An Investigation on How Classifiers Algorithm in Predictive Analytics Prevents Customers Forfeiture.

MSc Research Project
Data Analytics

Kennedy Osemede
16111974

School of Computing
National College of Ireland

Supervisor: Mr. Thibaut Lust

National College of Ireland
Project Submission Sheet – 2017/2018
School of Computing

Student Name:	Kennedy Osemede
Student ID:	16111974
Programme:	Data Analytics
Year:	2017
Module:	MSc Research Project
Lecturer:	Mr. Thibaut Lust
Submission Due Date:	11/12/2017
Project Title:	An Investigation on How Classifiers Algorithm in Predictive Analytics Prevents Customers Forfeiture.
Word Count:	XXX

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are encouraged to use the Harvard Referencing Standard supplied by the Library. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action. Students may be required to undergo a viva (oral examination) if there is suspicion about the validity of their submitted work.

Signature:	
Date:	10th December 2017

PLEASE READ THE FOLLOWING INSTRUCTIONS:

1. Please attach a completed copy of this sheet to each project (including multiple copies).
2. **You must ensure that you retain a HARD COPY of ALL projects**, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer. Please do not bind projects or place in covers unless specifically requested.
3. Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

An Investigation on How Classifiers Algorithm in Predictive Analytics Prevents Customers Forfeiture.

Kennedy Osemede
16111974
MSc Research Project in Data Analytics

10th December 2017

Abstract

This research is focused on predicting if a customer will leave or not based on using various machine learning techniques. The goal is to determine what accuracy each of those models will produce when intercompared against each other to predict if a customer will either leave or not and the determine the major factor behind their decision to leave. To achieve the above-mentioned goal, Logistic Regression, Naive Bayes, Decision Tree and Support Vector Machine models were implemented on a Customer Churn Dataset to determine the accuracy, precision and F-measure score. Logistic Regression achieved the best classification accuracy of 79.08 % and precision of 62.96% followed by Support Vector Machine which scored 77.24% Accuracy and 56.49% Precision.

1 Introduction

Identifying which customer will leave or stay has always been an issue for companies. Most organization have identified that it costs them more to acquire a new customer than to retain the service of their current and loyal customers.

The goal of retaining the services of current customers gives them better return in terms of net profit than doing the hard work of acquiring a new one. (Rust and Zahorik; 1993). There have been quite many data mining techniques which has been deployed to analyze which client is most likely to churn for actionable insights. (Hung et al.; 2006).

The predictive analytics market according to a research carried out by Insights will be valued at about $2.3billion by 2019. Therefore, a lot of organizations and prospective SME's have seen that predictive analytics is the future and they are starting to key into the idea of Predictive Analytics to get more Insight from their data and make informed decisions.

There are 2 distinct type of churners in ecommerce, Incidental and Complete Churn. Complete Churn is referred to as a customer not wanting to have anything to do with that company / product again while Incidental churn occurs when due to further circumstances, the customer cannot patronize the company again. It can be due to the fact the customer changed location etc. (Wu et al.; 2010) Churn is another word that is used to classify if a Customer will leave or not.

This research aims at proffering a solution to SME's and businesses by Investigating and Comparing Predictive Analytics techniques using machine learning to determine what accuracy the best model gives and further insights as to why they are likely to leave. The rest of the research is structured as follows, Section One highlights the specification of the project which also includes the research question, the study purpose and the necessary research variables.

1.1 Project Specification

1.1.1 Research Question

¨Which classifier model will be the best to use for the prevention of customer forfeiture with the application of machine learning.¨

1.1.2 Sub Research Question

Why would the model be the best fit and what other factors influences a customer to leave?

1.2 Purpose of Study

The purpose of this study is to Investigate with the application of various data mining techniques to help establish how classifier algorithms can help give insights on how to prevent customers from leaving.

2 Customer Churn

Customer churn is the process when a client / customer seizes to carry out business or further transaction with a service or company. The main goal behind acquiring customer loyalty is to make sure they are better satisfied with their current service and in that process, more can be sold to them.

There is a need for companies to have a specific method for performing churn analysis during a given period as Churn hinders growth. There are different ways at which various organizations calculate or predict customer churn. With the introduction and application of Machine Learning, predicting customer churn became easier and much more efficient and effective.

2.1 Machine Learning

Machine learning is the process of extracting insightful and relevant information by deploying certain algorithms that learn from the data. Priyadharshini (2017). Some examples of Machine learning problems are identifying different faces in images, Filtration of spams, Fraud detection and prevention, Segmentation of customers and many others.

Data mining and Machine learning go hand in hand, they also share striking similarity by usage of the same technique and algorithms, but their kind of prediction may differ. While data mining is known for its pattern recognition, Machine learning recreates established patterns and knowledge. There are two major machine learning methods which are being used today, they are Supervised and Unsupervised Machine learning. According to research, A lot of people carry out more of Supervised learning than Unsupervised machine learning methods.

Supervised Machine learning involves clearly identifying the inputs and outputs, while the algorithm is trained using examples that are labelled.

It takes the inputs and output to find the corresponding error, it modifies the model based on the inputs. It is commonly used in events when you are predicting the future based on historical data. Some type of supervised learning methods is Regression (Logistic Regression), Classification (Bayesian Network, Decision Tree, Support Vector Machine) and Gradient Boosting. Priyadharshini (2017).

Unsupervised learning is used on data without historical knowledge of anything. It is used to find patterns and clusters with certain attributes.

2.2 Logistic Regression

This is a model that uses variables with a finite set of values for prediction. (Apampa; 2016). Prediction lies on the application of many predictors which could either be categorical or numerical. Saedsayad (2017). It evaluates class probabilities by utilizing the logit function. Lets take an example, A Logistic Regression can be used to predict the outcome of an election. Based on the multiple input criteria the model can take the constituent the electorate is from, the party the candidate belongs to etc. Due to the historical data which involves the same criteria data input, it can score new instances based on probability that they fall into the category based on the outcome. (SearchBusinessAnalytics; 2017).

2.3 Naive Bayes

This is a concept of probability that predicts the outcome based on existing knowledge, it is a powerful algorithm that is utilized for classification task, very useful for the analysis of large dataset as it is easy to build and gives no hassles over iterative estimation of parameters. Saedsayad (2017). It assumes that all the features are conditionally independent given the class label.

2.4 Decision Tree

Decision Tree is utilized to tear apart problems that are complex, every branch of the tree can be an outcome with possibility of solving the problem.

The model is utilized in philosophy, forecasting of economy and finance to mention a few. The structure of the tree helps into drawing up a conclusion for any given problem which maybe or is complex in nature.

