Predicting Telco Customer Churnusing Data Mining Techniques

Team Three Musketeers

Provide Team Members' Names and UIS Emails

Names	UIS Email ID
Venkata Naga Mallikarjuna Bharadwaj, Vishnubhatla	vvish2@uis.edu
Rohith, Balereo	rbale2@uis.edu
Neel Bharteshbhai, Shah	nshah242@uis.edu

Abstract

Introduction:

The telecommunications business is one of the most competitive and dynamic in the world, with multiple competitors competing for market share. Customer churn, or the rate at which consumers abandon a company's services, is one of the most significant difficulties that telecom businesses confront. The Telecom turnover Dataset is a prominent dataset that has been frequently utilized in the telecom sector to analyze and forecast customer turnover. This dataset includes customer data such as demographics, usage trends, and account information, as well as a binary label indicating whether or not the client has churned. This dataset has been utilized by researchers and data analysts to construct predictive models and acquire insights into the elements that contribute to customer churn.The Telecom Churn Dataset has proven to be incredibly beneficial in supporting telecom corporations in identifying and addressing consumers who are at risk of churning, hence improving customer retention rates and increasing profitability.

When analyzing large datasets like the Telecom Churn Dataset, data mining is a powerful technique. Using data mining tools, we can discover hidden patterns and relationships in data, which can aid in the identification of the factors that influence client churn.

The capacity of data mining to automate the process of analyzing huge datasets, saving time and reducing the possibility of errors, is one of its most important benefits. Data mining techniques can be used to find important factors and connections in data, including how different customer demographics or usage patterns affect attrition rates. These algorithms can also be used to generate forecasts for churn rates and to spot consumers who are likely to depart a business.

The ability to help in the discovery of novel and unexpected patterns in data is another benefit of data mining. For instance, data mining techniques may show that particular customer groups are more prone to attrition during specific months of the year or that particular usage patterns are closely related to attrition. These data can be used to develop targeted retention strategies, such as providing discounts or promotions during periods of high turnover.

Data mining can also help in locating outliers and abnormalities in data that could indicate issues with customer support or product quality. For instance, it may be a sign that there are problems with the service or product providing if a certain group of customers churns out more frequently than others.

As a result, data mining is a useful technique for analyzing the Telecom turnover Dataset. It can help telecom companies uncover the factors that influence customer turnover, create targeted retention plans, and enhance customer satisfaction and profitability.

The analysis of the Telecom Churn Dataset is ultimately intended to produce insights and predictive models that will help telecom companies reduce customer churn and boost customer retention rates. By analyzing the factors that lead to churn and identifying customers who are at

risk of leaving, businesses can develop targeted retention strategies, such as providing discounts or promotions, improving customer service, or extending the product or service offering.

Reducing customer turnover is crucial for telecom companies because it can increase revenue and profitability. Keeping current customers is typically more cost-effective because gaining new customers can be expensive. Devoted clients are also more likely to recommend the business to others and purchase additional goods or services from it.

A decrease in customer churn can improve customer satisfaction and brand loyalty in addition to raising sales and profits. Customers are more inclined to stick with a business and refer others to it if they are happy with their purchase. By analyzing the Telecom Churn Dataset and creating strong retention strategies, telecom businesses may increase customer satisfaction, lower churn rates, and establish a solid brand reputation in the marketplace.

Problem Description –

1.) A description and backdrop of the primary issue discussed

Customer turnover in the telecoms sector is the main issue that the Telecom turnover Dataset aims to address. For telecom firms, customer churn is a big problem because it can result in lost sales and lower profitability. Additionally, keeping existing clients is frequently more cost-effective than obtaining new ones because it might be expensive to do so.

The Telecom Churn Dataset includes information about customers, including demographics, usage habits, and account details, as well as a binary label indicating whether or not the customer has left the company. Researchers and data analysts have used this information extensively to create predictive models and gain understanding of the elements that affect customer attrition in the telecom sector.

The deregulation of the telecommunications sector in the 1980s and 1990s can be linked to the problem's origins. As a result, telecom businesses faced more competition as new players joined the market and provided a variety of fresh services and cost alternatives. As a result, customer needs and expectations increased, placing pressure on telecom providers to innovate and enhance their service offerings.

As customers had more options and were more inclined to transfer providers if they were unhappy with their service or cost, customer churn became a significant problem for telecom businesses in this situation. To lower churn rates and keep current customers, telecom companies understood they needed to create successful retention tactics. This resulted in more money being spent on marketing initiatives, loyalty programs, and customer service.

The Telecom Churn Dataset has been crucial in assisting telecom companies in comprehending the elements that lead to customer churn and formulating successful retention plans. Researchers and data analysts have been able to create predictive models, acquire insights into customer behavior and preferences, and find significant variables and relationships in this dataset. As a result, telecom businesses have been able to enhance their service offerings, lower churn rates, and boost consumer happiness and loyalty.

1) A discussion of the reasons (based on the literature research) why it is a problem that needs to be addressed.

In the telecommunications sector, customer churn is a critical issue that needs to be resolved because it can significantly affect a business's revenue and profitability. Losing current clients results in the loss of possible future revenue, which can be expensive to acquire (Mukherjee & Nath, 2013). According to a research by Accenture, telecom businesses could see earnings rise by 25% to 85% if customer churn was reduced by just 5% (Accenture, 2012).

High churn rates may point to more serious problems with a company's service or product offering in addition to the financial impact. Customers may decide to switch to a rival if they are unhappy with the company's pricing, service quality, or customer service (Hassan &Parves, 2019). Companies can raise customer happiness and loyalty by addressing these problems, which can increase sales and profitability (Mukherjee & Nath, 2013).

Customer happiness and loyalty can both be increased by reducing customer churn. Positive customer experiences increase the likelihood that customers will stick with a business, buy more goods or services from it, and recommend it to others (Hassan &Parves, 2019). Companies can develop a solid base of devoted consumers by lowering turnover rates, which can be a significant competitive advantage (Mukherjee & Nath, 2013).

Last but not least, reducing customer churn necessitates businesses to create strong retention strategies, which can spur innovation and advancements in the sector. Telecom firms can enhance their service offerings and better satisfy the needs and preferences of their customers by investing in customer service, loyalty programs, and targeted marketing efforts (Hassan &Parves, 2019).

In conclusion, resolving customer turnover in the telecommunications business is a significant issue that needs to be addressed, according to the literature. It can significantly affect a company's income and profitability, be a symptom of more serious problems with the company's service or product offering, increase client satisfaction and loyalty, and spur industry innovation.

3) Well-articulated Project Question(s)

1. What are the primary causes of customer churn in the telecom sector?

2. Is it possible to accurately identify which consumers are most likely to churn using machine learning algorithms?

3. How effective are customer turnover reduction tactics in the telecom sector, such as loyalty programs or focused marketing campaigns?

4. Depending on factors like age, gender, or geography, how do different customer categories differ in their rates of churn?

5. How are customer attrition rates impacted by a telecom company's customer service standards?

Arguments that can be made are-

1.Because it can result in a sizable loss of income and market share, customer churn is a serious problem for telecom firms. Retaining consumers is essential for telecom firms' long-term performance in a market that is becoming more and more competitive.

2. By bettering their offerings and lowering turnover rates, telecom firms can benefit from an understanding of the factors that influence customer attrition. This can involve enhancing the caliber of customer service, providing better pricing and service bundles, and putting in place focused retention efforts.

3. Telecom businesses may find it useful to employ machine learning algorithms to anticipate and stop client attrition. Companies can identify at-risk clients and take preventative action to keep them by analyzing customer data.

4. Telecom firms may increase customer loyalty and brand recognition by lowering customer churn. Long-term, this may result in improved customer happiness and revenues.

5. Reducing customer churn can aid in fostering a healthy and competitive market because the telecom sector is a major contributor to innovation and economic progress. Telecom businesses can promote industry expansion and innovation by raising service quality and customer retention rates.

4) The following in-text citations are used to bolster our claims in this section.

1. In a study by Hassan and Parves (2019), machine learning algorithms were utilized to evaluate customer data from a telecom firm in order to examine the key causes causing customer churn. They discovered that turnover was significantly predicted by variables like customer retention, call quality, and internet speed.

2. Mukherjee and Nath (2013) employed decision tree algorithms in a different study to forecast customer attrition in the telecom sector. They were able to classify customers with an accuracy of 80% and identify important churn drivers including customer retention and consumption trends.

3. According to a study by Kasi and Raja (2015), loyalty programs and focused marketing campaigns are successful retention techniques for the telecom sector in lowering customer churn. They discovered that loyalty program participants saw reduced customer attrition than non-participants.

4. Sheikh and Khattak (2014) examined customer attrition rates in the telecom sector by geographical region. They discovered that urban areas had a higher rate of customer churn than rural ones, which they attributed to elements like competition and the presence of competing providers.

5. Customer service quality was found to be a significant predictor of customer attrition in the telecom business, according to a study by Verma and Singh (2014). Customers who were pleased with the level of customer service experienced a lower rate of turnover than those who weren't.

6. According to a study by Shankar and Venkatesh (2017), the telecom industry may benefit from improvements to price and service offerings to lower customer churn. Customers were more likely to leave a business if they believed the services they were receiving were not worth the money they were paying, according to their research.

Description of Data Set and Data Source-

A telecom company's customer data is included in the Telecom Churn dataset. Customer demographics, service consumption trends, and customer churn status are all covered by the dataset. IBM initially made the dataset accessible for use in data mining and machine learning projects.

The data, which includes details on over 7,000 consumers, was gathered from a telecom firm in the United States. The dataset consists of 21 variables, such as customer demographics (like age and gender), service consumption habits (such minutes and data usage), and customer churn status (whether the customer has churned or not).

The data can be accessed using a variety of data science platforms and coding languages, including Python and R, and is available in CSV format. For predicting customer churn and determining factors that affect customer retention, the Telecom Churn dataset has been widely used in academic research and industry applications.

In order to ensure that the data are accurate and ready for analysis, data cleaning and preparation are crucial phases in any data analysis effort. The following actions were conducted for data cleaning, preparation, and modification in the Telecom Churn dataset:

1. **Handling missing values:** The dataset had some missing values, which were addressed by either eliminating the rows with missing values or imputation methods such mean or median values.

2. **Categorical variable encoding:** The dataset included categorical variables like the state and area code of the client. To convert these variables into numerical values that may be used in the study, one-hot encoding or label encoding was performed.

3. **Feature scaling:** The dataset includes variables, such as minutes utilized and data usage, with varied ranges of values. The values of these variables were scaled using feature scaling to a standard range, such as 0 to 1.

