TABLE OF CONTENTS

1.0 Introduction

Customers are one of a business's most valuable assets, and their satisfaction is critical to the company's overall profitability and ability to compete successfully in the market. Customers have access to a wide variety of products and service providers, which is a positive development despite the intense competition in the industry. Surveys indicate that it is often more expensive to recruit new customers than it is to retain relationships with ones that already exist. Maintaining a long-term relationship with existing customers will increase a company's revenue. Hence, in order to keep their dominant position in the market, it is now essential for companies to figure out how to make the most of the consumers they already have and how to keep from losing any of their current customers.

2.0 Business Understanding

In this assignment, the business that will be studied is a leading online e-commerce company, a non-store online retailer. The majority of the products the company offers include apparel, foodstuffs, computers and accessories, and mobile phones. Since their inception, the company has seen a steady decline in the number of customers they have. Therefore, for the company to keep their position as the leading online e-commerce company and maximise its profits, the company wishes to research the causes for the consumers to leave and take precautionary measures to prevent additional existing customers from leaving. As a result, the company's goal is to strengthen its customers' loyalty and build a predictive model to identify consumers who are likely to churn.

3.0 Aims and Objectives

The primary aim of this project is to create a predictive machine learning model by utilising supervised classification algorithms and recommend appropriate marketing techniques to the online e-commerce company to reduce the number of customer churns. The objectives of this project are:

i. To construct a predictive machine learning model to classify the consumers who are going to churn by analysing the purchase behaviour of such customers.

ii. To create individualised marketing strategies to retain customers who are more inclined to churn.

4.0 Metadata of the Dataset

The dataset used for this assignment was obtained from Kaggle, which is available at https://www.kaggle.com/datasets/ankitverma2010/ecommerce-customer-churn-analysis-and-prediction. The dataset consists of 20 attributes and 5630 instances, which includes the Customer ID, and the independent variable "Churn", where 1 indicates that the customer had churned, and 0 indicates that the customer is still active. The metadata of the dataset is shown in Table 1.

Table 1. The metadata of the e-commerce dataset.

Variable Name	Description	Data Type	Length	Sample Data
Customer ID	Unique customer ID	Numeric	8	50001, 50002, 50003...
Churn	Churn flag	Numeric	8	1; 0
Tenure	Tenure of customer in the organisation	Numeric	8	4, 0, 13...
Preferred Login Device	Preferred login device of customer	Character	12	Mobile Phone; Computer
City Tier	City tier of customer	Numeric	8	1; 2; 3
Warehouse To Home	Distance in between warehouse to home of customer	Numeric	8	6, 8, 30...
Preferred Payment Mode	Preferred payment method of customer	Character	16	Cash on Delivery; Credit Card; Debit Card; E-wallet; UPI
Gender	Gender of customer	Character	6	Male; Female
Hour Spend On App	Number of hours spent on mobile app or website	Numeric	8	1, 2, 3...

Number Of Device Registered	Total number of devices registered per customer	Numeric	8	3, 4, 5...
Preferred Order Category	Preferred order category of customer in the last month	Character	18	Mobile Phone; Laptop & Accessory; Fashion; Grocery; Others
Satisfaction Score	Service satisfaction score given by customer	Numeric	8	2, 3, 4...
Marital Status	Marital status of customer	Character	8	Single; Married; Divorced
Number Of Address	Total number of addresses added by customer	Numeric	8	9, 7, 6...
Complain	Any complain raised by customer in the last month	Numeric	8	1; 0
Order Amount Hike From Last Year	Percentage increases in order compared to last year	Numeric	8	11, 15, 14...
Coupon Used	Total number of coupons used in the last month	Numeric	8	1, 0, 4...
Order Count	Total number of orders placed in the last month	Numeric	8	1, 6, 15...
Day Since Last Order	Number of days since last order by customer	Numeric	8	5, 3, 7...
Cashback Amount	Average cashback received by customer in the last month	Numeric	8	160, 121, 120...

5.0 Exploratory Data Analysis

Importing the File into SAS Enterprise Miner

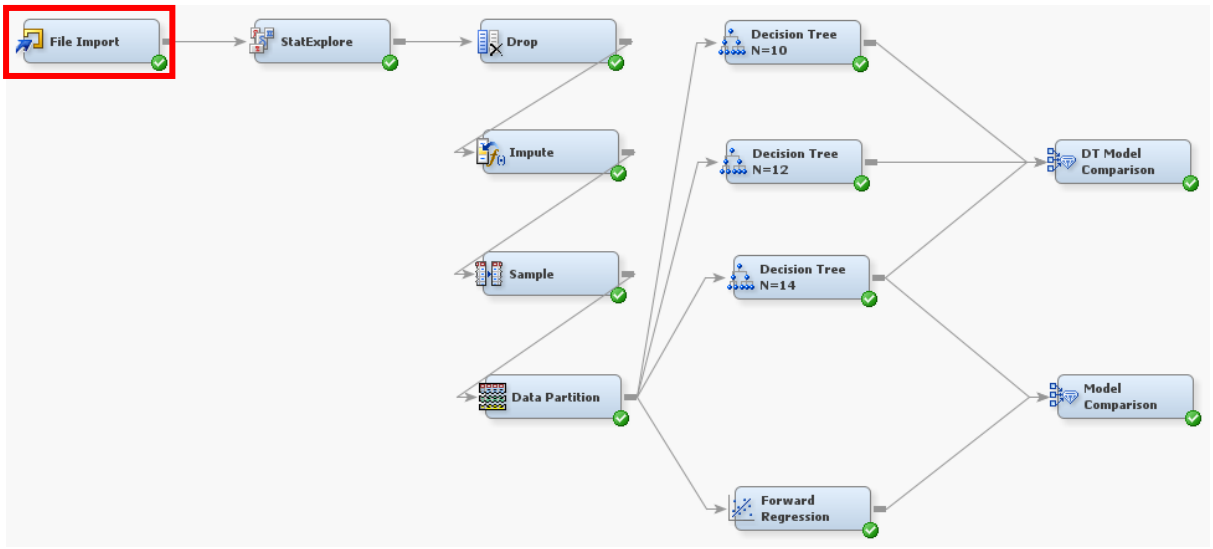

Figure 1. The e-commerce dataset was imported using File Import.