Just like other models, DT can also be used to solve classification and regression problems also, the main idea behind using decision tree is to build a training model that can be used to predict a target variable or class value by decision rules which will be learnt from the training data itself. The tree always tries to solve a problem by using Internal Node and Leaf Node, every internal node of the tree is an attribute and leaf node is a class label. A node displays the analysis via its objective on a 1D display. As soon as an ideal relationship has been spotted between values that are inputted, they are grouped together to give decision tree model. (Wagle et al.; 2017)

2.5 Support Vector Machine

Created by Vapnik Vladimir, Support Vector Machine is utilized to solve both classification problems and regression problems. Utilizing both kernel functions of Nonlinear and Linear, it gets information by looking for the hyperplane which is the point of separation between two data. (Radhimeenakshi; 2016)

It can be used as a tool for prediction by searching for a line or a boundary of decision which is also known as Hyperplane which separates the classes or data easily to avoid overfitting of the data. (Somvanshi and Chavan; 2016)

3 Related Work

(Vafeiadis et al.; 2015) carried out a comparison of different machine learning techniques on a telecommunications churn prediction dataset, the techniques are no other than Support Vector Machine(SVM), Decision tree learning, Naive Bayes, Artificial Neural Networks (ANN) and Regression Analysis.

The Dataset utilized in conducting this experiment was a churn dataset which was obtained from UCI Machine Learning Repository which included binary attributes (Yes/No). It also contained 18 predictors inclusive of the churn binary attribute and 5000 records. To mention a few of the attributes contained in the dataset, they are namely: - Account Length, Total Evening Charge, Area Code, Total Night Minute, Total Evening Calls, Total Evening Minutes, Churn.

Based to their research, Performance Evaluation was initially carried out on each of the models to check for their Recall, Accuracy, Precision and F-measures. The boosting algorithm was also carried to check the performance of each classifier algorithm, the main reason for using a boosting algorithm was to improve the f-measure. At the end of the research based on cross validation, the researchers found out that the top two performing models without boosting were Decision Tree which give an accuracy of 94% and F-measure score of 77% and Support Vector Machine giving an accuracy of 93% and F-measure score of 73%.
Immediately Ada boosting was applied, SVM(SVM-Poly) came out tops with an accuracy of 97% and F-measure score of over 84%, they cited that Naive Bayes and Logistic Regression could not be boosted due to lack of parameters that could be tuned to.

(Apampa; 2016) carried out another research on how banks can make their Customer marketing response effective utilizing Logistic Regression, Decision Tree, Naive Bayes and Random forest predictive models. The models were applied to both balance and unbalanced dataset provided by the bank. The Dataset that was used to conduct the research was a Bank dataset which had about 45,211 records and 17 attributes. To mention a few of the attributes contained in the dataset, they are namely: - Age, Balance, Day, Campaign, Job and the predictor being 'y' which indicated if a customer subscribed to the campaign or not.

After running series of model testing and 10-fold cross validations across each model, he found out that Logistic Regression and Naive Bayes algorithm had Area Under Curve (AUC) of 75.7% and 75.6% respectively. The Random Forest model returned an AUC of 74.2%. The Author cited that the results were way better for the balanced dataset. He further concluded that for all the classifications performed well on AUC, Precision and Recall but Random forest returned a lover AUC score in comparison to Decision Tree and that Random forest does not improve the overall performance of Decision Tree.

(Chen et al.; 2015) carried out a research using predictive analytics techniques to ascertain how the length, frequency and other circumstances in a logistic company influence customer churn. The researchers stressed the importance of performing a customer churn analysis in other to save companies of having to spend on acquiring a new one. The predictive analytics methods utilized in this research were Logistic Regression, Decision Trees, Support Vector Machine and Artificial Neural Network. The initial data set contained about 106,747 customers who had made over 210 million transactions between a specific period. After the performance of Data cleaning and preprocessing, a total of 69,170 customer data was the final output. They segregated the people who had been inactive and defined various reason as to why they could have been inactive.

The performance of the selected models was measured based on their Accuracy, Precision, Recall and F-measure. After the implementation of the selected models, the researchers noted that Decision tree model outclassed other models. Based on their dataset and in conclusion of their research, the however denoted that the influencing factor for customer churn were the Customer Recency, Customer Length and Customer Monetary.

Definition of Key Terms

Accuracy: This is simply the ratio of the observation of the predicted values against the total observation. It is calculated using the formula (TP + TN) / (TP + TN + FP + FN)

Note: TP stands for True Postive, TN stands for True Negative, FP stands for False Positive and FN stands for False Negative.

Precision: This is the ratio of positive predictive observation against the total positive predicted observation. It is calculated using the formula, TP / (TP+FP).

F-Measure: This is the average of precision and recall. It is calculated using the formula, 2*TP / 2 * TP + FP+ FN.

True Positive (True Yes): This is the amount of the correctly predicted values by the model.

True Negative (True No): This is the amount of the correctly predicted negative values.

False Positive (False Yes): This is the amount of falsely predicted values that belong to a correct predicted value. i.e. When the actual value is of class No and it is predicted to be Yes.

False Negative (False No): This is the amount of falsely predictive vales that doesn't belong to the correct predicted values. i.e. If the Actual value predicts Yes and the predicted values says No.

4 Methodology

This research was conducted and structured in the CRISP-DM (Cross Industry Standard Process for Data Mining) structure.

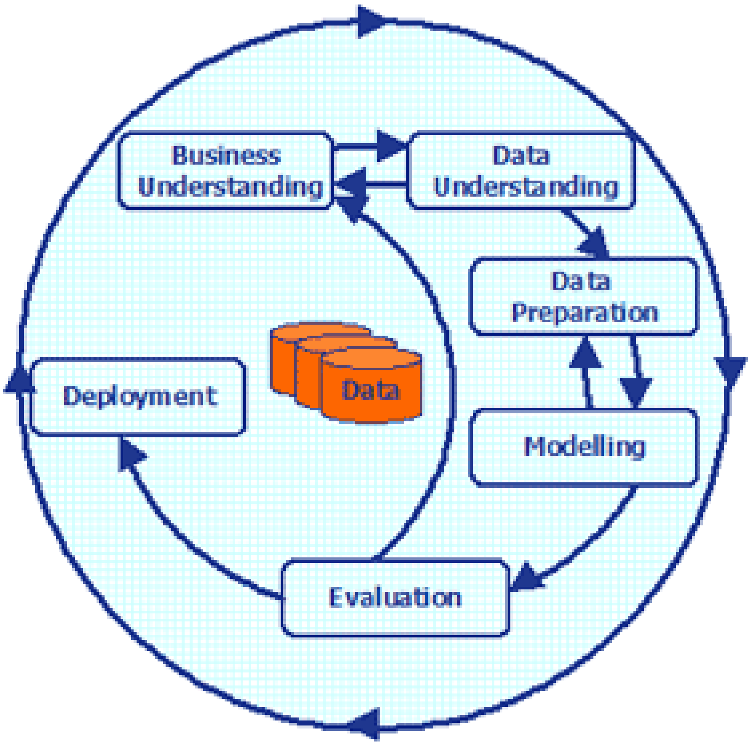


Figure 1: The CRISP-DM Process Flow

The crisp-DM follows the process of Business Understanding to Data Understanding then Data Preparation and Modelling which follows by Model Evaluation and deployment of your model.