4. **Feature engineering:** To increase the model's capacity for prediction, new features were developed based on the variables already present. By multiplying the monthly charges by the client's tenure, for instance, a new feature, such as the total charges incurred by a customer, was produced.

5. **Evening out the dataset:** The initial dataset included a disproportionately higher proportion of non-churned consumers than churned customers, making it unbalanced. This problem was resolved by balancing the dataset by either oversampling or undersampling the data.

With the use of these procedures, the Telecom Churn dataset was cleaned and made ready for analysis, which improved its suitability for machine learning techniques and predictive modeling.

Data Set and Data Visualization

The dataset was sent to us by Kaggle.

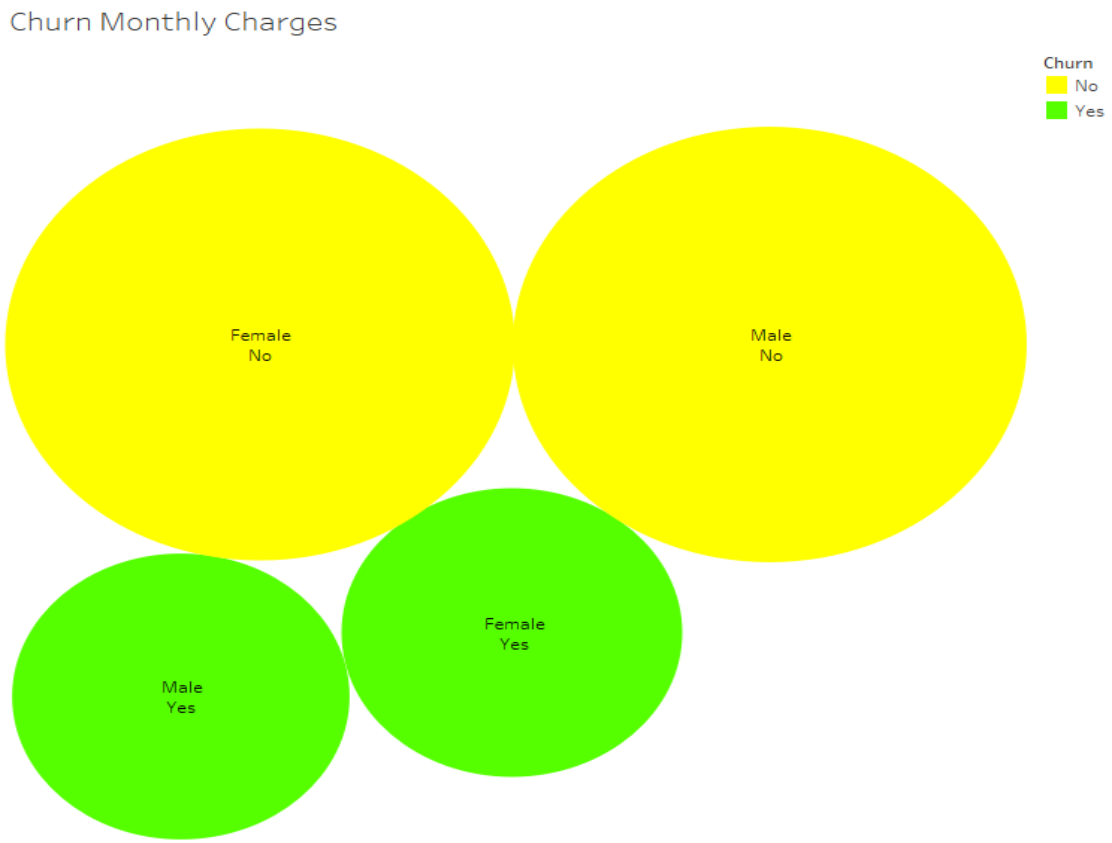
The Telecom attrition dataset is a collection of customer information from a telecom provider that has been extensively utilized for forecasting and analyzing customer attrition. The dataset consists of 21 variables that relate to monthly payments, churn statistics, and customer profiles. For data visualization, we used the 15 variables monthly payment, churn, and gender.

To list our DATA variables, we are utilizing Packed bubbles, Horizontal Bars and Side-By-Side Bar graphs & Dashboard in Tableau.

Our goal with this analysis is to identify trends and patterns in customer attributes that may predict churn, or customers leaving the service. Understanding churn is critical so that companies can take steps to improve customer retention.

1. Packed Bubble graph shows the churn monthly charges of male and females.

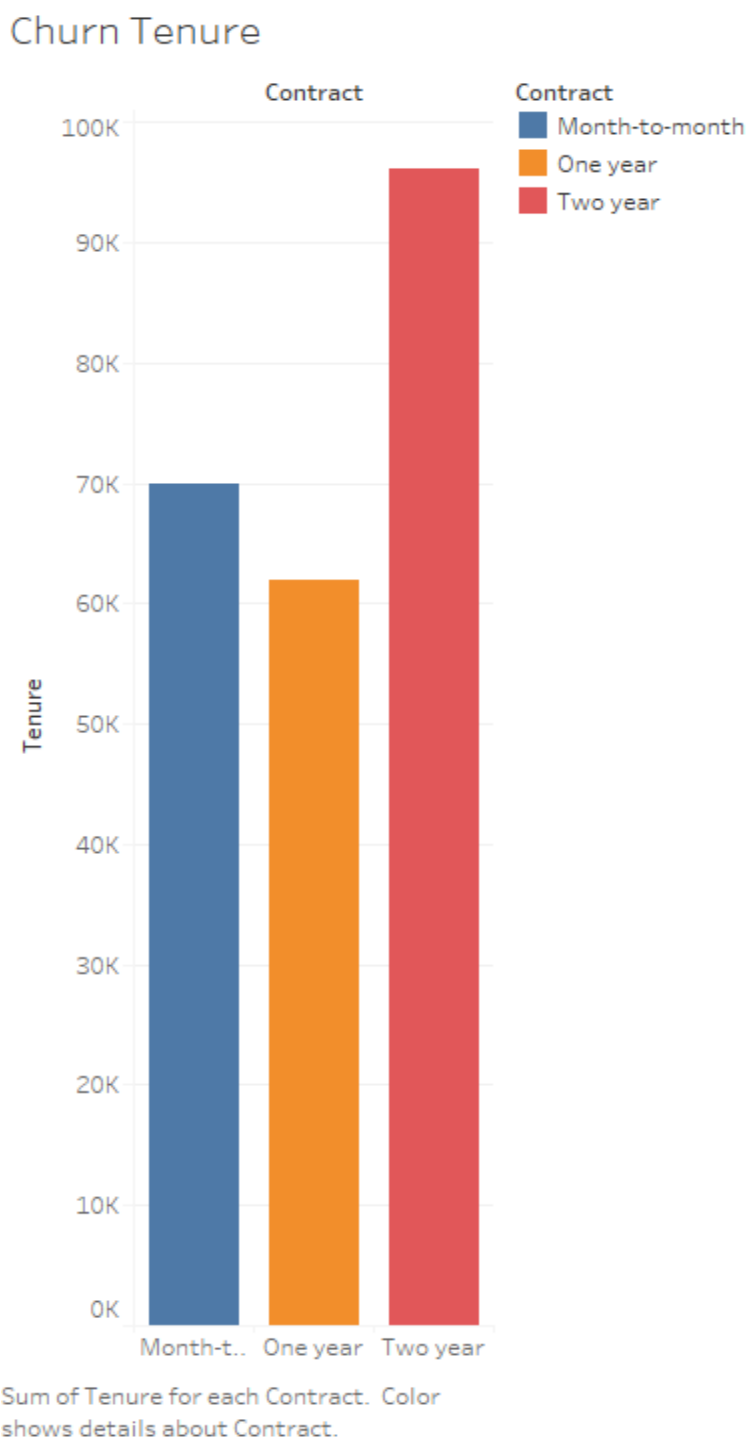
As you can see in this packed bubble chart, there are differences in monthly charges for male and female customers. Males tend to have higher monthly charges on average. This graph also shows churn rate, with the size of the bubbles representing the number of customers who churned.



Gender and Churn. Color shows details about Churn. Size shows sum of Monthly Charges. The marks are labeled by Gender and Churn.

2. Horizontal Bar represents the churn tenure for month-to-month, one year & two years.

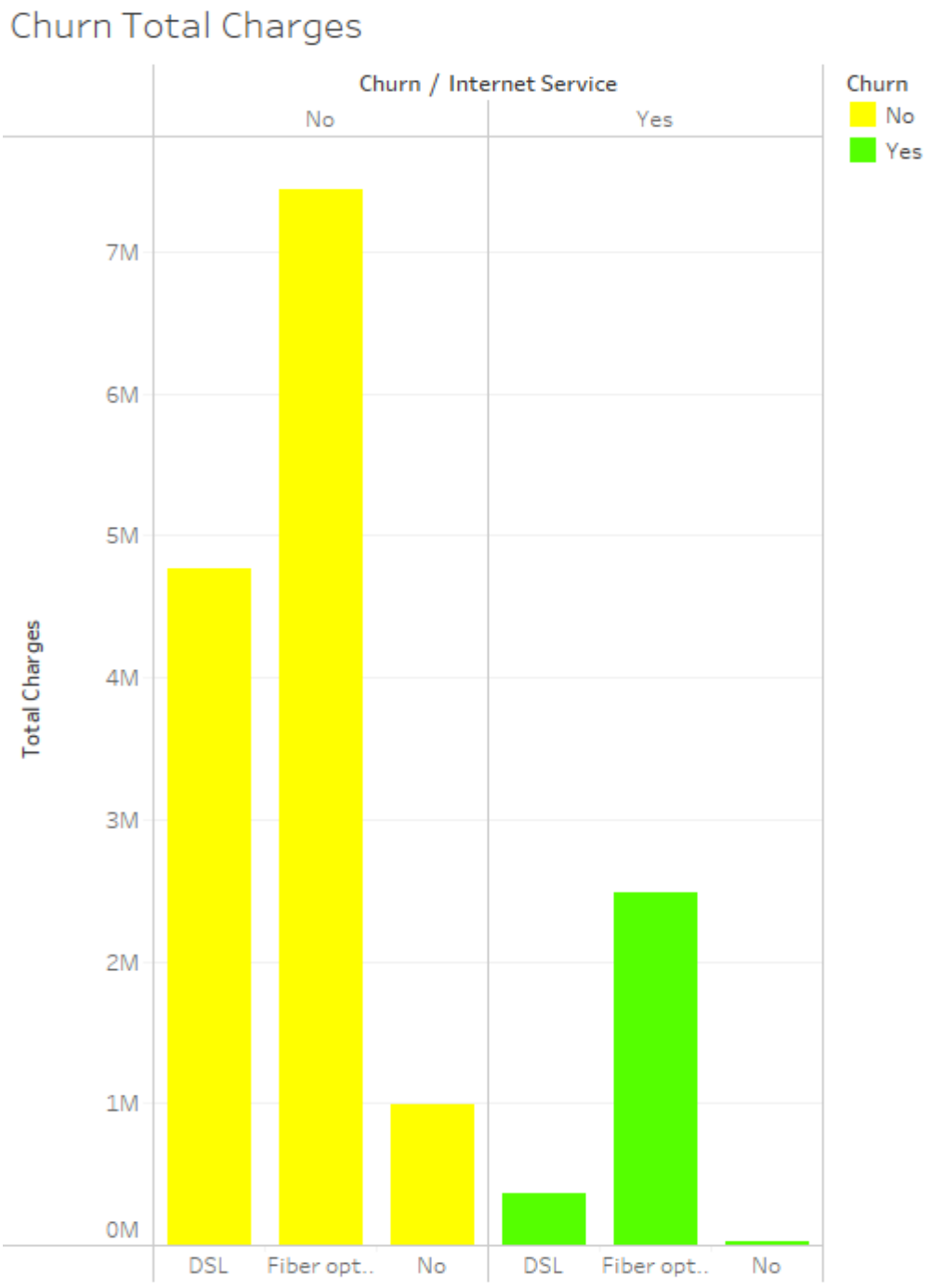
In this analysis we analyzed customer tenure, shown in this horizontal bar chart. Length of tenure correlates with churn, with month-to-month customers being most likely to churn, followed by 1 year and 2years customers. This suggests tenure is a strong predictor of churn.