Name	Role	Level	Report	Order	Drop	Lower Limit	Upper Limit
CashbackAmoun	Input	Interval	No		No	.	.
Churn	Target	Binary	No		No	.	.
CityTier	Input	Interval	No		No	.	.
Complain	Input	Interval	No		No	.	.
CouponUsed	Input	Interval	No		No	.	.
CustomerID	ID	Interval	No		No	.	.
DaySinceLastOrd	Input	Interval	No		No	.	.
Gender	Input	Nominal	No		No	.	.
HourSpendOnAp	Input	Interval	No		No	.	.
MaritalStatus	Input	Nominal	No		No	.	.
NumberOfAddre	Input	Interval	No		No	.	.
NumberOfDevice	Input	Interval	No		No	.	.
OrderAmountHik	Input	Interval	No		No	.	.
OrderCount	Input	Interval	No		No	.	.
PreferedOrderC	Input	Nominal	No		No	.	.
PreferredLoginD	Input	Nominal	No		No	.	.
PreferredPayme	Input	Nominal	No		No	.	.
SatisfactionScor	Input	Interval	No		No	.	.
Tenure	Input	Interval	No		No	.	.
WarehouseToHo	Input	Interval	No		No	.	.

Figure 2. The independent variable's role was changed to "Target" and the level was changed to "Binary".

Firstly, the dataset was imported to SAS Enterprise Miner using the File Import function. In the variables tab, the role of the independent variable – "Churn", was changed to "Target". Its level was changed to binary, as we will be predicting whether the customer will churn according to the demographic data and the customer's purchasing behaviour.

6.0 Data Understanding

6.1 Initial Data Exploration

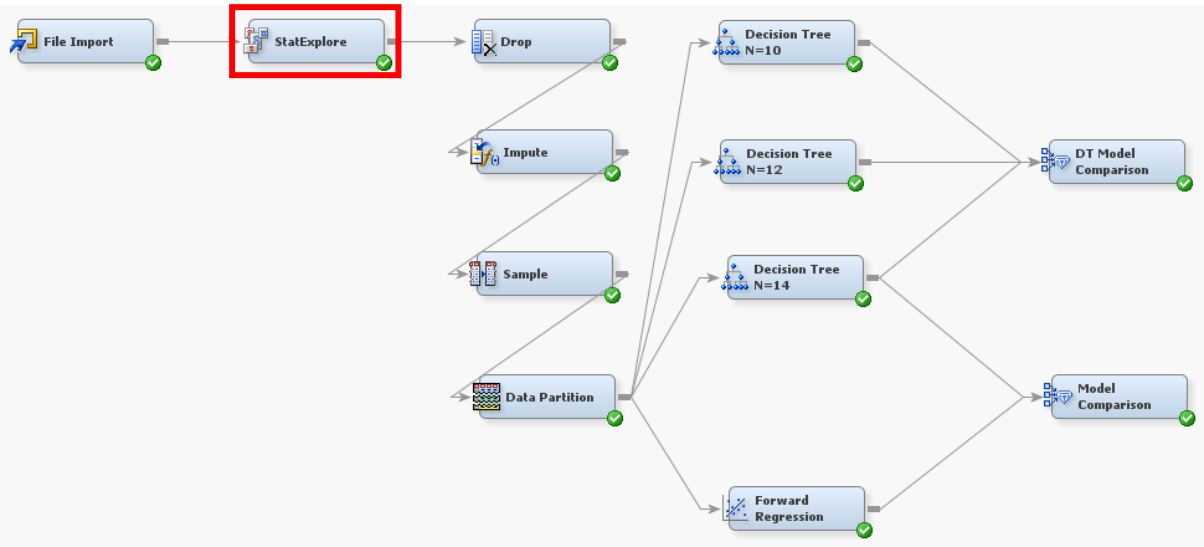

Figure 3. StatExplore was used for initial data exploration.

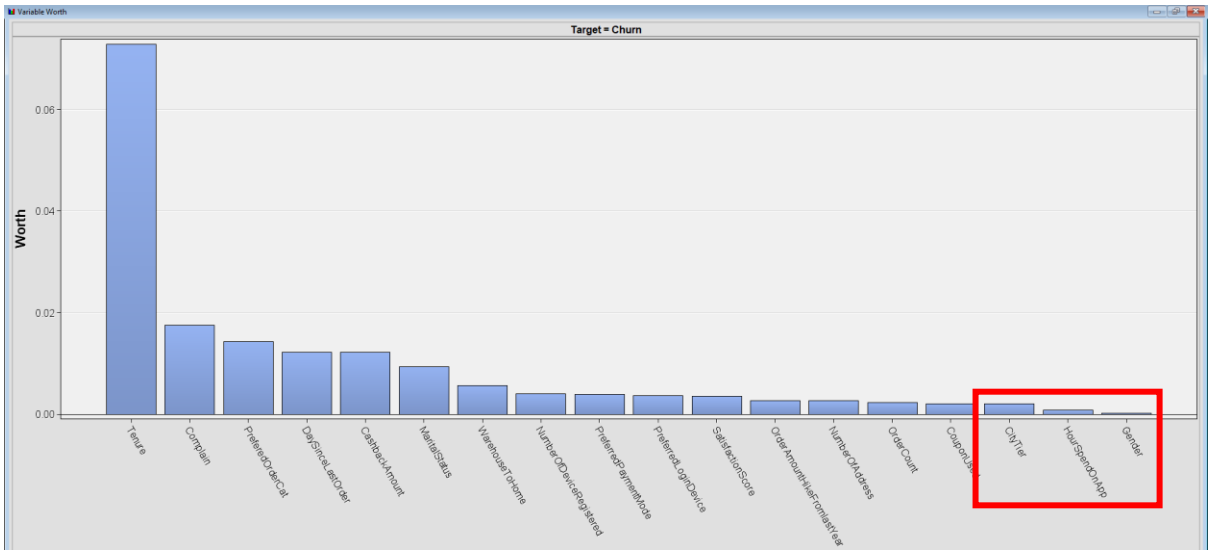

Figure 4. Results of StatExplore: CityTier, HourSpendOnApp, and Gender have the lowest variable worth.