The Business Understanding Process: - The phase places emphases on making sure that requirements and objectives of the business are well laid out and understood from the business perspective. After adequate understanding, the knowledge will be used to achieve the data mining objectives which has been set out. Nadali et al. (2011). In this phase, it was well understood that the essence of application of machine learning was to give us better insights as to which customer is leaving and their reasons for leaving. This in turn will make the company or business take further actions as to how to retain their services.

Data Preparation and Understanding: - It is highly important that the data is understood before any process or analysis is performed on it. Discovery of hidden knowledge to form your basis of hypothesis. The data must be prepped in other for the analysis to be performed on it, at this stage the cleaning and transformation is done. Decision on which attribute will be utilized during the analysis is made and NULL Values are removed. (Zapata and Gil; 2011). The dataset utilized is a Telecommunications churn dataset which was downloaded from IBM Watson Analytics with over 7,000 rows and 17 columns. It contains Customer data about their monthly and yearly expenditure with the telco and the plans they subscribed on for that month with other features alongside their likelihood to leave them or not which was a Binary category (Yes/No). The attributes of the above-mentioned dataset to mention a few includes Customer ID, Gender, Marital Status, Dependents, Phone Services, Type of Contract, Payment Method, Monthly Charge, Total Charge, Churn.

Amongst the 7,000 plus rows in the dataset, the churn column contained 5175 No Instances and 1870 Yes Instances. The 5175 No instances indicates that the likelihood of those customers leaving the company was 0% while the 1870 Yes instances indicates that the likelihood of the customers leaving the company was 100%.

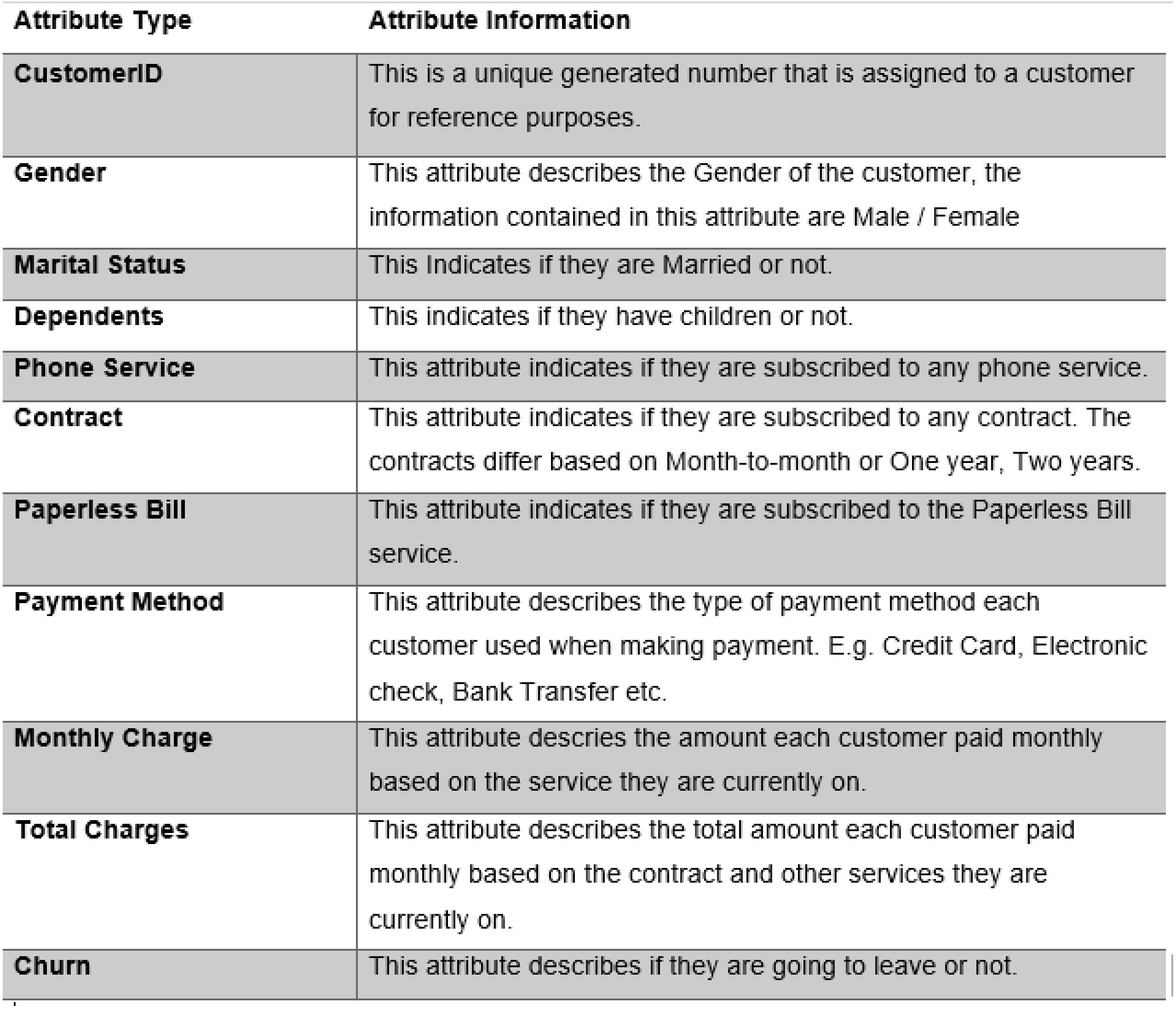
Modelling and Evaluation: Some data mining modelling techniques are considered and selected in this phase, the optimal values of the parameters are selected and applied. (Nadali et al.; 2011). Hitherto to model deployment, evaluation and review of the model is very necessary. It must be compared to tally with the business objective. (Zapata and Gil; 2011). The models that were utilized for this project were Support Vector Machine, Nave Bayes, Decision Tree and Logistic Regression. Each of the model were evaluated against each other and insightful information was gathered from the implementation of the models vis--vis bearing in mind that the business objective is at stake.

Deployment: It was implemented using RapidMiner Studio 7.6.001 Version. After importation of the dataset into the Rapid Miner data repository, data Mining techniques were performed on the data sets to display findings and make recommendations.

5 Implementation

Data Understanding and Preparation

With the acquired dataset, A few data mining models were applied to it to intercompare each of them against one another thereby taking note which of the models gives the best



Attribute Type	Attribute Information
CustomerID	This is a unique generated number that is assigned to a customer for reference purposes.
Gender	This attribute describes the Gender of the customer, the information contained in this attribute are Male / Female
Marital Status	This Indicates if they are Married or not.
Dependents	This indicates if they have children or not.
Phone Service	This attribute indicates if they are subscribed to any phone service.
Contract	This attribute indicates if they are subscribed to any contract. The contracts differ based on Month-to-month or One year, Two years.
Paperless Bill	This attribute indicates if they are subscribed to the Paperless Bill service.
Payment Method	This attribute describes the type of payment method each customer used when making payment. E.g. Credit Card, Electronic check, Bank Transfer etc.
Monthly Charge	This attribute descries the amount each customer paid monthly based on the service they are currently on.
Total Charges	This attribute describes the total amount each customer paid monthly based on the contract and other services they are currently on.
Churn	This attribute describes if they are going to leave or not.