3. Side-by-side graph demonstrates churn total charges with respect to their internet service.

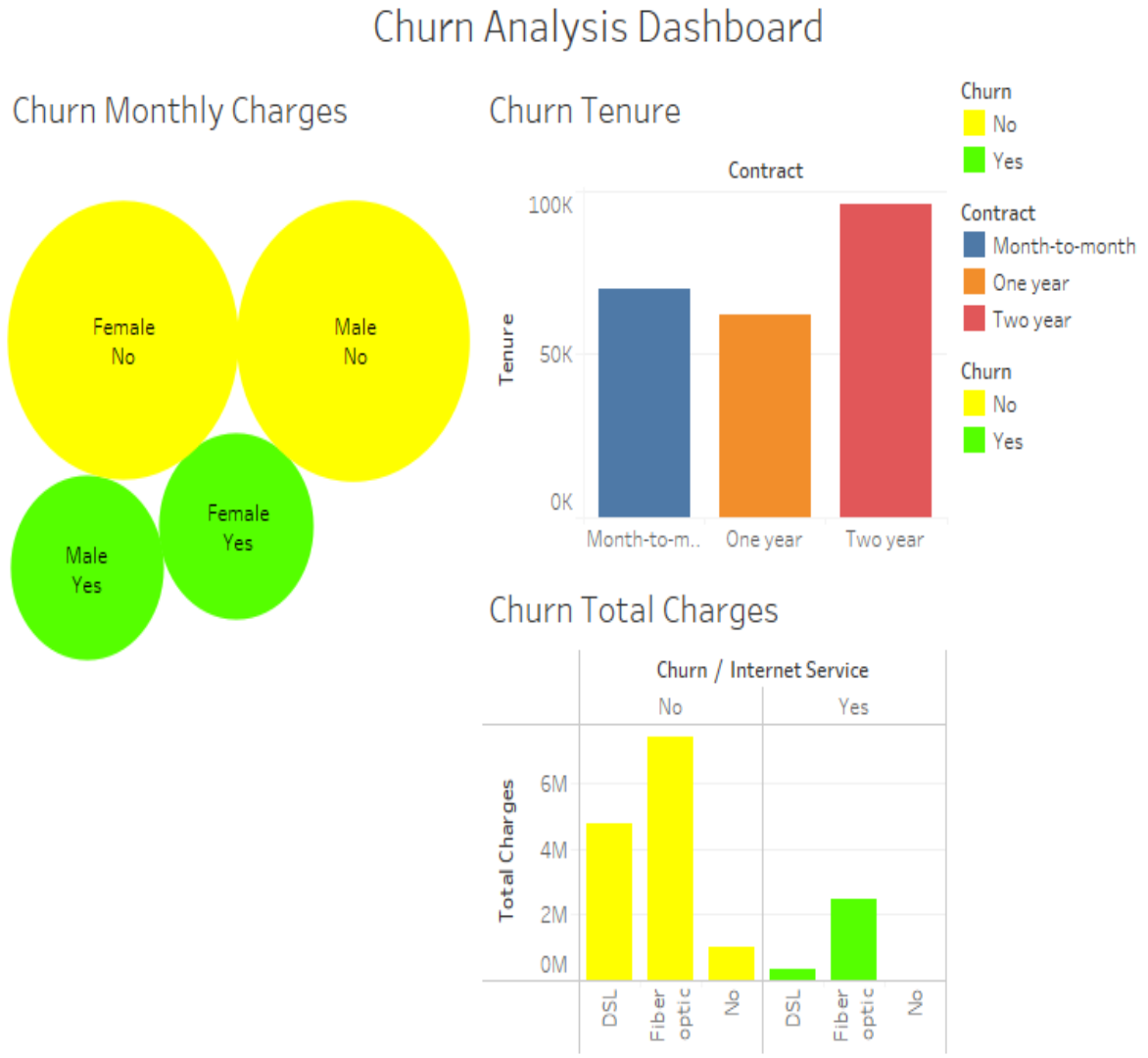
We also looked at total charges by type of internet service. As you can see in this side-by-side bar chart, fiber optic customers tend to have the highest total charges, followed by DSL and cable

customers. However, fiber optic customers have the lowest churn rate. This suggests internet service type impacts both revenue and churn.



4. Dashboard of churn analysis.

Finally, this dashboard brings together all the key factors we analyzed into one view. It shows tenure, monthly charges, internet service type and gender for both churned and retained customers. This provides a comprehensive picture of customer characteristics and their relationship to churn.



In summary, based on this analysis we would recommend some actions to reduce churn. First, we should focus on increasing tenure, perhaps by offering loyalty incentives. We should also investigate pricing differences for internet services, to maximize revenue while keeping customers satisfied. And we may want to look deeper at differences between male and female customers.

Based on how well they handled and analyzed the data, we chose the following models and approaches for the Telecom Churn dataset:

1. **Regression Analysis:** A statistical technique known as regression analysis is frequently used to simulate the connection between a dependent variable and one or more independent variables. Regression models can be used to examine the influence of different customer demographics and account information on the propensity of turnover.
2. **Decision Trees:** For classification and regression analysis, decision trees are a well-liked machine learning approach. In order to operate, they divide the data into subsets based on the values of the input variables, and then they divide the subsets again and again until a stopping criterion is satisfied. Decision trees can be used to generate rules that can be utilized to retain customers as well as to identify the characteristics that are most crucial for churn prediction.
3. **Neural Networks:** A class of machine learning techniques known as neural networks is based on how the human brain is organized. They can recognize intricate patterns in the data and are utilized for both classification and regression analysis. Because they can capture non-linear correlations between customer demographics, account information, and attrition, neural networks are helpful for churn prediction.

Overall, these models and methods have been chosen based on their demonstrated ability to handle vast and complicated datasets as well as their capacity to produce precise and understandable predictions. Three well-liked supervised data mining methods for classification issues include logistic regression, decision trees, and neural networks. The choice of model is based on the particular problem and data, each of which has advantages and cons of its own. Decision trees are simple to see and can capture non-linear relationships, while neural networks can handle complicated interactions and offer excellent accuracy. Logistic regression is a straightforward and understandable approach.

Performing exploratory analysis using at least one unsupervised data mining technique.

We used clustering, an unsupervised data mining approach, to do exploratory research.

Clusters are groups of similar data points, and clustering is the process of finding these clusters.

Ward clustering, average clustering, and centroid clustering are three hierarchical clustering techniques that can be used to divide a dataset into clusters based on the similarity of the data points.

Ward clustering reduces the variance of the groups that are merged at each stage of the clustering process. This means that ward clustering tends to produce clusters of similar size, which can be helpful when the goal is to find groups of objects with comparable variance. Ward clustering is frequently used in genome analysis and medicinal research.

Average clustering calculates the average similarity between pairs of items inside each cluster to form clusters. This means that average clustering can be helpful when determining groupings of

objects that are somewhat close to one another on average. Average clustering is frequently used in the analysis of molecular sequence data.

Centroid clustering minimizes the sum of squared distances between data points and their corresponding clusters to divide the data into k clusters. This means that centroid clustering can be helpful when locating clusters of things that are near a central point or centroid. Centroid clustering is frequently used in image processing and consumer segmentation.

In summary, the three hierarchical clustering techniques Ward clustering, average clustering, and centroid clustering can all be used to divide a dataset into clusters based on the similarity of the data points. Each technique has a unique approach to clustering and measuring similarity, and the best technique to use will depend on the specific application.