Interval Variable Summary Statistics
(maximum 500 observations printed)

Data Role=TRAIN

Variable	Role	Mean	Standard Deviation	Non Missing	Missing	Minimum	Median	Maximum	Skewness	Kurtosis
CashbackAmount	INPUT	177.223	49.20704	5630	0	0	163.23	324.99	1.149846	0.974505
CityTier	INPUT	1.654707	0.915389	5630	0	1	1	3	0.735326	-1.40153
Complain	INPUT	0.284902	0.451408	5630	0	0	0	1	0.953347	-1.09152
CouponUsed	INPUT	1.751023	1.894621	5374	256	0	1	16	2.545653	9.132281
DaySinceLastOrder	INPUT	4.543491	3.654433	5323	307	0	3	46	1.191	4.023964
HourSpendOnApp	INPUT	2.931535	0.721926	5375	255	0	3	5	-0.02721	-0.66708
NumberOfAddress	INPUT	4.214032	2.583586	5630	0	1	3	22	1.088639	0.959229
NumberOfDeviceRegistered	INPUT	3.688988	1.023999	5630	0	1	4	6	-0.39697	0.582849
OrderAmountHikeFromlastYear	INPUT	15.70792	3.675485	5365	265	11	15	26	0.790785	-0.28038
OrderCount	INPUT	3.008004	2.93968	5372	258	1	2	16	2.196414	4.718466
SatisfactionScore	INPUT	3.066785	1.380194	5630	0	1	3	5	-0.14263	-1.12514
Tenure	INPUT	10.1899	8.557241	5366	264	0	9	61	0.736513	-0.00737
WarehouseToHome	INPUT	15.6399	8.531475	5379	251	5	14	127	1.619154	9.98693

Figure 5. Results of StatExplore: There are missing values in some of the variables.

After importing the dataset into SAS Enterprise Miner, StatExplore was used for initial data exploration to observe if there were any missing values in the dataset, or if the variables in the dataset were highly skewed. Besides, the worth of the variables was also determined, and it is shown that “CityTier”, “HourSpendOnApp”, and “Gender” have the least variable worth. Hence, these variables will be dropped in the following step. It is also noticed that there are missing values in “CouponUsed”, “DaysSinceLastOrder”, “HourSpendOnApp”, “OrderAmountHikeFromLastYear”, “OrderCount”, “Tenure”, and “WarehouseToHome”; thus, data imputation will be performed on these variables.

7.0 Data Preparation

7.1 Data Cleaning

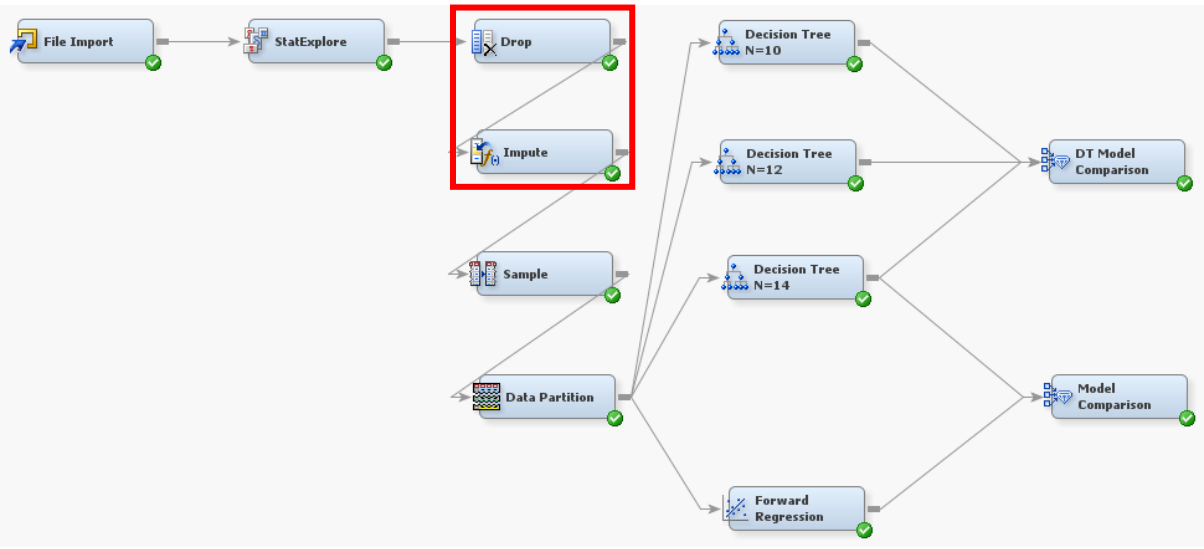

Figure 6. Drop and Impute methods were used in the data cleaning step.

Name	Drop	Role	Level
CashbackAmoun	**Default**	Input	Interval
Churn	**Default**	Target	Binary
CityTier	**Yes**	Input	Interval
Complain	**Default**	Input	Interval
CouponUsed	**Default**	Input	Interval
CustomerID	**Default**	ID	Interval
DaySinceLastOr	**Default**	Input	Interval
Gender	**Yes**	Input	Nominal
HourSpendOnAp	**Yes**	Input	Interval
MaritalStatus	**Default**	Input	Nominal
NumberOfAddre	**Default**	Input	Interval
NumberOfDevice	**Default**	Input	Interval
OrderAmountHik	**Default**	Input	Interval
OrderCount	**Default**	Input	Interval
PreferedOrderC	**Default**	Input	Nominal
PreferredLoginD	**Default**	Input	Nominal
PreferredPayme	**Default**	Input	Nominal
SatisfactionScor	**Default**	Input	Interval
Tenure	**Default**	Input	Interval
WarehouseToHc	**Default**	Input	Interval

Figure 7. "CityTier", "Gender" and "HourSpendOnApp" were dropped.