Figure 2: The description of each attribute in the Dataset

accuracy, precision and F-measure score to determine if a Customer will leave or not. The aim of taking this approach is to apply predictive analytics to gain more insight from any given dataset vis-a-vis saving organizations the cost of having to reacquire their loyal customers.

The churn data was split into two cross validation folds (Training data and Testing data) with the 70% of the data being used as training data while the other 30% was used as testing across each model.

5.1 Model Application Results

5.1.1 Logistic Regression

The application of Logistic Regression on the Churn dataset has given the above results, it can be noted that Logistic Regression model gave an Accuracy of 79.08% and precision

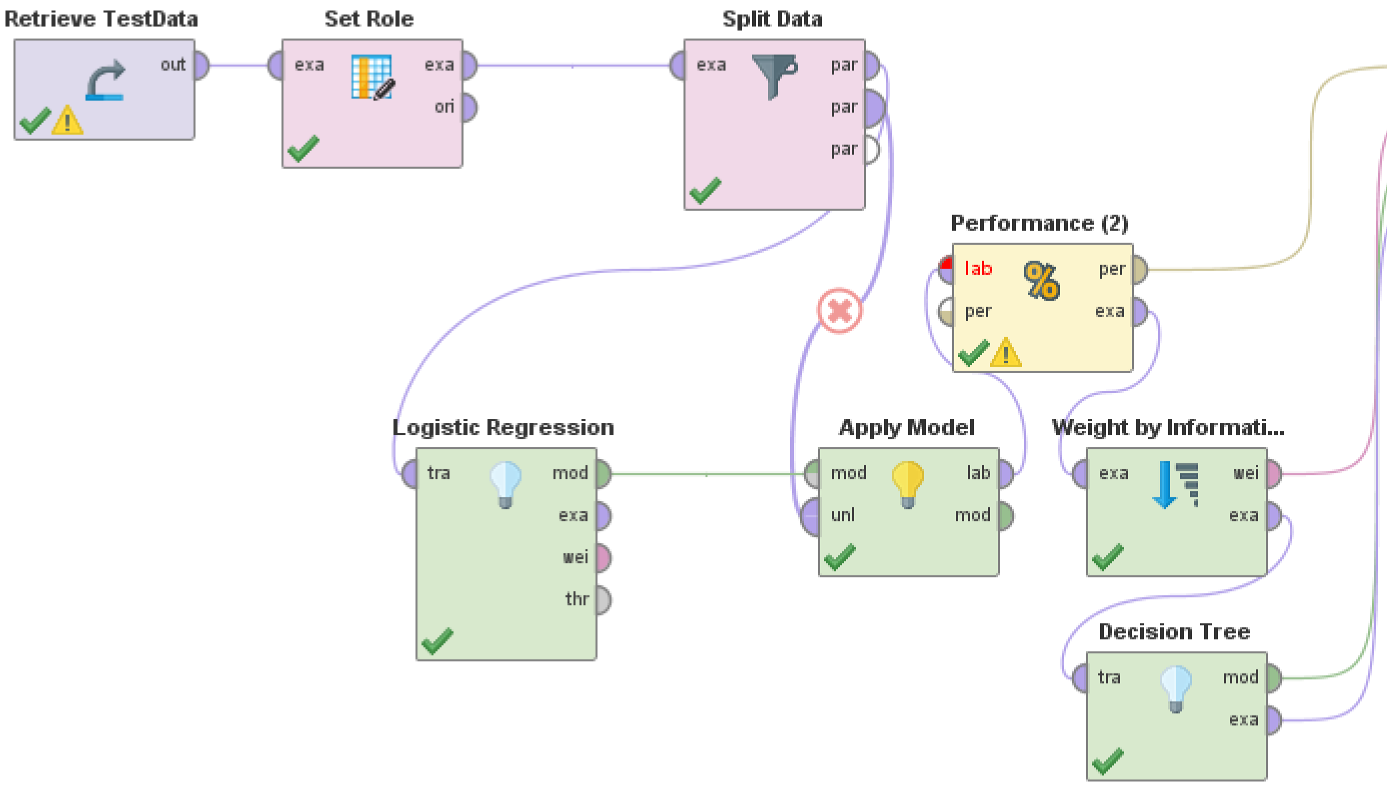
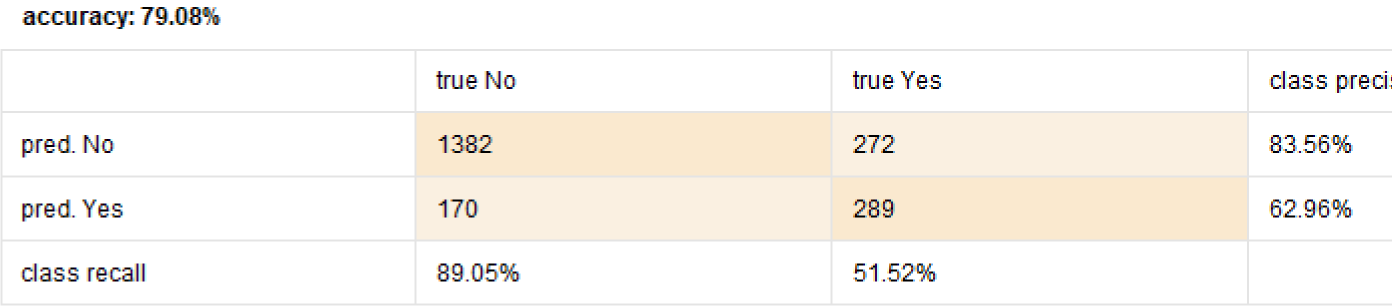


Figure 3: The Process Flow For Logistic Regression Implementation on RapidMiner

of 62.96%. The model classified True Positive (True Yes): 289, True Negative (True No): 1382, False Positive (False Yes): 272, False Negative (False No) 170.