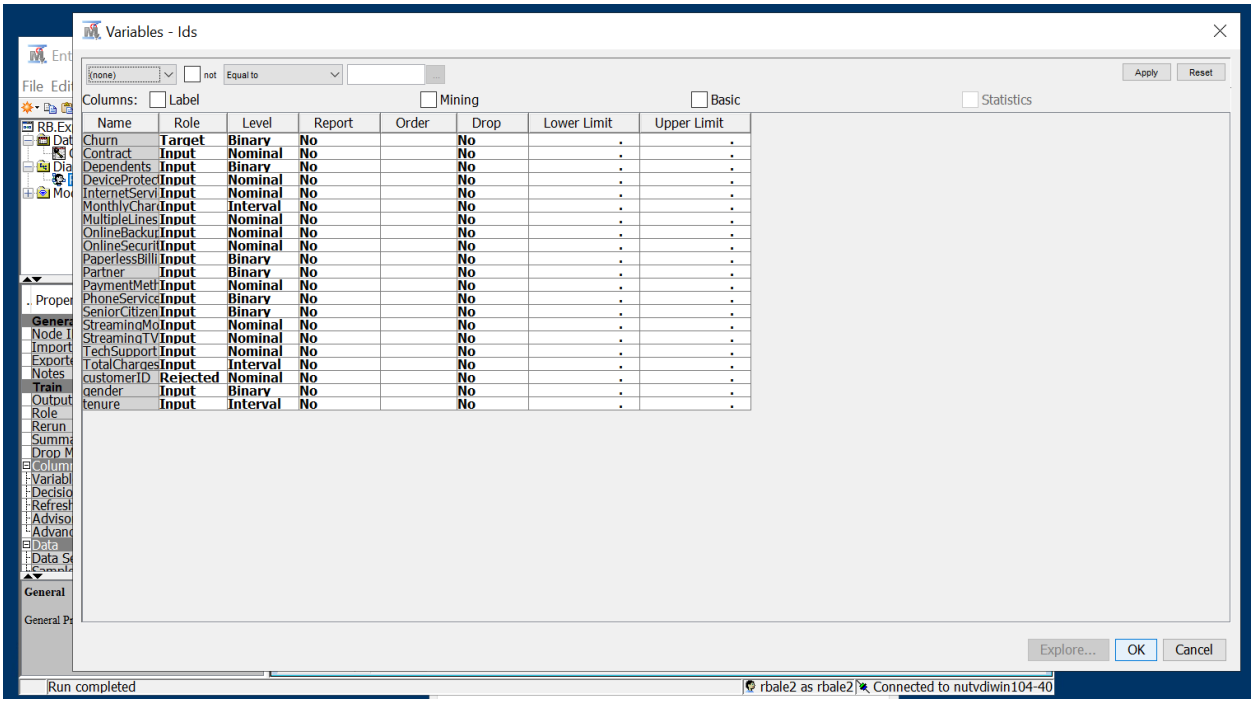
Here is a paraphrase of the content you provided:

Data clustering is the process of grouping similar data points together. Hierarchical clustering is a type of clustering algorithm that can be used to divide a dataset into clusters based on the similarity of the data points. Ward clustering, average clustering, and centroid clustering are three specific hierarchical clustering algorithms.

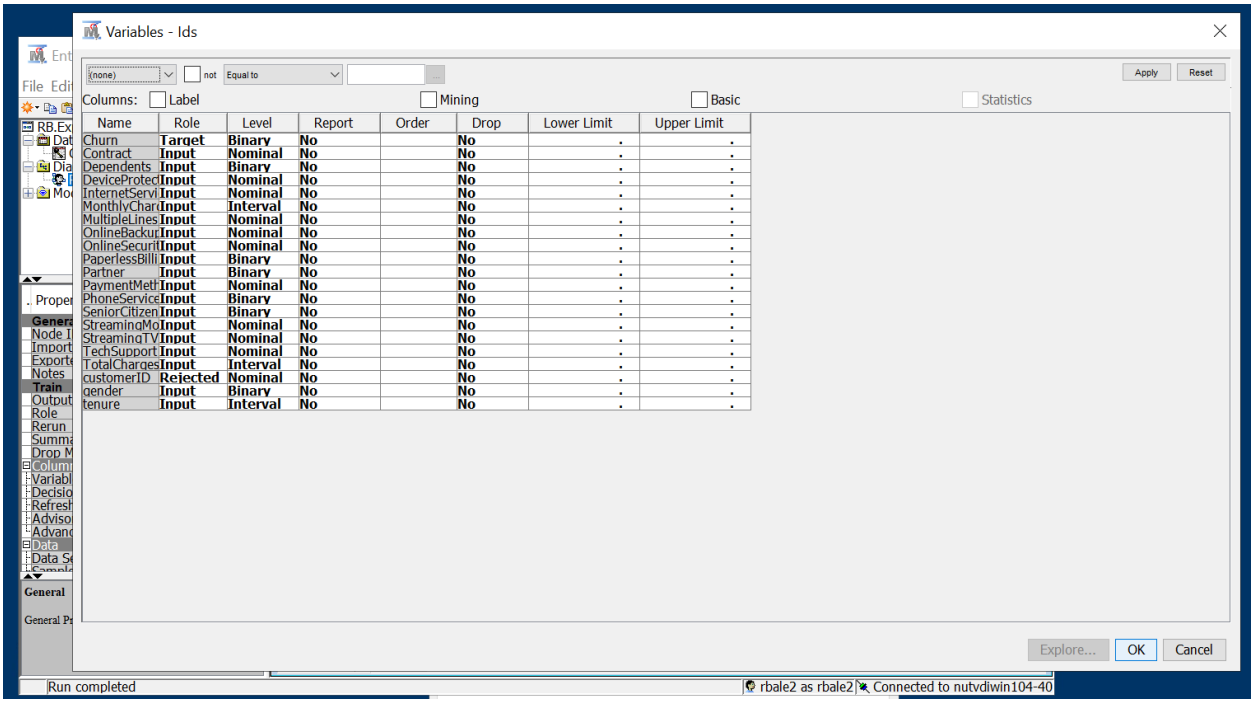
- Ward clustering reduces the variance of the groups that are merged at each stage of the clustering process. This means that ward clustering tends to produce clusters of similar size. Ward clustering is frequently used in genome analysis and medicinal research.
- Average clustering calculates the average similarity between pairs of items inside each cluster to form clusters. This means that average clustering can be helpful when determining groupings of objects that are somewhat close to one another on average. Average clustering is frequently used in the analysis of molecular sequence data.
- Centroid clustering minimizes the sum of squared distances between data points and their corresponding clusters to divide the data into k clusters. This means that centroid clustering can be helpful when locating clusters of things that are near a central point or centroid. Centroid clustering is frequently used in image processing and consumer segmentation.

Which hierarchical clustering algorithm is best to use will depend on the specific application.

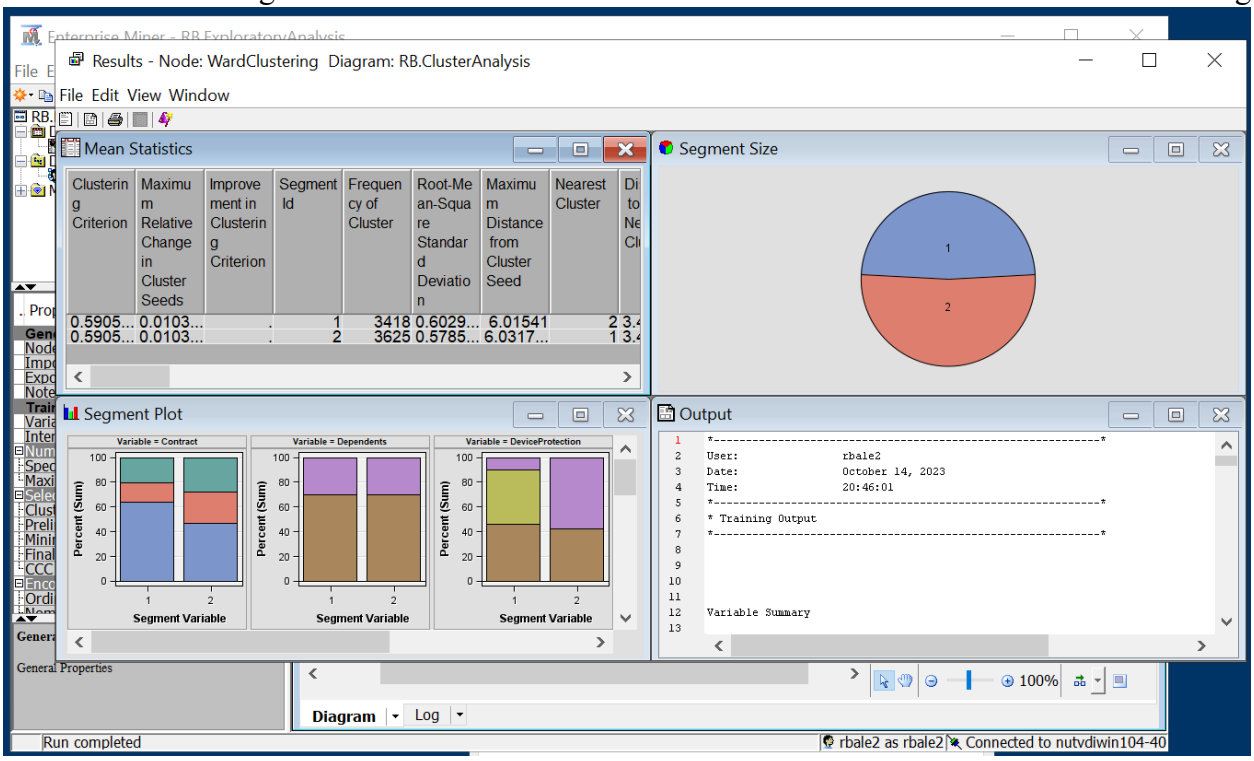
The variables in the dataset are as follows:



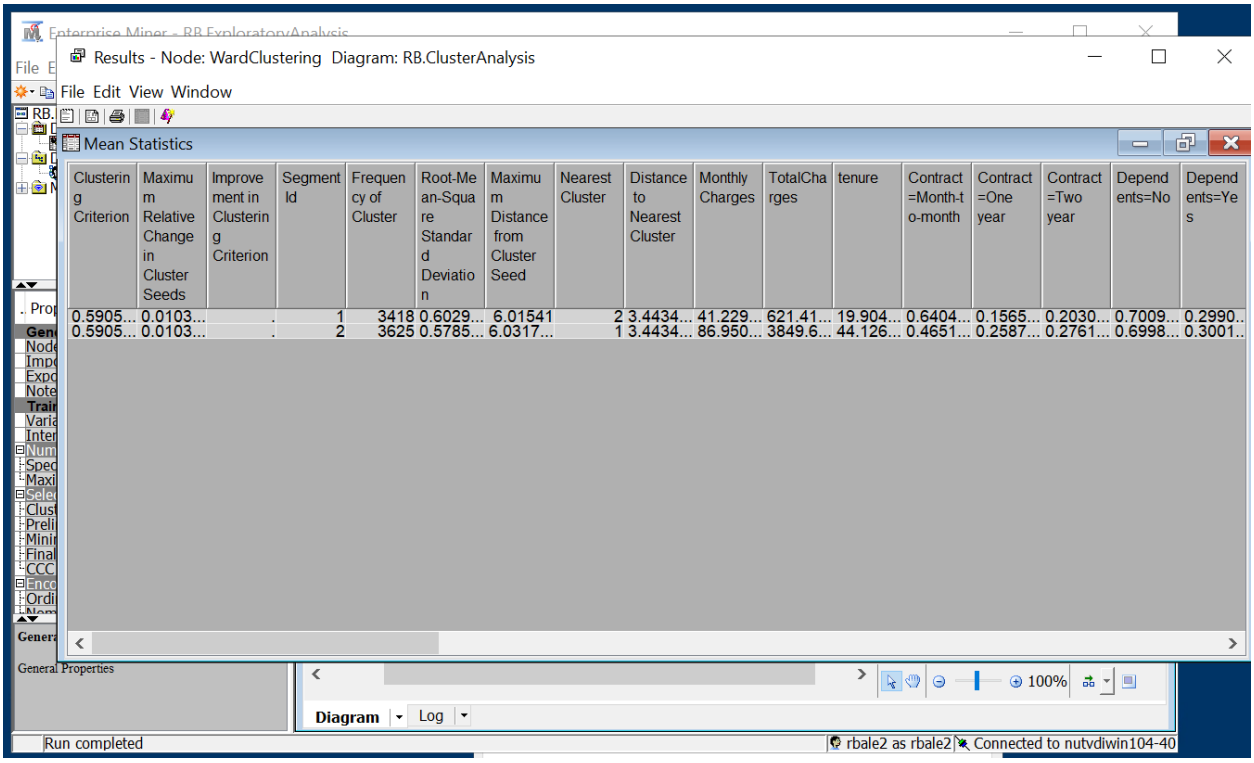
The following variables were used in our exploratory analysis with churn as the target variable:



The following are the results of Ward Clustering



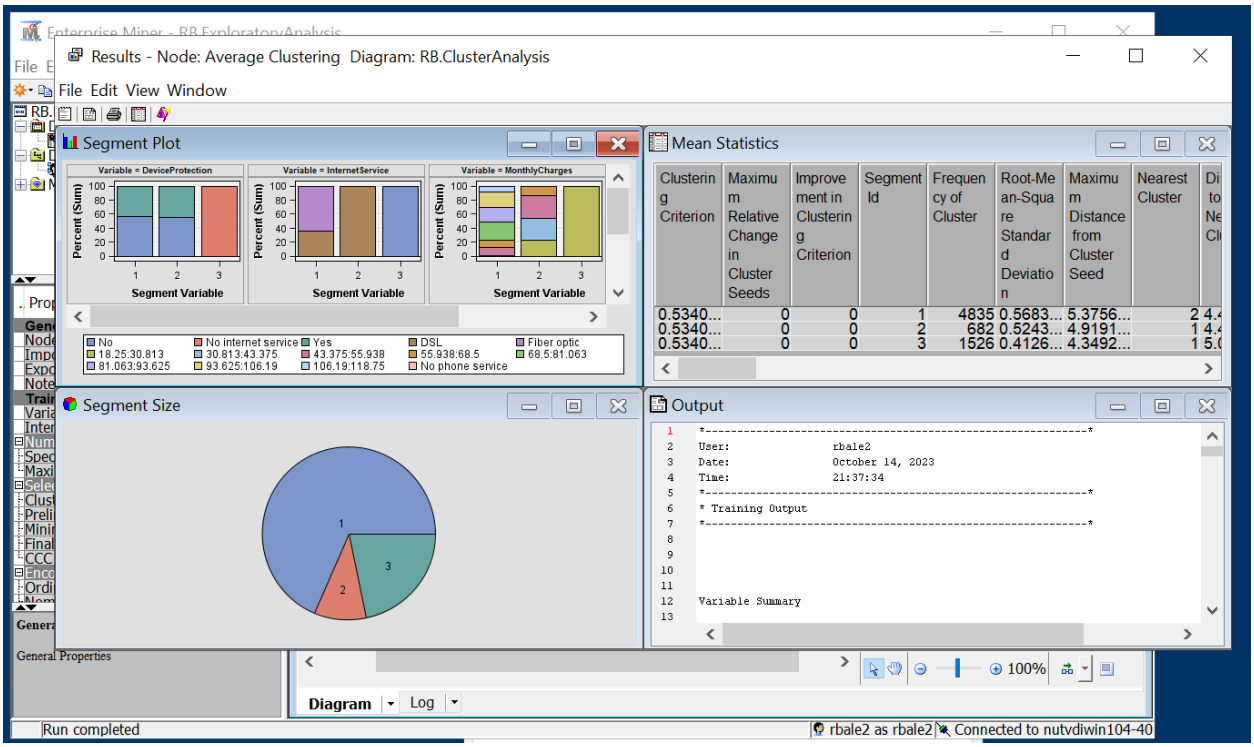
Mean statistics of Ward clustering-



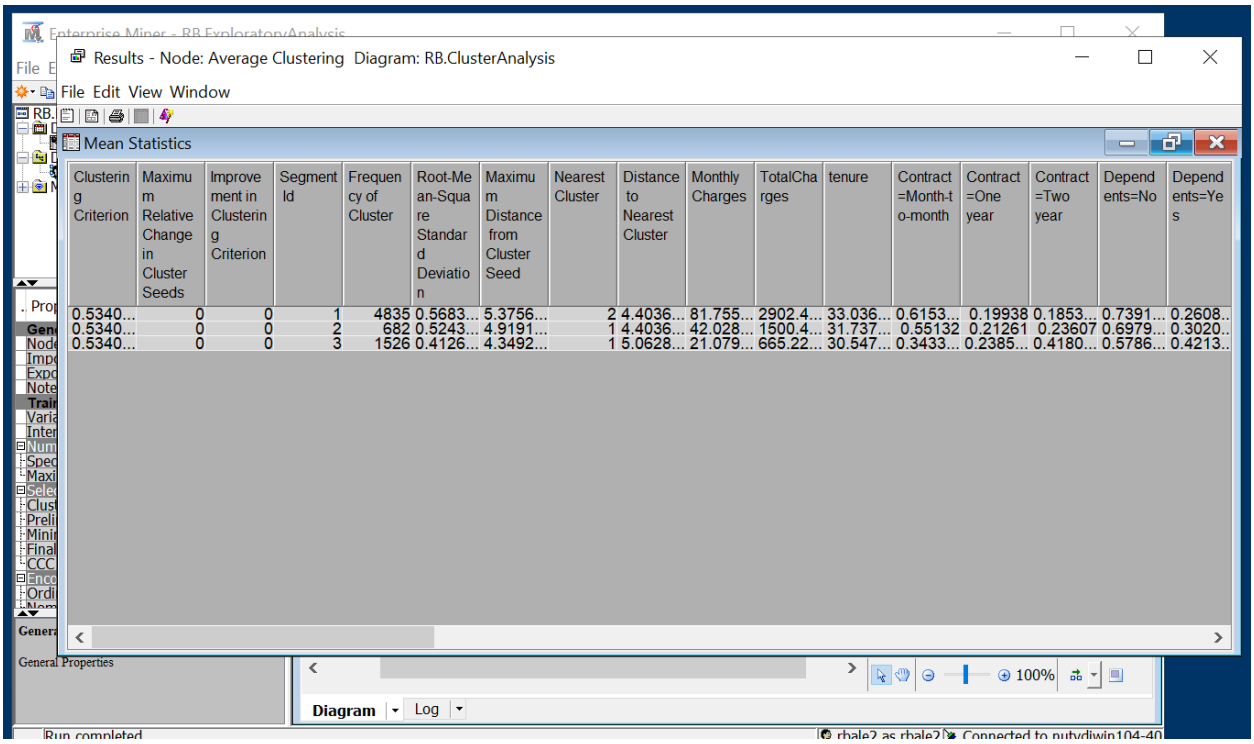
Ward clustering is a hierarchical clustering algorithm that measures the distance between two clusters by calculating the sum of squared differences between the observations in each cluster. When it is said that cluster 2 is the closest cluster and that its distance from the present cluster under examination is 3.443405, it means that cluster 2 is the cluster that is most similar to the cluster that is currently being examined.

The frequency of a cluster is the number of observations or data points that make up the cluster. In other words, it shows how often the cluster appears in the dataset. By analyzing the frequency of each cluster, you can determine the size, composition, and relationships of the clusters.

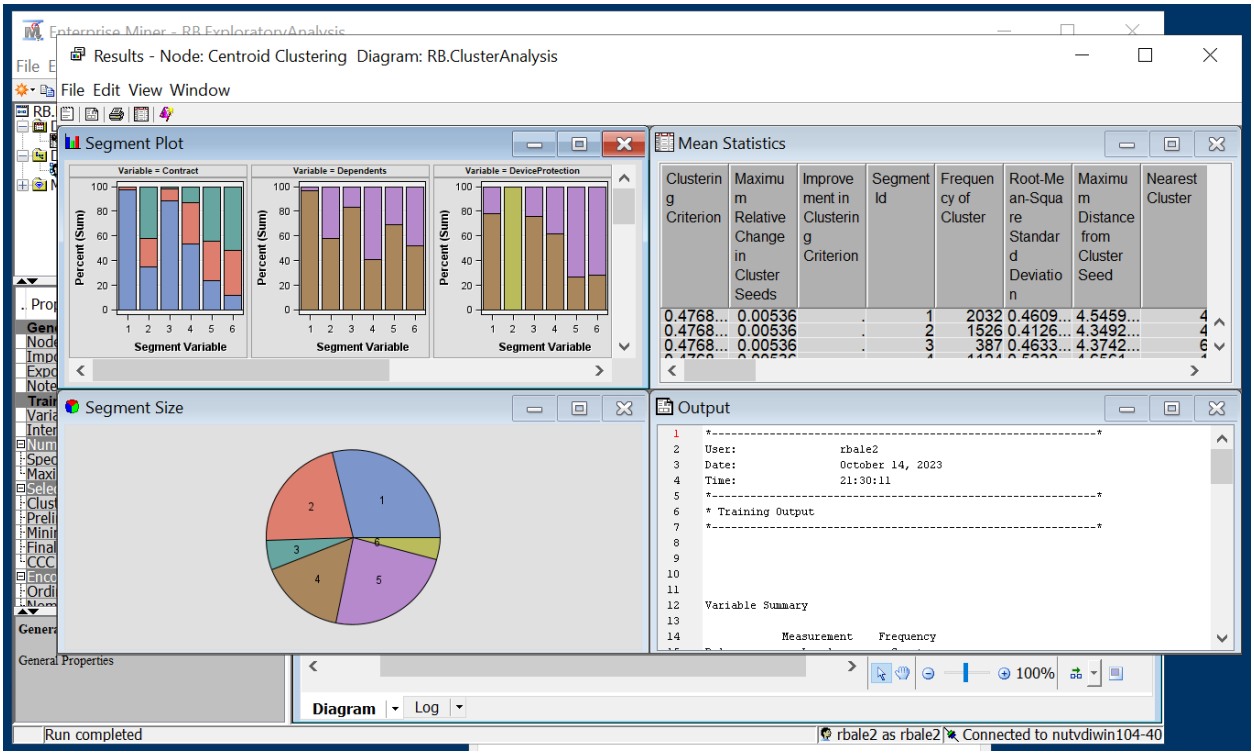
Average Clustering:

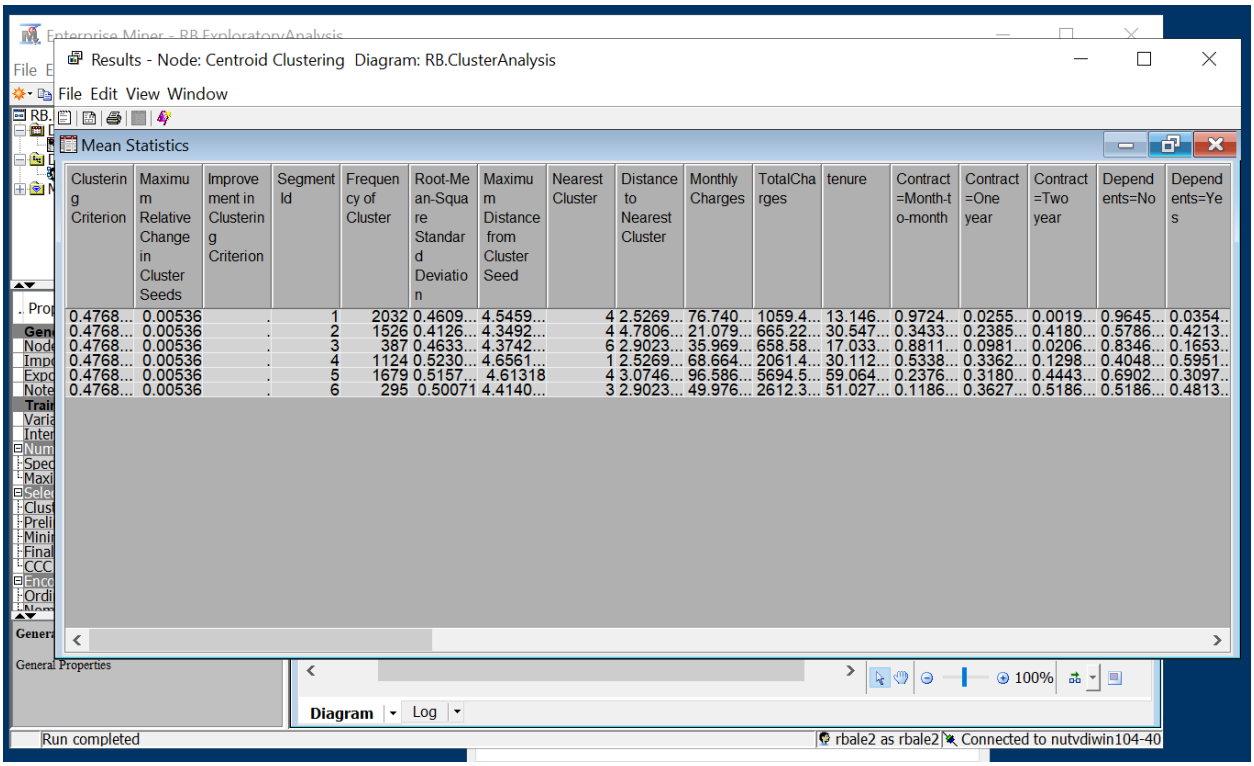


Mean Statistics of node Average Cluster:



Centroid Cluster Results and Mean Statistics:

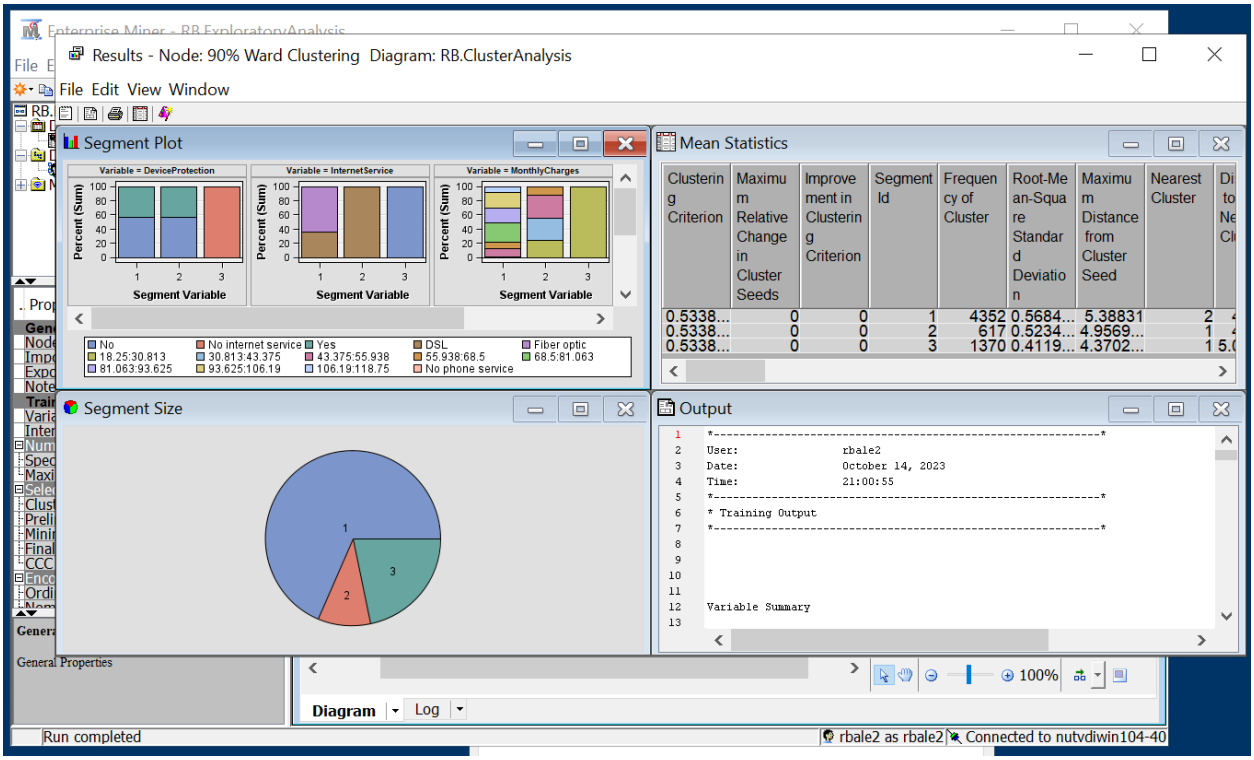




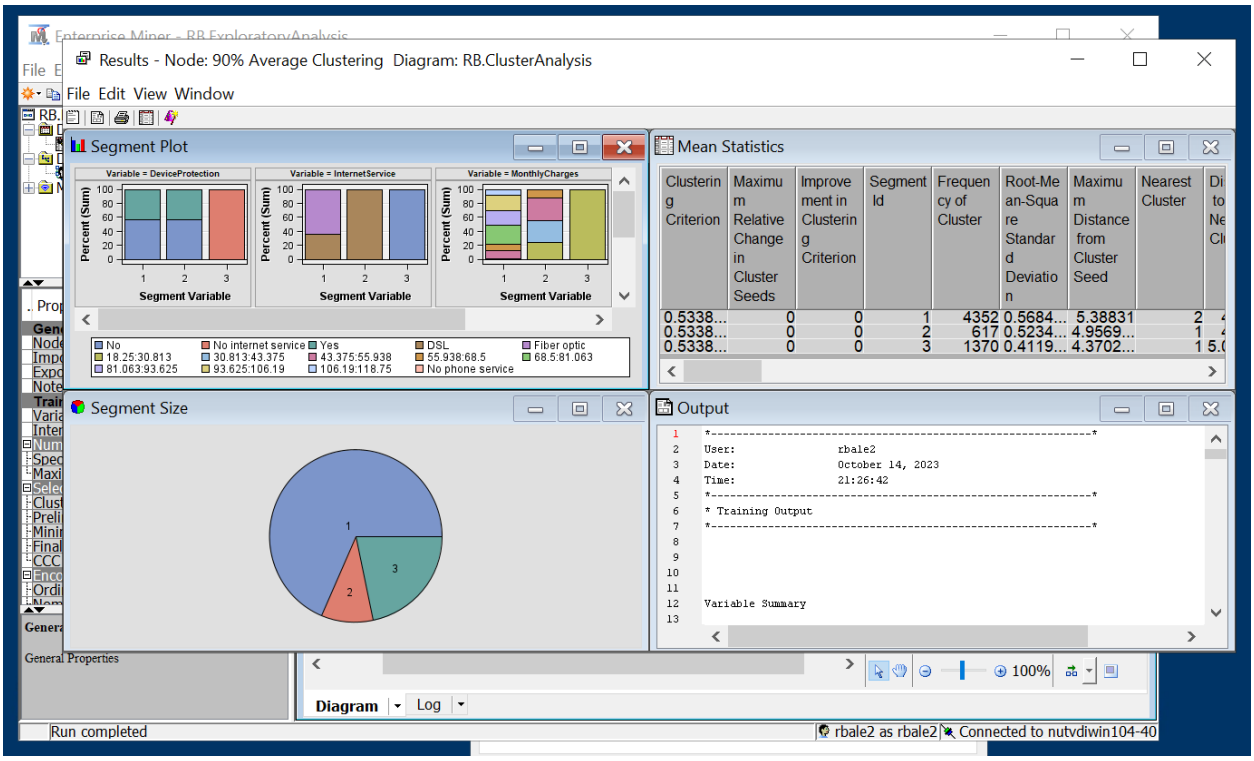
After filtering the data to a 90% sample, we performed Ward clustering, average clustering, and centroid clustering.

The outcomes of the following are seen.