Name	Use	Method	Use Tree	Role	Level
CashbackAmoun	Default	Default	Default	Input	Interval
Churn	Default	Default	Default	Target	Binary
Complain	Default	Default	Default	Input	Interval
CouponUsed	Default	Median	Default	Input	Interval
DaySinceLastOr	Default	Median	Default	Input	Interval
MaritalStatus	Default	Default	Default	Input	Nominal
NumberOfAddre	Default	Default	Default	Input	Interval
NumberOfDevice	Default	Default	Default	Input	Interval
OrderAmountHik	Default	Median	Default	Input	Interval
OrderCount	Default	Median	Default	Input	Interval
PreferedOrderC	Default	Default	Default	Input	Nominal
PreferredLoginD	Default	Default	Default	Input	Nominal
PreferredPayme	Default	Default	Default	Input	Nominal
SatisfactionScor	Default	Default	Default	Input	Interval
Tenure	Default	Median	Default	Input	Interval
WarehouseToHo	Default	Mean	Default	Input	Interval

Figure 8. The median and mean imputation methods were used on the variables with missing values.

In the initial data exploration step, the variable worth result showed that "CityTier", "HourSpendOnApp", and "Gender" have the least variable worth, hence the data scientist decided to drop the values in order to decrease the runtime of the model, as well as to improve the performance of the model. For imputation, the data scientist decided to use the median imputation method for "CouponUsed", "DaysSinceLastOrder", "OrderAmountHikeFromlastYear", "OrderCount", and "Tenure". This is because as shown in the output of the StatExplore in Figure 5, the min and max values of these variables have a big difference. Thus, to avoid increasing the skewness of the variables, the median imputation method is used. On the other hand, as "WarehouseToHome" is a continuous variable, the mean imputation method is used.

7.2 Class Imbalance

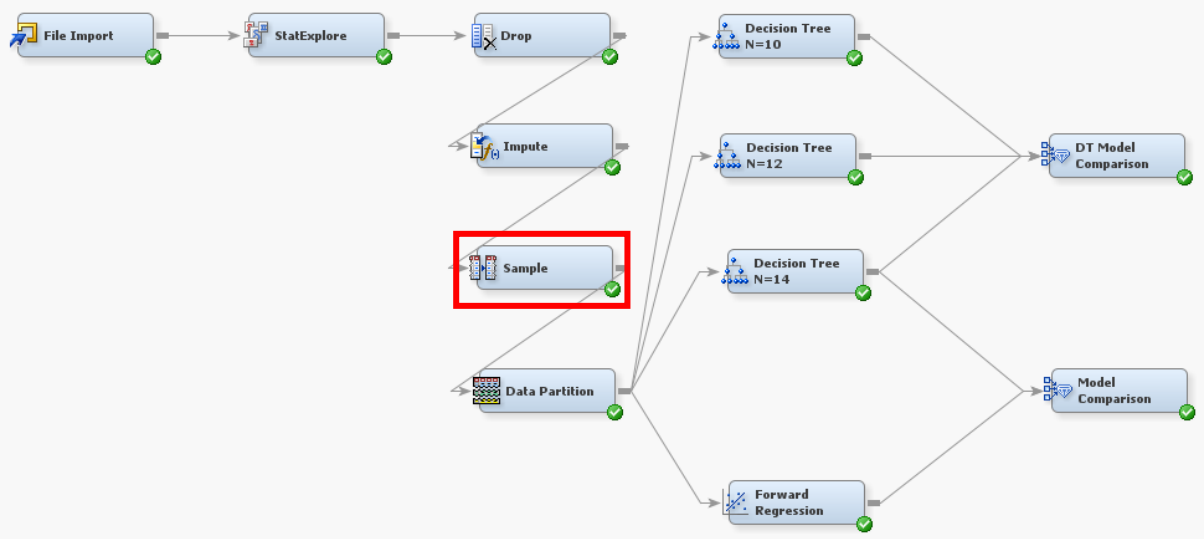

Figure 9. The Sample method is used to solve the class imbalance problem.

Property	Value
Train	
Variables	
Output Type	Data
Sample Method	Default
Random Seed	12345
Size	
Type	Percentage
Observations	
Percentage	50.0
Alpha	0.01
PValue	0.01
Cluster Method	Random
Stratified	
Criterion	Equal
Ignore Small Strata	No
Minimum Strata Size	5
Level Based Options	
Level Selection	Event
Level Proportion	100.0
Sample Proportion	50.0

Figure 10. The percentage and criterion properties were adjusted.

Summary Statistics for Class Targets
(maximum 500 observations printed)

Data=DATA

Variable	Numeric Value	Formatted Value	Frequency Count	Percent	Label
Churn	0	0	4682	83.1616	Churn
Churn	1	1	948	16.8384	Churn

Data=SAMPLE

Variable	Numeric Value	Formatted Value	Frequency Count	Percent	Label
Churn	0	0	948	50	Churn
Churn	1	1	948	50	Churn

Figure 11. The target variable is now balanced.

As shown in the figures above, the Sample method is used to solve the class imbalance problem. In the original dataset, there are 4682 existing customers and 948 churned customers. Therefore, there is a huge difference between the two flags, causing a class imbalance problem. After employing the Sample method by altering the percentage and criterion properties to 50% and "equal", the class imbalance problem is now solved.

