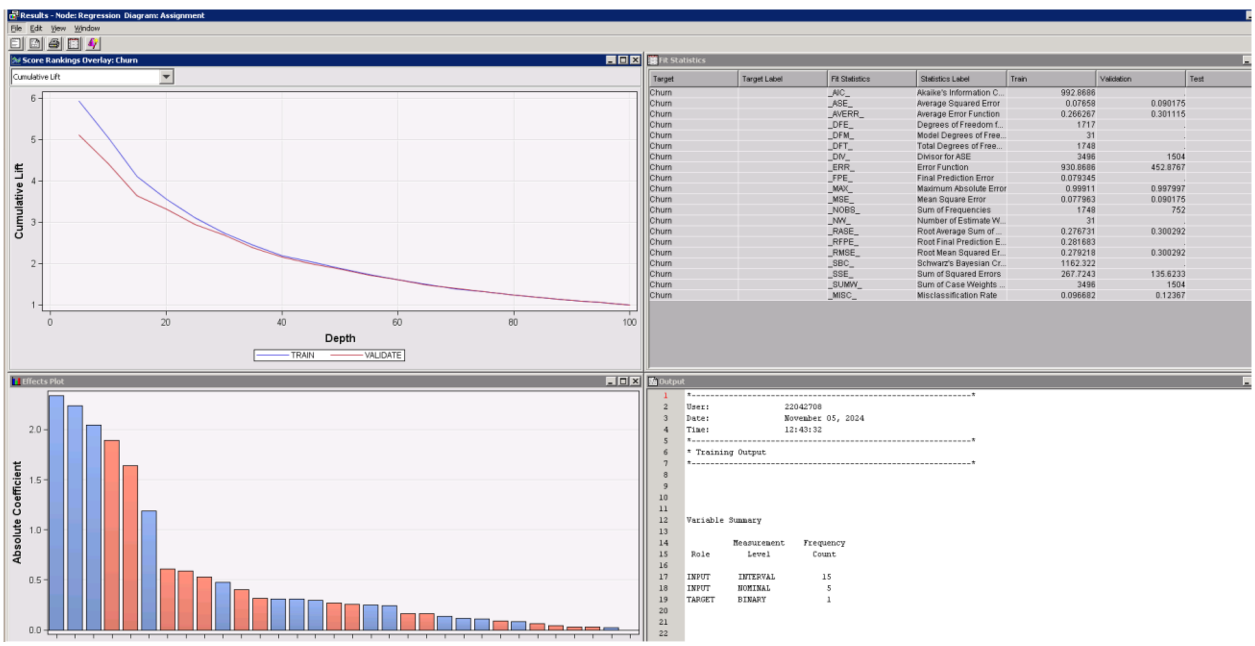
Flow of Nodes in SAS



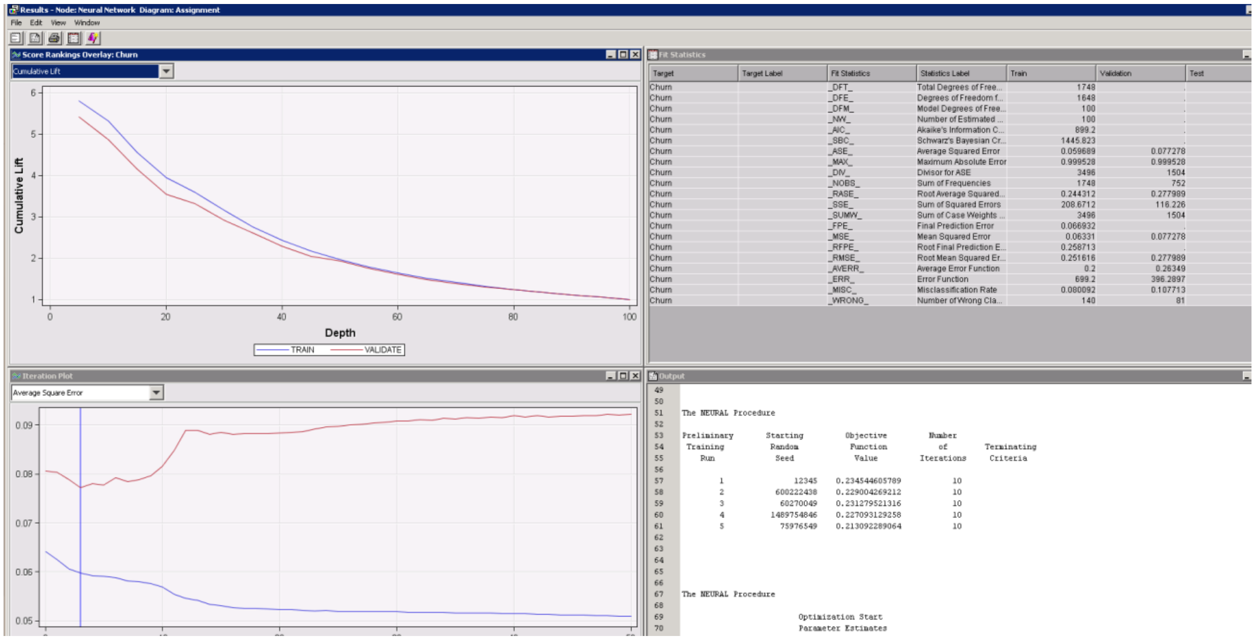
Results of Decision Tree Node



Results of Regression Node



Results of Neural Network Node



Results of Model Comparison node