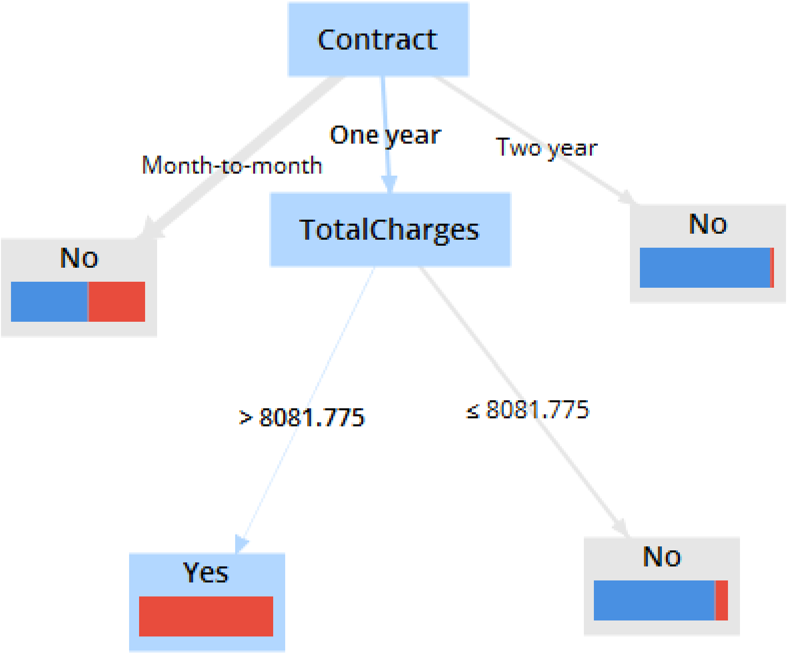


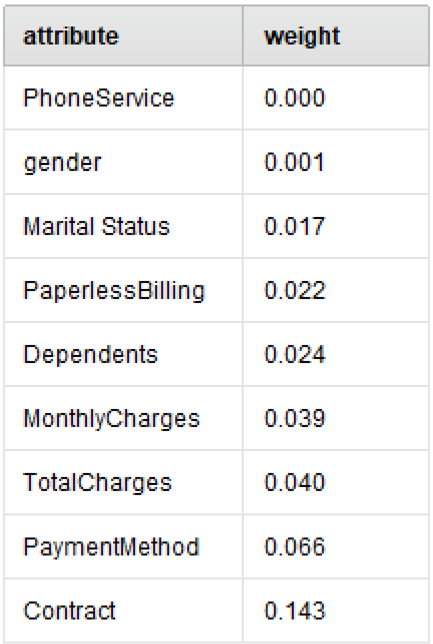
accuracy: 79.08%

	true No	true Yes	class preci
pred. No	1382	272	83.56%
pred. Yes	170	289	62.96%
class recall	89.05%	51.52%	

Figure 4: The Results from Logistic Regression Analysis

Applying Decision Tree to the model paved way for a drill down effect to recognize the most important factor amongst the attributes, it can be noted that the type of contract the customer is currently only plays a major influence in their decision to stay or remain with the telco company. Contract being the main node, has 3 sub attributes (Month-to Month, One Year and Two Years). The chances of a customer churning one a month to month contract is relatively low vis--vis two-year basis. If the total charge based on the contract the customer is on is greater than 8081.775 the chances of the customer leaving is very high, while if it is less than 8081.775 the chances of the customer leaving is very low.





attribute	weight
PhoneService	0.000
gender	0.001
Marital Status	0.017
PaperlessBilling	0.022
Dependents	0.024
MonthlyCharges	0.039
TotalCharges	0.040
PaymentMethod	0.066
Contract	0.143

Figure 5: Decision Tree Insight Into Logistic Regression with Feature Selection

5.1.2 Naive Bayes

After the successful implementation of Logistic Regression on the data set, Naive Bayes model was applied to get further insights in terms of how accurate and precise it would

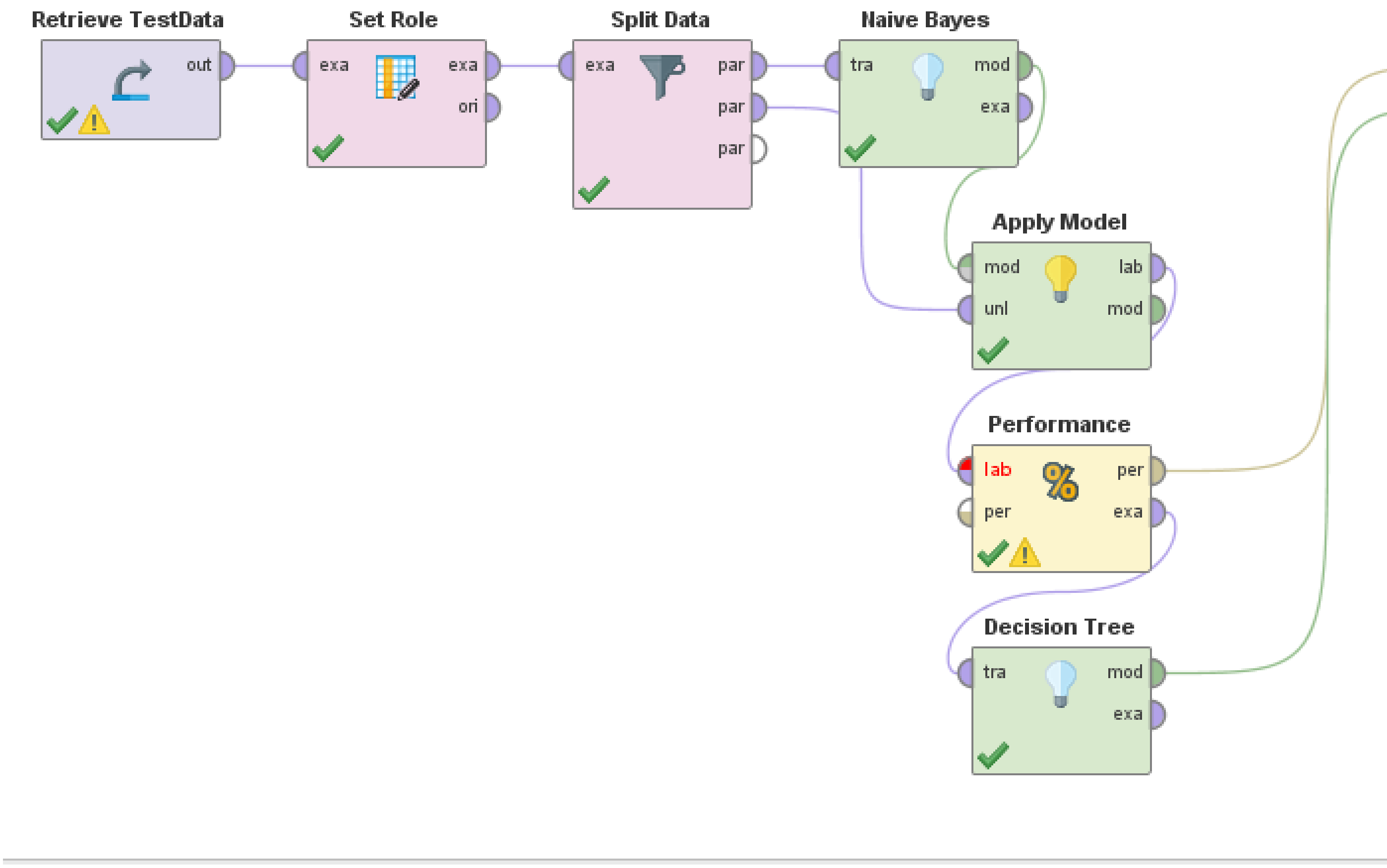


Figure 6: Naive Bayes Process on RapidMiner

determine if a customer will leave based on previous knowledge. The accuracy after implementing Naive Bayes model was 76.38% while the precision was 54.86%. Based on classification, it classified True Positive (True Yes): 350, True Negative (True No): 1264, False Positive (False Yes): 211, False Negative (False No) 288.

5.1.3 Decision Tree

The data was retrieved from the local computer and split into two cross validation folds (70% Training data and 30% Testing data). The Set role was used to set Churn as the label for prediction, Decision tree classification was utilized setting both pre-and post-pruning in other to avoid overfitting of the tree.