7.3 Data Splitting

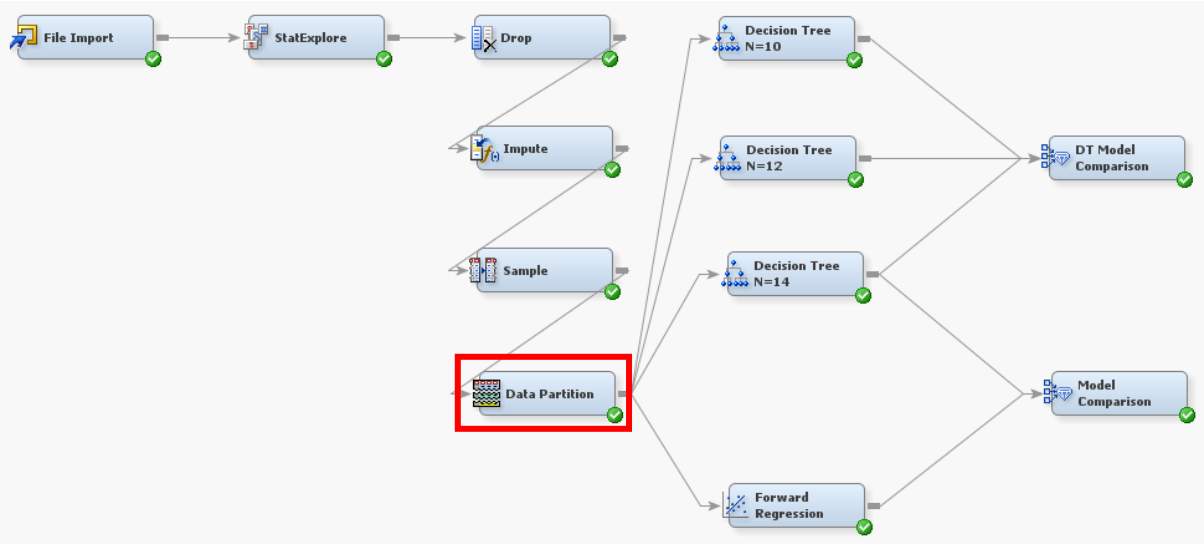

Figure 12. Data splitting is performed using the Data Partition method.

Property	Value
General	
Node ID	Part
Imported Data	
Exported Data	
Notes	
Train	
Variables	
Output Type	Data
Partitioning Method	Default
Random Seed	12345
Data Set Allocations	
Training	70.0
Validation	30.0
Test	0.0
Report	
Interval Targets	Yes
Class Targets	Yes
Status	
Create Time	11/22/22 11:11 AM
Run ID	a9ba14ef-04d7-364a-aaca-

Figure 13. The data is split into a 70:30 ratio.

Partition Summary

Type	Data Set	Number of Observations
DATA	EMWS7.Smpl_DATA	1896
TRAIN	EMWS7.Part_TRAIN	1327
VALIDATE	EMWS7.Part_VALIDATE	569

Figure 14. Data splitting is successful.

After performing data sampling, the dataset is now ready to be split into the training set and the test set. The Data Partition method is employed for this step. The dataset is split into a 70:30 ratio, with 70% observations for the training set, and 30% observations for the test set. It can be seen in Figure 14 that the data-splitting procedure is now completed.

8.0 Modelling

8.1 Decision Tree

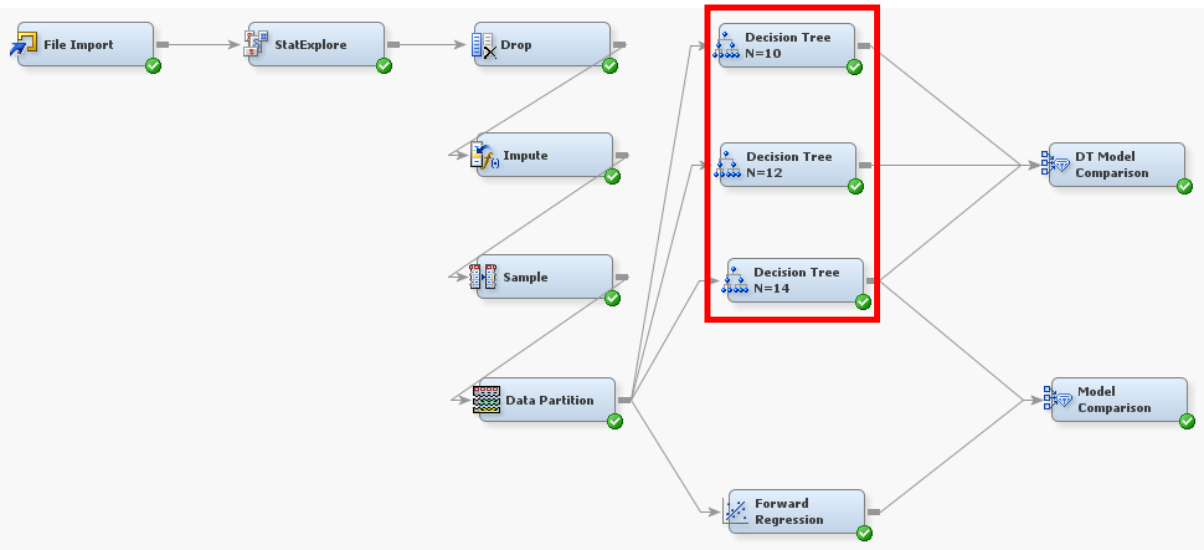

Figure 15. Three decision tree models with different number of leaves were built (N = 10, 12, 14).

Figure 16. Subtree Assessment Plot: Misclassification Rate of decision tree N = 14.

First, in the modelling stage, three decision trees with different number of leaves (N = 10, 12, 14) were built. Then, the third decision tree with 14 leaves was chosen as it can be seen in Figure 16 that the misclassification rate of the tree stabilises at around 14 leaves.