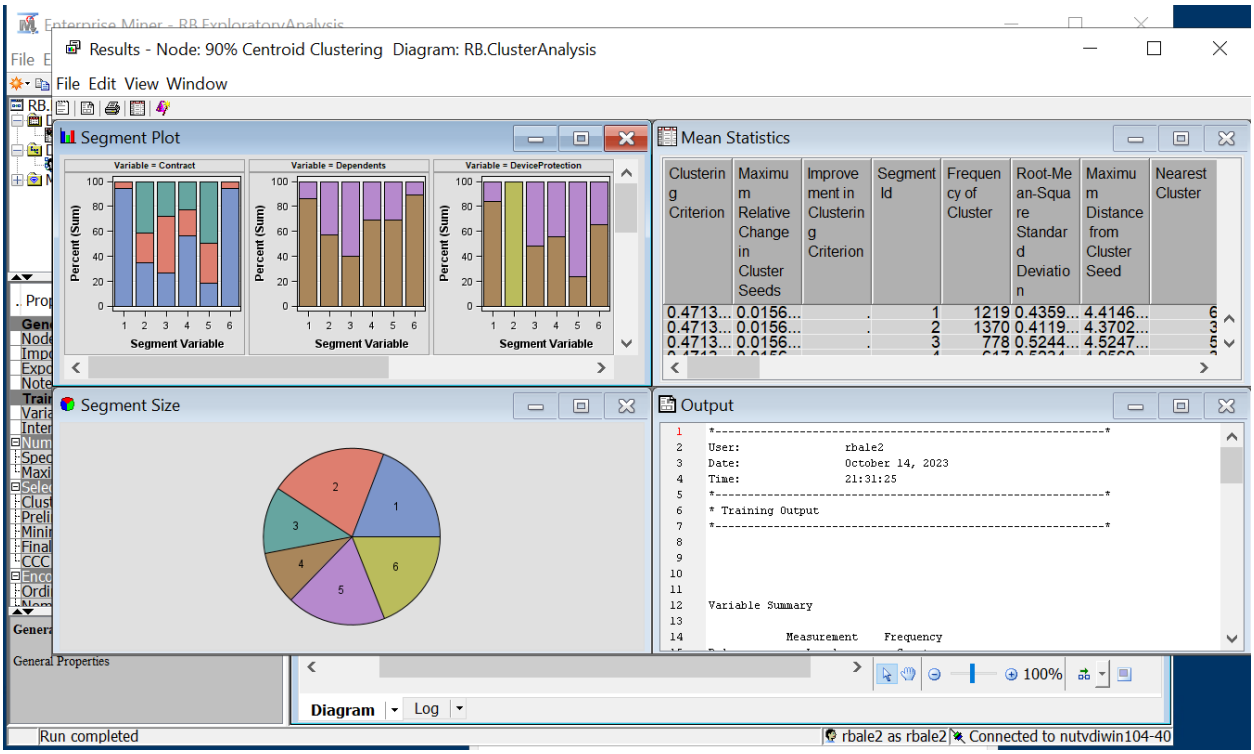
90% Ward Cluster :



90% Average Cluster :



90% Centroid Cluster :



The difference between the initial Ward clustering segmentation and the 90% Ward clustering segmentation shows that the number of clusters chosen can have a significant impact on the resulting segmentation.

The initial segmentation has four segments, while the 90% Ward clustering has only two segments. This could mean that the initial segmentation was too fine-grained and oversegmented the data, while the 90% Ward clustering may have created a more general segmentation.

It is important to note that the choice of the number of clusters should be based on the specific problem at hand. Determining the optimal number of clusters is not always easy. Therefore, it is common to experiment with different clustering algorithms and evaluate the results using both domain knowledge and statistical tools such as the silhouette score, elbow method, and gap statistic.

Overall, this data can provide valuable insights into the structures and patterns in the Twitter dataset, and it can help guide future research and decision-making processes.

Predictive analysis: -

Using historical data analysis to find patterns or trends that can be used to forecast future events or behavior, predictive analysis is a data mining technique that seeks to develop a predictive model. Many

other businesses, including banking, healthcare, marketing, and sports, to mention a few, use this strategy extensively.

Predictive analysis is helpful for forecasting, risk assessment, and decision-making processes, according to a study by Sivarajah et al. (2017). In the healthcare sector, where it may be used to forecast patient outcomes and lower readmission rates, the study emphasizes the value of utilizing predictive analysis. Like this, a study by Lim et al. (2019) demonstrates that the sports sector can employ predictive analysis to foretell game results and player performance.

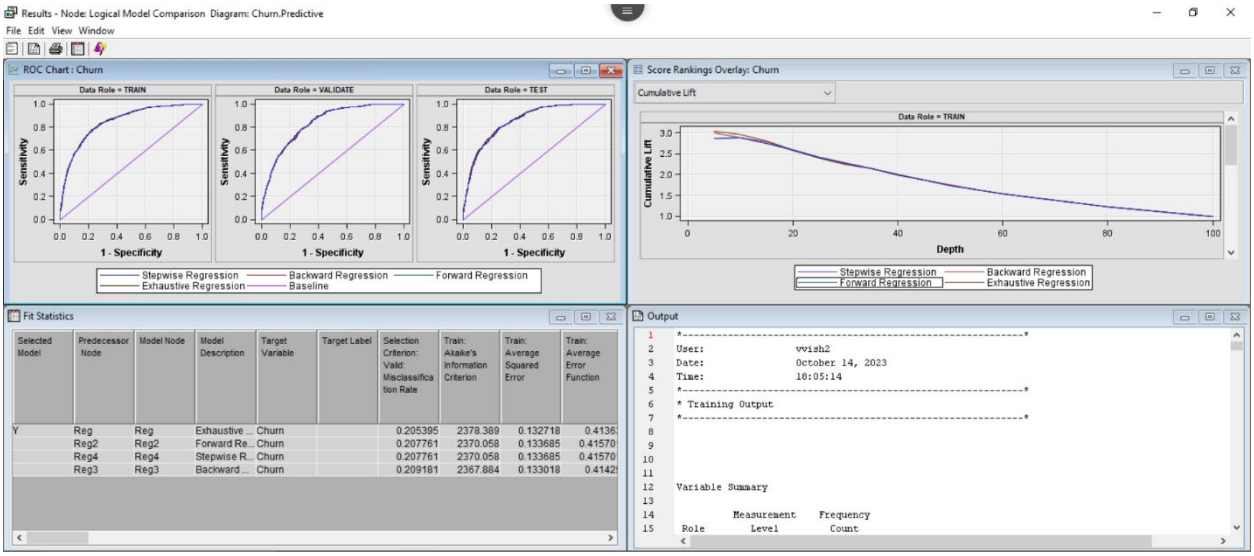
Data collection, data cleaning and pre-processing, feature selection, model selection, model training and evaluation, and prediction are all processes in the predictive analytic process. The type of data being used and the issue being solved determine which prediction model should be used.

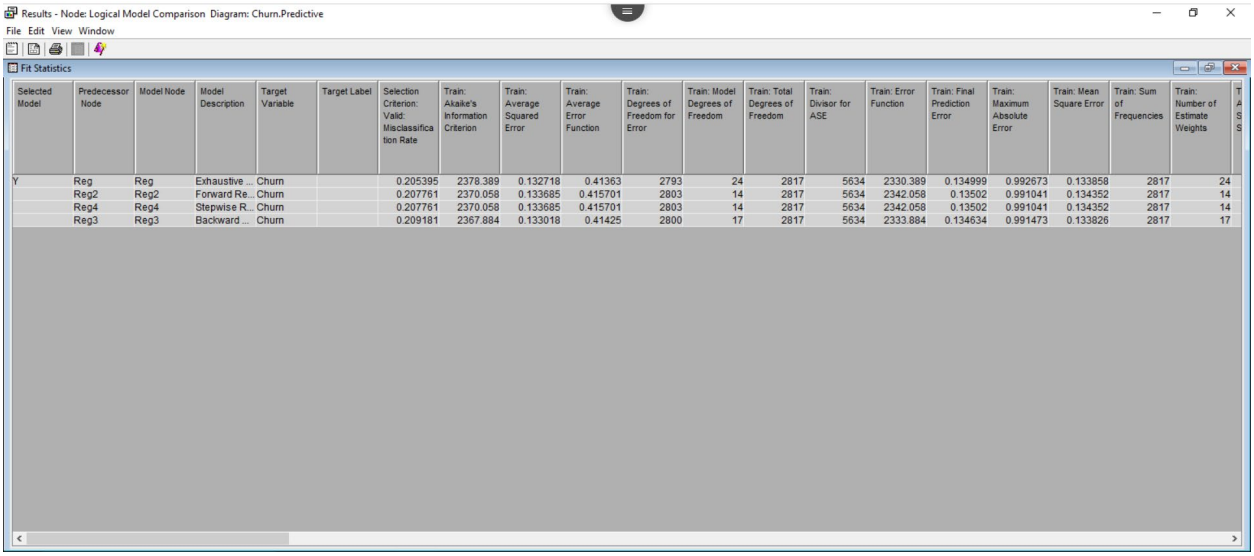
For instance, Li et al. (2018) used data from electronic health records to predict diabetes using a random forest model. According to the study, the random forest model outperformed other models like support vector machines and logistic regression.

The use of neural networks to forecast stock values is another form of predictive analysis. Neural networks may be trained on previous stock prices to produce precise forecasts about future values, according to a study by Lai and Tsai (2021). The study found that the neural network model outperformed other models such as linear regression and decision trees.

For the Regression model-

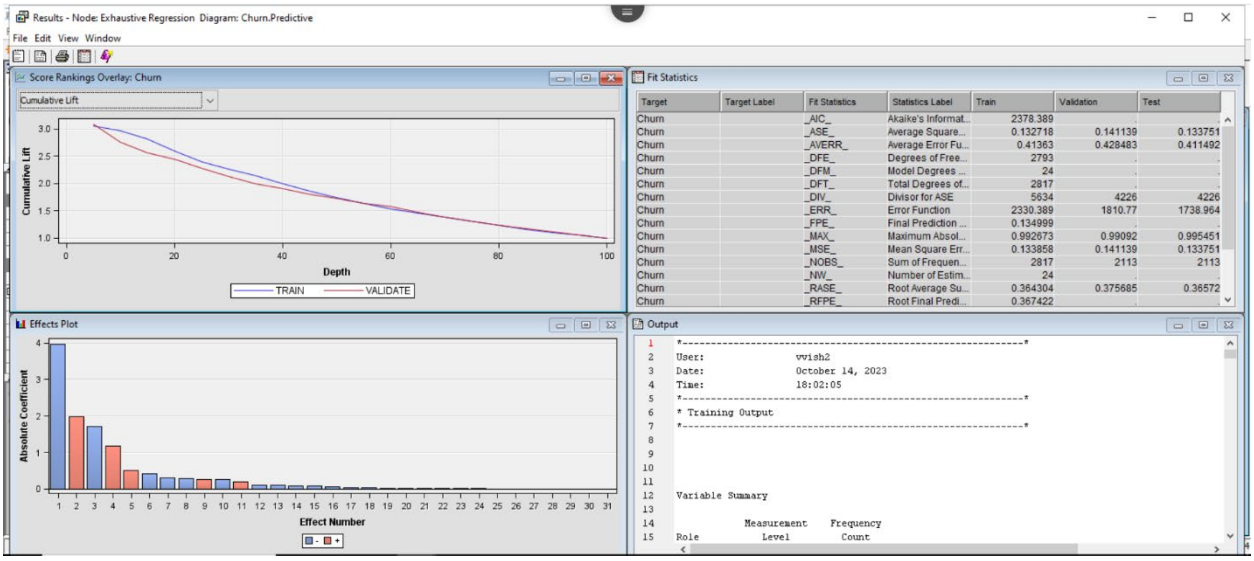
Regression comparison- The best regression model is Exhaustive regression.