The Performance of the model was extracted after successful execution and feature selection was retrieved from the model in other to determine the most insightful information from it. It gave an accuracy of 73.41% and 100% precision. Based on classification, it classified True Positive (True Yes): 2, True Negative (True No): 3617, False Positive (False Yes): 1311, False Negative (False No) 0.

It can be denoted that decision tree is a weak learner as it was only able to identify and classify only 2 customers as True Positives from the testing dataset leading to an accuracy of 100%. Decision Tree was able to identify Contract as the Node with Monthly

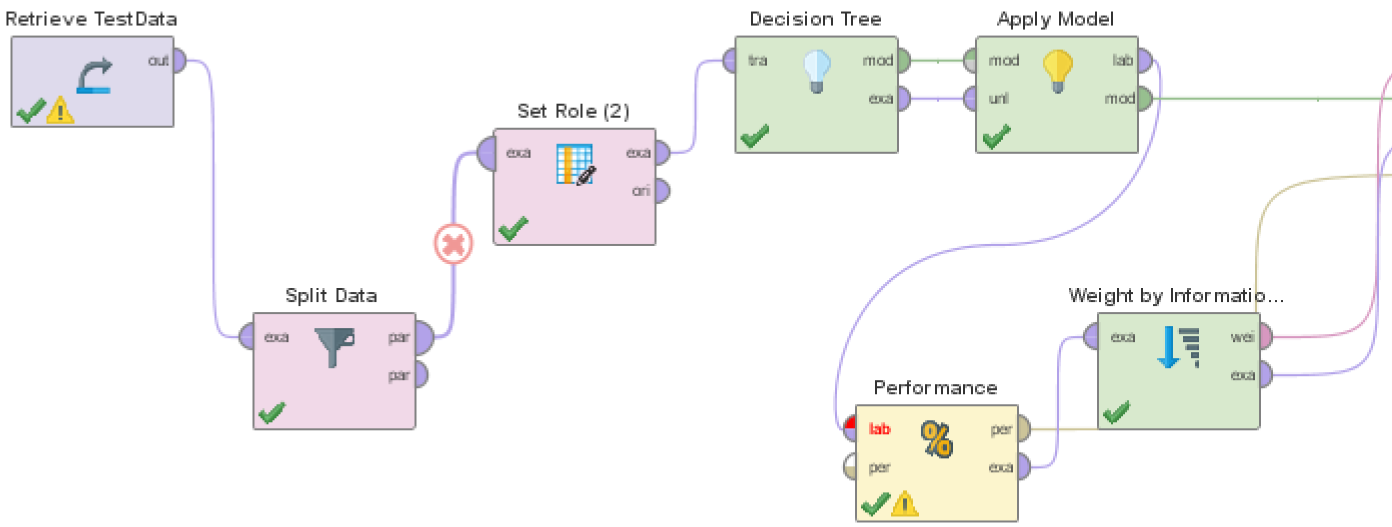
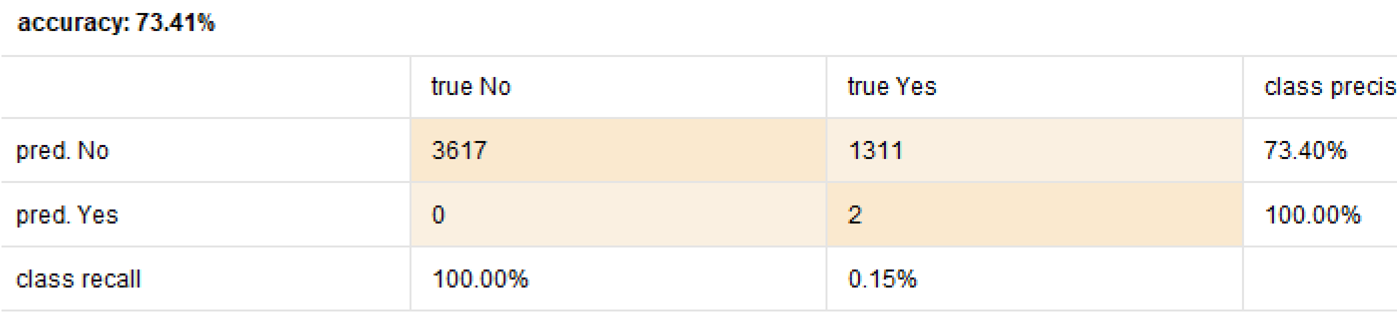


Figure 7: Decision Tree Process on RapidMiner



accuracy: 73.41%

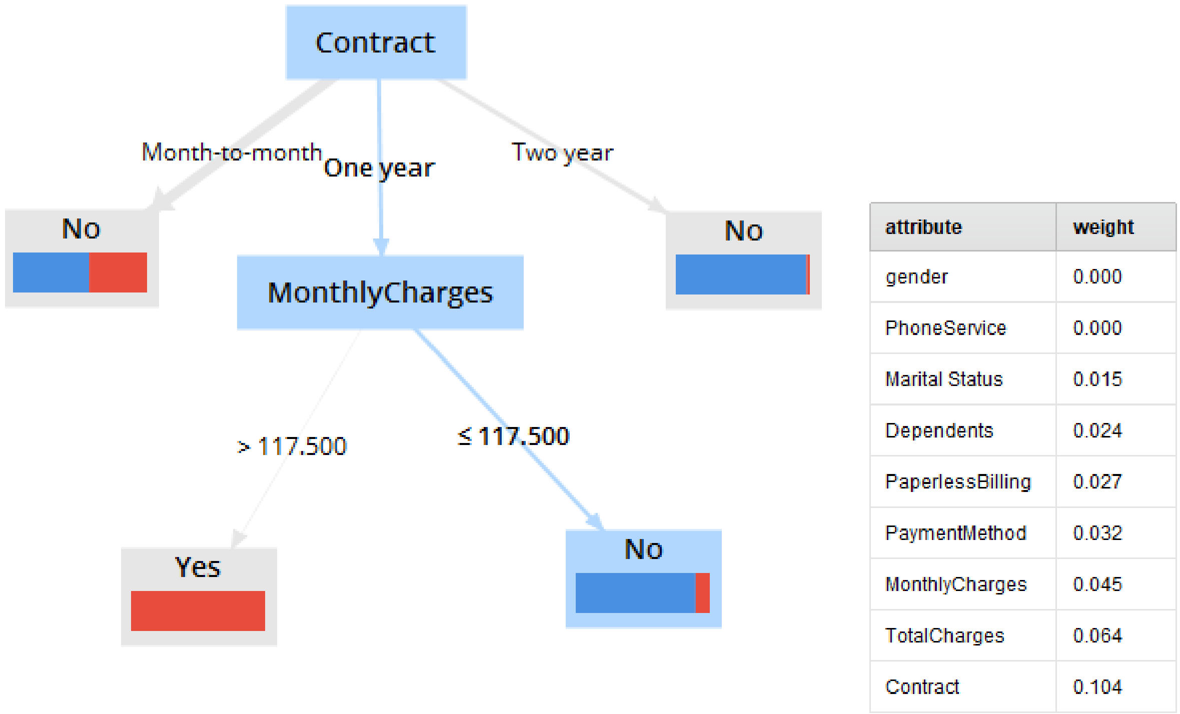
	true No	true Yes	class precis
pred. No	3617	1311	73.40%
pred. Yes	0	2	100.00%
class recall	100.00%	0.15%	

Figure 8: Decision Tree Model Performance

Charges being the leaf node. Just as Decision tree was applied to Logistic Regression for a tree down view of how well the model perform and gather further insight. Based on the type of contract which varies from month to month, one year and two years. The customers on month to Month and Two years contract are not likely to leave based on the result from the model. The customers who are on a one-year contract are likely to leave if their monthly charges are greater than or equal to 117.50 and are less likely to leave if their monthly charges are less than or equal to 117.50.