8.1.1 Interpretation of the Decision Tree Model

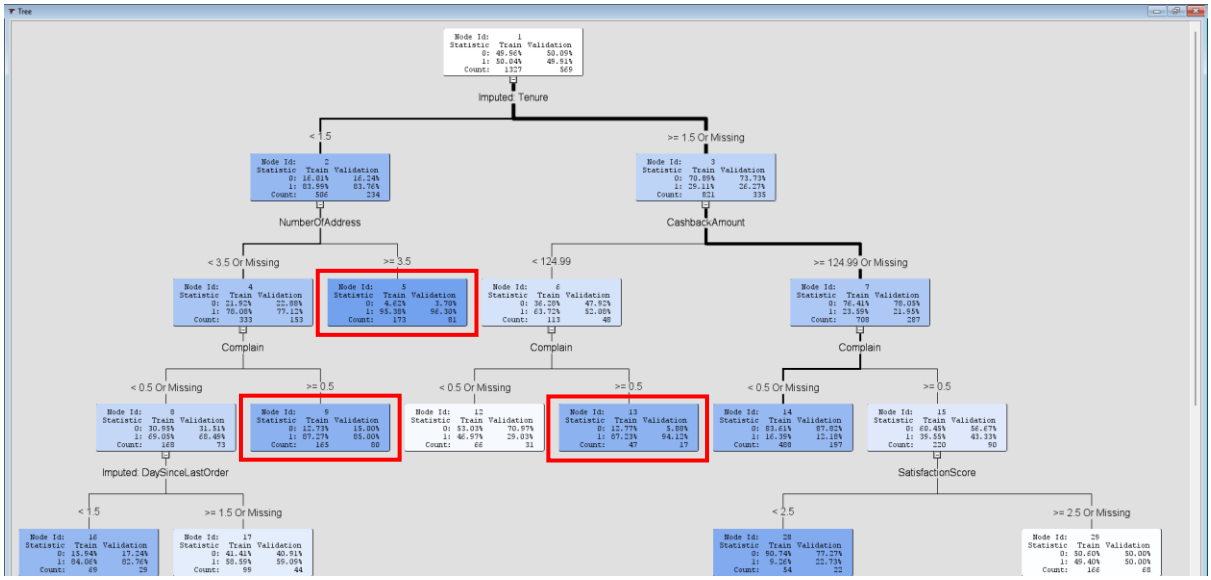

Figure 17. Decision Tree N = 14 tree results.

Figure 17 shows the tree results of the Decision Tree N = 14. As it can be seen, the boxes highlighted with the red boxes have a darker shade of blue colour, this indicates that the nodes have a higher probability of outcome. Hence, the node rules of the nodes are shown and explained below.

```
*------------------------------------------------------------*
 Node = 5
*------------------------------------------------------------*
if NumberOfAddress >= 3.5
AND Imputed: Tenure < 1.5
then
 Tree Node Identifier   = 5
 Number of Observations = 173
 Predicted: Churn=1 = 0.95
 Predicted: Churn=0 = 0.05
```

According to the node 5 shown above, it can be seen that if the number of address entered by the customer is more than or equal to 3.5, and tenure is less than 1.5, there is a 95% chance that the customer will churn.

```
*-------------------------------------------------------------*
 Node = 9
*-------------------------------------------------------------*
if NumberOfAddress < 3.5 or MISSING
AND Imputed: Tenure < 1.5
AND Complain >= 0.5
then
 Tree Node Identifier   = 9
 Number of Observations = 165
 Predicted: Churn=1 = 0.87
 Predicted: Churn=0 = 0.13
```

According to the node 9 shown above, it can be seen that if the number of address entered by the customer is less than 3.5, tenure is less than 1.5, and complain is more than or equal to 0.5, there is an 87% chance that the customer will churn.

```
*-------------------------------------------------------------*
 Node = 13
*-------------------------------------------------------------*
if Imputed: Tenure >= 1.5 or MISSING
AND Complain >= 0.5
AND CashbackAmount < 124.99
then
 Tree Node Identifier   = 13
 Number of Observations = 47
 Predicted: Churn=1 = 0.87
 Predicted: Churn=0 = 0.13
```

According to the node 13 shown above, it can be seen that if tenure is more than or equal to 1.5, complain is more than or equal to 0.5, and the cash back amount received by the customer is less than $124.99, then there is an 87% chance that the customer will churn.

8.2 Forward Logistic Regression

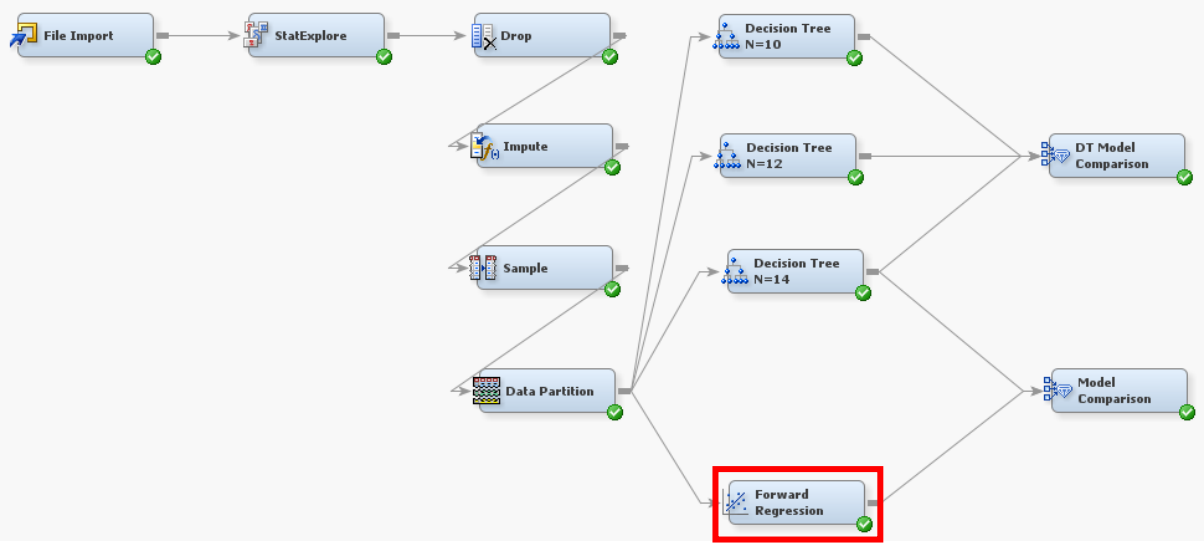

Figure 18. The forward logistic regression model was built.

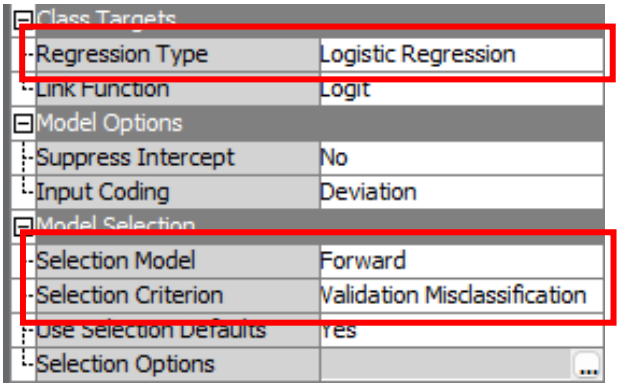

Figure 19. The selection model is set to "Forward" and the selection criterion is set to "Validation Misclassification".

Based on the figures shown above, it can be seen that the next model that was built is the forward logistic regression model, with the selection model set to "Forward" and the selection criterion set to "Validation Misclassification".

8.2.1 Interpretation of the Model

Output

Summary of Forward Selection

Step	Effect Entered	DF	Number In	Score Chi-Square	Pr > ChiSq	Validation Misclassification Rate
1	IMP_Tenure	1	1	305.0787	<.0001	0.2935
2	Complain	1	2	89.8370	<.0001	0.2408
3	NumberOfAddress	1	3	33.0679	<.0001	0.2267
4	PreferedOrderCat	5	4	60.7397	<.0001	0.1880
5	MaritalStatus	2	5	40.3493	<.0001	0.1986
6	IMP_WarehouseToHome	1	6	21.5062	<.0001	0.1986
7	IMP_DaySinceLastOrder	1	7	19.0246	<.0001	0.1968
8	IMP_OrderCount	1	8	25.2782	<.0001	0.2039
9	SatisfactionScore	1	9	19.3937	<.0001	0.1968
10	PreferredPaymentMode	6	10	25.6997	0.0003	0.2039
11	NumberOfDeviceRegistered	1	11	10.1856	0.0014	0.1916
12	CashbackAmount	1	12	12.8364	0.0003	0.1916

The selected model, based on the misclassification rate for the validation data, is the model trained in Step 4. It consists of the following effects:

Intercept Complain IMP_Tenure NumberOfAddress PreferedOrderCat

Likelihood Ratio Test for Global Null Hypothesis: BETA=0

-2 Log Likelihood Intercept Only	-2 Log Likelihood Intercept & Covariates	Likelihood Ratio Chi-Square	DF	Pr > ChiSq
1839.612	1303.974	535.6376	8	<.0001

Based on the summary of the forward selection, it can be seen that the model that was selected for the logistic regression is the step 4 model. Besides, for the Chi-square test, the p-value is less than 0.05, indicating that the model is adequate and significant.

Output

Type 3 Analysis of Effects

Effect	DF	Wald Chi-Square	Pr > ChiSq
Complain	1	82.5610	<.0001
IMP_Tenure	1	212.8406	<.0001
NumberOfAddress	1	46.3511	<.0001
PreferedOrderCat	5	57.9262	<.0001

According to the type 3 analysis of effects, it can be seen that “Complain”, “Tenure”, “NumberOfAddress”, and “PreferredOrderCat” are significant and important to the model as all their p-values are less than 0.05 for the Chi-square test.

Output

Analysis of Maximum Likelihood Estimates

Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq	Standardized Estimate	Exp(Est)
Intercept		1	0.2712	0.1844	2.16	0.1413		1.312
Complain		1	1.3108	0.1443	82.56	<.0001	0.3518	3.709
IMP_Tenure		1	-0.1871	0.0128	212.84	<.0001	-0.8128	0.829
NumberOfAddress		1	0.1981	0.0291	46.35	<.0001	0.2799	1.219
PreferedOrderCat	Fashion	1	0.0273	0.1802	0.02	0.8796		1.028
PreferedOrderCat	Grocery	1	0.00845	0.3343	0.00	0.9798		1.008
PreferedOrderCat	Laptop & Accessory	1	-1.0195	0.1533	44.26	<.0001		0.361
PreferedOrderCat	Mobile	1	0.3406	0.1742	3.82	0.0506		1.406
PreferedOrderCat	Mobile Phone	1	-0.1270	0.1538	0.68	0.4088		0.881

Odds Ratio Estimates

Effect		Point Estimate
Complain		3.709
IMP_Tenure		0.829
NumberOfAddress		1.219
PreferedOrderCat	Fashion vs Others	0.476
PreferedOrderCat	Grocery vs Others	0.467
PreferedOrderCat	Laptop & Accessory vs Others	0.167
PreferedOrderCat	Mobile vs Others	0.651
PreferedOrderCat	Mobile Phone vs Others	0.408

Based on the analysis of maximum likelihood estimates and odds ratio estimates results, it can be seen that "Complain" has the strongest influence on the target variable. As the point estimate value of "Complain" is positive 3.709, it means that as Complain increases by 1 unit, the odds of the customer churning increase by 2.709 as compared to the customer not churning.

9.0 Evaluation – Critical Analysis of the Models

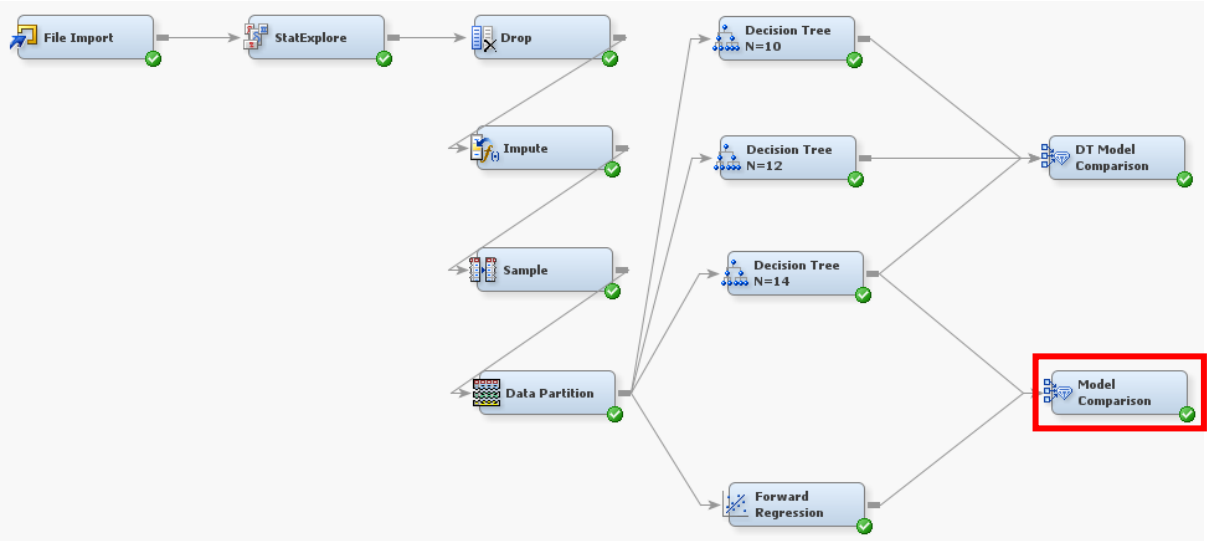

Figure 20. Model comparison of Decision Tree N = 14 and Forward Logistic Regression.