5.1.4 Support Vector Machine

Data was retrieved from the Data bank and Churn attribute was set as the label for which the prediction will be performed on. The missing Value operator was utilized should any value be missing, it will be replaced. Support Vector Machine doesn't perform any analysis on Binary Operator which was our Churn Label(Yes/No), Nominal to Numerical Operator was utilized to replace that process thereby performing Dummy encoding so that the model could perform its analysis. Dummy encoding is the process of utilizing categorical variable predictors in multiple ways of model estimation.



attribute	weight
gender	0.000
PhoneService	0.000
Marital Status	0.015
Dependents	0.024
PaperlessBilling	0.027
PaymentMethod	0.032
MonthlyCharges	0.045
TotalCharges	0.064
Contract	0.104

Figure 9: Decision Tree Like View with Feature Selection Performance

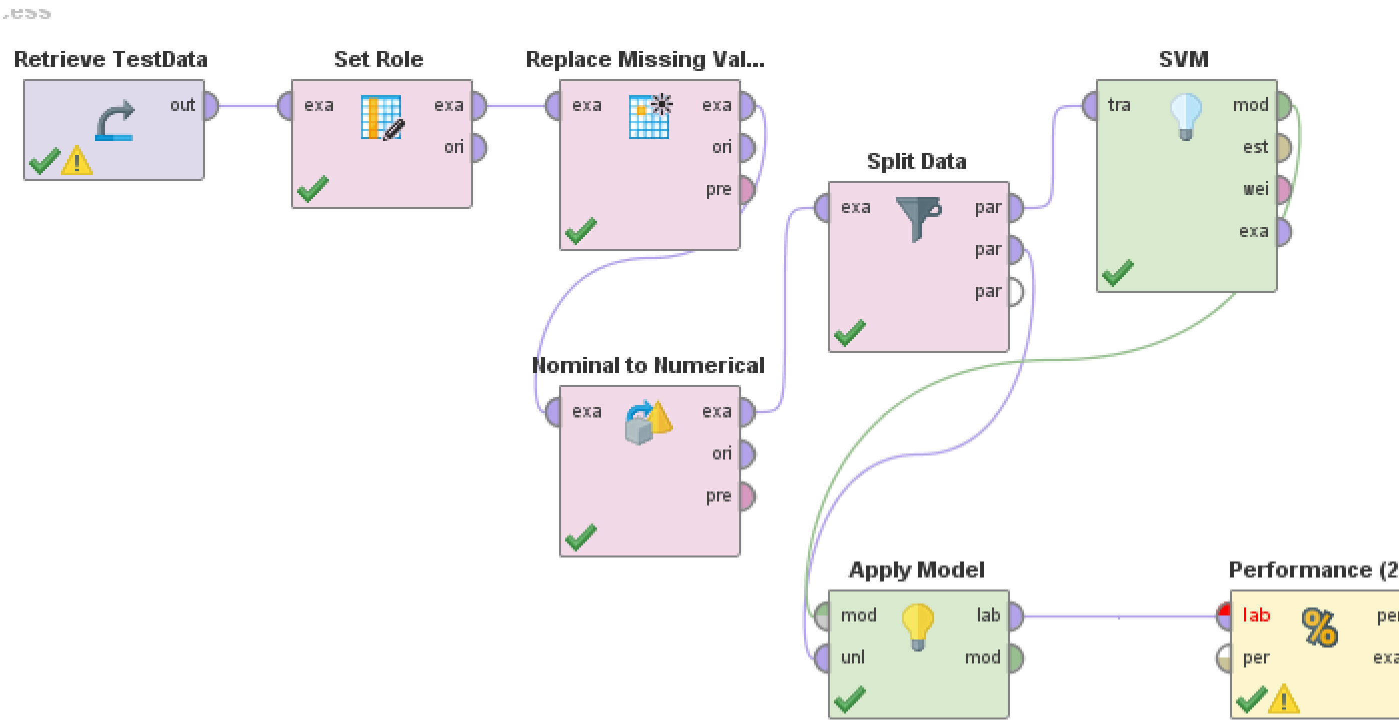
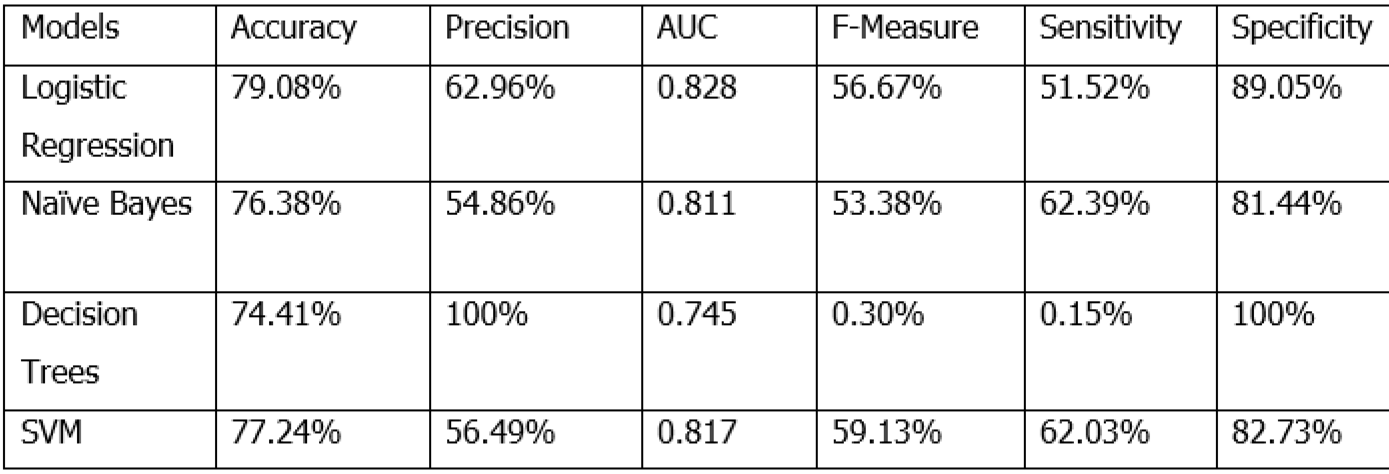


Figure 10: Support Vector Process on RapidMiner

6 Evaluation

Each model was utilized on the same dataset following the same procedure and their accuracy, precision and other factors were noted down after cross validation. It can be seen from the image below that Logistic Regression performed better that all other models namely (Naive Bayes, Decision Tree and Support Vector Machine).



Models	Accuracy	Precision	AUC	F-Measure	Sensitivity	Specificity
Logistic Regression	79.08%	62.96%	0.828	56.67%	51.52%	89.05%
Naïve Bayes	76.38%	54.86%	0.811	53.38%	62.39%	81.44%
Decision Trees	74.41%	100%	0.745	0.30%	0.15%	100%
SVM	77.24%	56.49%	0.817	59.13%	62.03%	82.73%

Figure 11: Evaluation of All Models

It attained an accuracy of 79.08 and a precision of 62.96 with an AUC score of 0.282% and F measure of 56.7%. The sensitivity and Specificity of the model were 51.52% and 89.05%, for a model to considered good, it must be Precise and Accurate.

The Second-best model after Logistic regression to predict if a customer is likely to leave was Support Vector Machine, it gave an accuracy of 77.24% and a precision of 56.49% with an AUC of 0.817% and F-measure of 59.13%. It performed better than Logistic Regression on Sensitivity which was 62.03% and its specificity was 82.73% respectively.

The third model best suitable for the predicting of customer churn is Nave Bayes model, this model gave an accuracy of 76.38% and a precision of 54.86% with an Area Under the Curve(AUC) score of 0.811 and an F-measure score of 53.38%. It performed better than both Logistic Regression and Support Vector Machine based on the Sensitivity (True Positive Rate) which was 62.39% and its specificity was 81.4%.

The fourth best model suitable for predicting customer churn is no other than Decision tree, which gave a model accuracy of 74.41% and a precision value of 100%. Note, a model cannot be 100%. Decision tree only classified 2 (TPs) out of the many data points in the test data. It had the lowest F-measure score which was 0.30% and A sensitivity of 0.15% and 100% specificity.

7 Conclusion and Future Work

With the amount of data generated now and in the future, the need for data mining is the top most agenda for every successful corporation. Data mining in conjunction with predictive analytics is the future, Meaningful information needs to be extracted from data in other to assist businesses or organizations grow and further enlarge its revenue and make valuable decisions as to where they should head to. Having firsthand information from your data using predictive analytics gives you a major edge over your competitors.

The results from this research shows that Logistic Regression Model outperformed the other 3 models when compared against each other. It outperformed them based on its accuracy and precision at predicting if a customer will leave or not. This is however valuable as managers will feel confident going into the office knowing fully well where to focus their resources on based on customer retention. Based on other metrics Logistic Regression classified 51.52% True Positive Rate(Sensitivity) of the data which is less than Support Vector Machines, True Positives rate of 62.03% and its F-measure score of 59.13%. Overall it Had a True Negative Rate Prediction (Specificity) score of 89.05%.

Based on feature selection, it was noted that the type of contract the customers are on plays a major factor on if they are likely to leave or not. With Churn variable being the target variable, feature selection analysis was carried out to gather further insight as to what factors influences them to leave.

A decision tree operator was placed to see what type of contract is influencing them to leave, Customers on a month-to-month and a year contract are less likely to leave even though there are some likely to leave on a month to month basis but the customers who will stay are more than the ones who will leave. Customers who pay more than 8081.775 as total charges are more likely to leave since it is expensive and customers who pay less than 8081.775 are less likely to leave.

There is need for further development of this project as further insights and other metrics of customer behaviors can be analyzed and valuable piece of analysis could be made from it. Other classification models can be applied on the dataset to determine which outperforms the other and recommendation can then be made based on their findings.

8 Acknowledgment

I will like to use this medium to my express my gratitude to Almight God who made it possible to be where I am today. Secondly, I will like to extend my gratitude to my Family (The Osemede's), friends and Loved ones because with their push and support I was able to break barriers and finish this research in due time. Lastly I will like to show a big appreciate to the best supervisor a thesis student can have in the person of Mr. Thibaut Lust, Your vision and guidance were key factors towards the completion of this research. Thank you Sir for making me understand so much during this little time.

References

Apampa, O. (2016). Evaluation of classification and ensemble algorithms for bank customer marketing response prediction, *Journal of International Technology Information Management* **25**(4).

Chen, K., Hu, Y.-H. and Hsieh, Y.-C. (2015). Predicting customer churn from valuable b2b customers in the logistics industry: a case study, *Information Systems and e-Business Management* **13**(3): 475–494.

Hung, S.-Y., Yen, D. C. and Wang, H.-Y. (2006). Applying data mining to telecom churn management, *Expert Systems with Applications* **31**(3): 515–524.

Nadali, A., Kakhky, E. N. and Nosratabadi, H. E. (2011). Evaluating the success level of data mining projects based on crisp-dm methodology by a fuzzy expert system, *Electronics Computer Technology (ICECT), 2011 3rd International Conference on*, Vol. 6, IEEE, pp. 161–165.

Priyadharshini (2017). Machine learning: What it is and why it matters.
URL: *https://www.simplilearn.com/what-is-machine-learning-and-why-it-matters-article*

Radhimeenakshi, S. (2016). Classification and prediction of heart disease risk using data mining techniques of support vector machine and artificial neural network, *Computing for Sustainable Global Development (INDIACom), 2016 3rd International Conference on*, IEEE, pp. 3107–3111.

Rust, R. T. and Zahorik, A. J. (1993). Customer satisfaction, customer retention, and market share, *Journal of retailing* **69**(2): 193–215.

Saedsayad (2017).
URL: *http://www.saedsayad.com/logistic$_r$egression.htm*

SearchBusinessAnalytics (2017). What is logistic regression? - definition from whatis.com.
URL: *http://searchbusinessanalytics.techtarget.com/definition/logistic-regression*

Somvanshi, M. and Chavan, P. (2016). A review of machine learning techniques using decision tree and support vector machine, *Computing Communication Control and automation (ICCUBEA), 2016 International Conference on*, IEEE, pp. 1–7.

Vafeiadis, T., Vafeiadis, K. I., Sarigiannidis, G. and Chatzisavvas, K. C. (2015). A comparison of machine learning techniques for customer churn prediction, *Simulation Modelling Practice and Theory* p. 19.

Wagle, M., Yang, Z. and Benslimane, Y. (2017). Bankruptcy prediction using data mining techniques, *Information and Communication Technology for Embedded Systems (IC-ICTES), 2017 8th International Conference of*, IEEE, pp. 1–4.

Wu, H.-L., Zhang, W.-W. and Zhang, Y.-Y. (2010). An empirical study of customer churn in e-commerce based on data mining, *2010 International Conference on Management and Service Science* .

Zapata, J. C. M. and Gil, N. (2011). Incorporation of both pre-conceptual schemas and goal diagrams in crisp-dm, *Computing Congress (CCC), 2011 6th Colombian*, IEEE, pp. 1–6.