Fit Statistics

Selected Model	Predecessor Node	Model Node	Model Description	Target Variable	Target Label	Selection Criterion: Valid: Misclassification Rate	Train: Misclassification Rate	Valid: Misclassification Rate
Y	Tree	Tree	Decision ...	Churn	Churn	0.168717	0.188395	0.168717
	Reg	Reg	Forward ...	Churn	Churn	0.188049	0.217031	0.188049

Figure 21. Comparison of the misclassification rate of the two models.

Event Classification Table

Model Selection based on Valid: Misclassification Rate (_VMISC_)

Model Node	Model Description	Data Role	Target	Target Label	False Negative	True Negative	False Positive	True Positive
Reg	Forward Regression	TRAIN	Churn	Churn	124	499	164	540
Reg	Forward Regression	VALIDATE	Churn	Churn	44	222	63	240
Tree	Decision Tree N=14	TRAIN	Churn	Churn	111	524	139	553
Tree	Decision Tree N=14	VALIDATE	Churn	Churn	35	224	61	249

Figure 22. Event classification table of the two models.

Decision Tree	0 = Stay	1 = Churn
0 = Stay	TP (249)	FP (61)
1 = Churn	FN (35)	TN (224)

Forward Regression	0 = Stay	1 = Churn
0 = Stay	TP (240)	FP (63)
1 = Churn	FN (44)	TN (222)

Using the event classification table, the accuracy of the models for the validation set can be calculated using the formula $\frac{TP+TN}{TP+TN+FP+FN}$. The accuracy of the decision tree model is 83.1%, whereas the accuracy of the forward regression model is 81.2%. Besides, based on the misclassification rate for the validation set, the decision tree model has a lower misclassification rate than the forward regression model. Therefore, it can be said that the decision tree model is slightly better compared to the forward regression model.

According to the event classification table, it can be seen that both of the models have a higher false positive than false negative predictions, which means the models are more prone to labelling a loyal customer as a customer that is going to churn. In an e-commerce business setting, it is better for the model to have more false positives than false negatives as it is more costly to think that a customer is going to churn is a loyal customer, as this would require a lot of expenses for the business to do target marketing or giving promotions to the customers who

are going to churn. Hence, based on the accuracy and misclassification rate, it can be said that the decision tree model would be a more suitable model to be used to predict customer churn. The results, evaluations, and interpretations of the models are summarised in the table below:

Table 2. Results and interpretations of the models.

Model Name	Settings / Properties	Results (Model Performance)	Interpretation of Outcomes and Suggestions
Decision Tree	Method: N Number of Leaves: 14 Assessment Measure: Decision	MISC Train: 0.188 MISC Valid: 0.169 Accuracy Train: 81.2% Accuracy Valid: 83.1%	There is a 95% chance that a customer will churn if the number of addresses entered by the customer is more than or equal to 3.5 and tenure is less than 1.5. There is an 87% chance that a customer will churn if the number of addresses entered by the customer is less than 3.5, tenure is less than 1.5, and complaints are more than or equal to 0.5. There is also an 87% chance that a customer will churn if tenure is more than or equal to 1.5, the complaint is more than or equal to 0.5, and the cashback amount received by the customer is less than $124.99.
Forward Regression	Regression Type: Logistic Regression Selection Model: Forward Selection Criterion: Validation Misclassification	MISC Train: 0.217 MISC Valid: 0.188 Accuracy Train: 78.3% Accuracy Valid: 81.2%	The model is relatively strongly influenced by the variable “Complain”. This means the customer will likely churn if he has raised any complaints. Hence, the e-commerce business must always be mindful when a customer posts a complaint and take action immediately to avoid the customer from churning. Good customer service must be given to the customers to retain them.

10.0 Discussion and Conclusion

In conclusion, the decision tree model is better suitable than the forward regression approach for predicting customer churn. This is because the decision tree model provides the probabilities of all possible outcomes and specifies the factors that impact those probabilities, making it easier to comprehend. The decision tree model is more practical since it can generate nodes for each attribute, which can also be modified to meet specific needs. In addition, the e-commerce company would like to identify the traits and behaviours of customers who are likely to depart, and the decision tree model could help with this. With a logistic regression model, the business would only know which variable is significant in predicting the outcome. Still, when the company wants to know the exact probability of a customer churning under specific conditions, it is provided with very little information. By deploying a decision tree model, the e-commerce company can achieve its business objectives of identifying customers who are likely to churn and taking preventative action based on the probability assigned to each customer.

A few conclusions are drawn based on the results obtained from the decision tree and logistic regression models. Firstly, if the customer's tenure is less than 1.5, or if the cashback amount received in the last month is less than $124.99, the customer is more likely to churn. Besides, according to the nodes in the decision tree model and the odds ratio estimates of the logistic regression model, it is noticed that the variable "Complain" plays a significant role in determining whether the customer will churn or not. Based on the decision tree model results, it can be seen that if the customer raised more than or equal to 0.5 complaints in the last month, the customer has a very high chance to churn.

Therefore, based on the conclusion drawn, there are a few preventive actions that the e-commerce company can take to counter the situation. First and foremost, if the company receives any customer complaints, it must take action immediately and assist the customers in their post-purchasing journey. After the customer purchase products from the company, the company can also follow up with the customers and ask for their feedback on the products; if there are defects in the products or dissatisfaction among the customers, the company can immediately assist the customer in improving the customers' satisfaction. Besides, the company can also build customer interaction through social media or digital marketing by

posting its latest products and how it differs from other e-commerce businesses. This will increase customer engagement and, at the same time, increase the company's revenue. Moreover, the company can also send personalised emails to the customers according to the products they purchased the most from the company.