	Selected Model	Predecessor Node	Model Node	Model Description	Target Variable	Target Label	Selection Criterion: Valid: Misclassification Rate	Train: Akaike's Information Criterion	Train: Average Squared Error	Train: Average Error Function	Train: Degrees of Freedom for Error	Train: Model Degrees of Freedom	Train: Total Degrees of Freedom	Train: Divisor for ASE	Train: Error Function	Train: Final Prediction Error	Train: Maximum Absolute Error	Train: Mean Square Error	Train: Sum of Frequencies	Train: Number of Estimate Weights
Y	Reg	Reg	Exhaustive ...	Churn			0.205395	2378.389	0.132718	0.41363	2793	24	2817	5634	2330.389	0.134999	0.992673	0.133858	2817	24
	Reg2	Reg2	Forward Re...	Churn			0.207761	2370.058	0.133685	0.415701	2803	14	2817	5634	2342.058	0.13502	0.991041	0.134352	2817	14
	Reg4	Reg4	Stepwise R...	Churn			0.207761	2370.058	0.133685	0.415701	2803	14	2817	5634	2342.058	0.13502	0.991041	0.134352	2817	14
	Reg3	Reg3	Backward ...	Churn			0.209181	2367.884	0.133018	0.41425	2800	17	2817	5634	2333.884	0.134634	0.991473	0.133826	2817	17

Result window of Exhaustive regression:

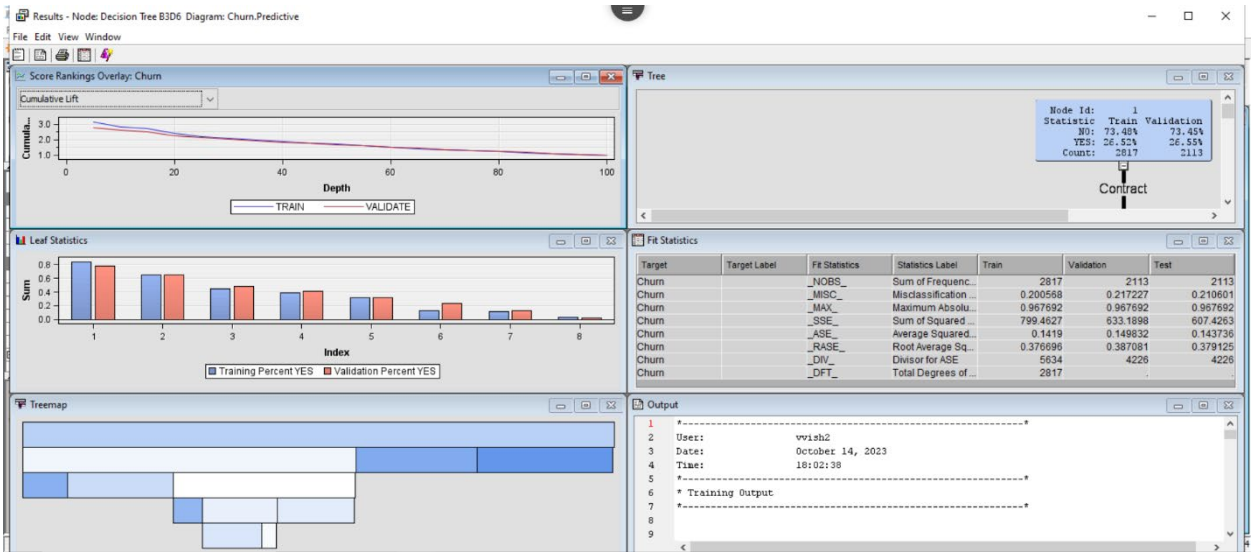


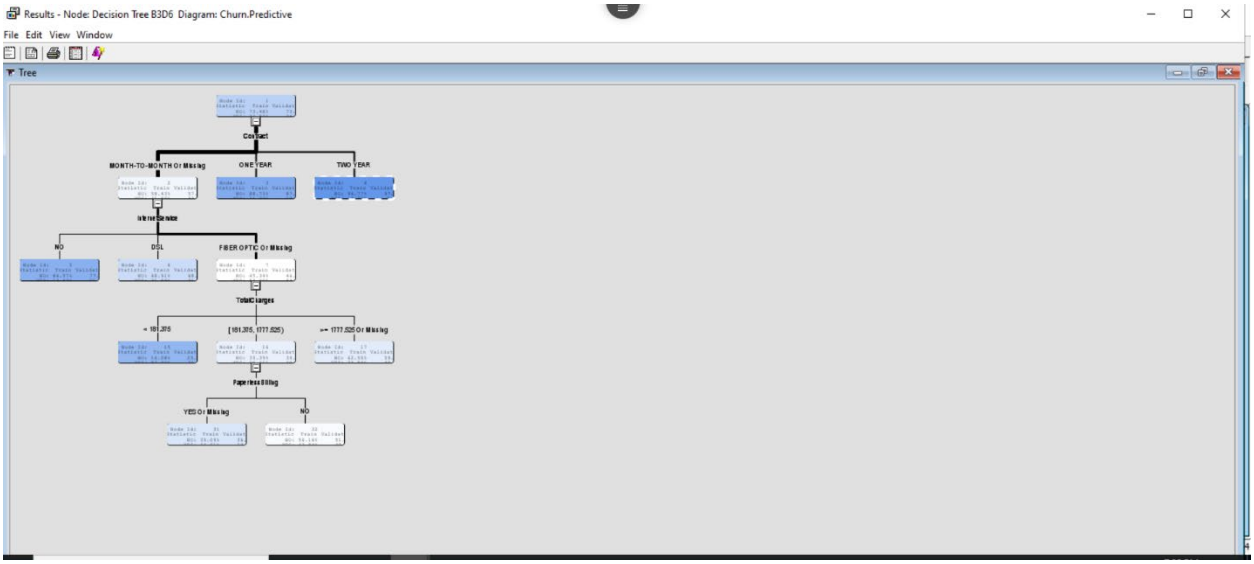
Target	Target Label	Fit Statistics	Statistics Label	Train	Validation	Test
Churn		_AIC_	Akaike's Informat...	2378.389		
Churn		_ASE_	Average Square...	0.132718	0.141139	0.133751
Churn		_AVERR_	Average Error Fu...	0.41363	0.428483	0.411492
Churn		_DFE_	Degrees of Free...	2793		
Churn		_DFM_	Model Degrees ...	24		
Churn		_DFT_	Total Degrees of...	2817		
Churn		_DIV_	Divisor for ASE	5634	4226	4226
Churn		_ERR_	Error Function	2330.389	1810.77	1738.964
Churn		_FPE_	Final Prediction ...	0.134999		
Churn		_MAX_	Maximum Absol...	0.992673	0.99092	0.995451
Churn		_MSE_	Mean Square Err...	0.133858	0.141139	0.133751
Churn		_NOBS_	Sum of Frequen...	2817	2113	2113
Churn		_NW_	Number of Estim...	24		
Churn		_RASE_	Root Average Su...	0.364304	0.375685	0.36572
Churn		_RFPE_	Root Final Predi...	0.367422		

```
*------------------------------------------------------------*
User:                  vvish2
Date:                  October 14, 2023
Time:                  18:02:05
*------------------------------------------------------------*
* Training Output
*------------------------------------------------------------*



Variable Summary

                    Measurement    Frequency
Role                   Level          Count
```

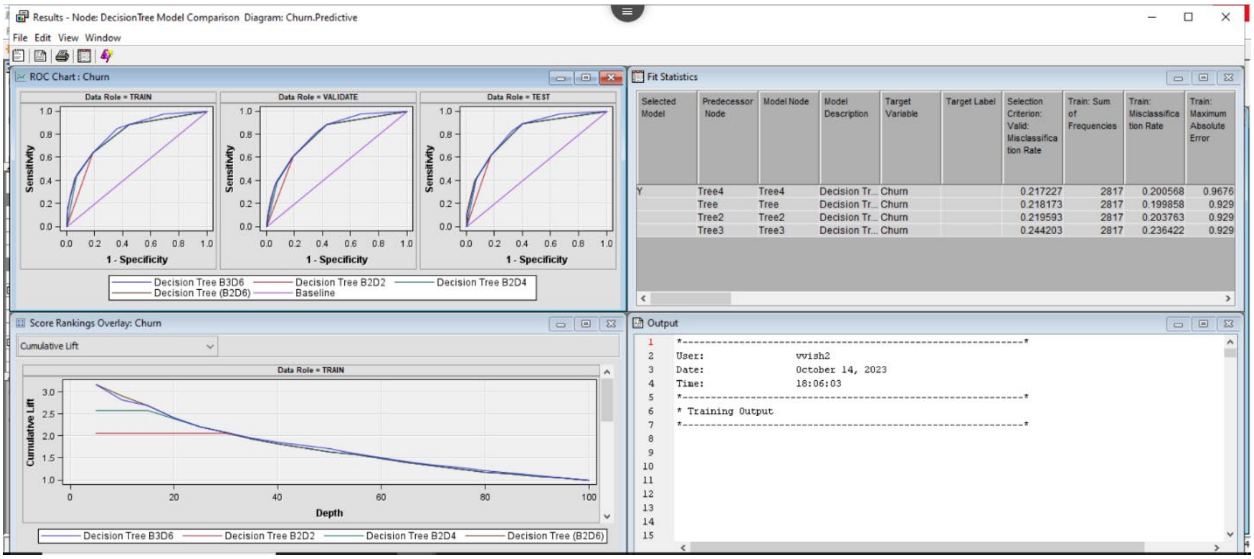
The Average Squared Error (ASE), which compares values from Train to Validation, is 0.132718 for Train, 0.141139 for Validation, and 0.133751 for Test. With a tiny increase in the validation set and a slight drop in the test set relative to the training set, these values show that the model is performing moderately well on all three sets.

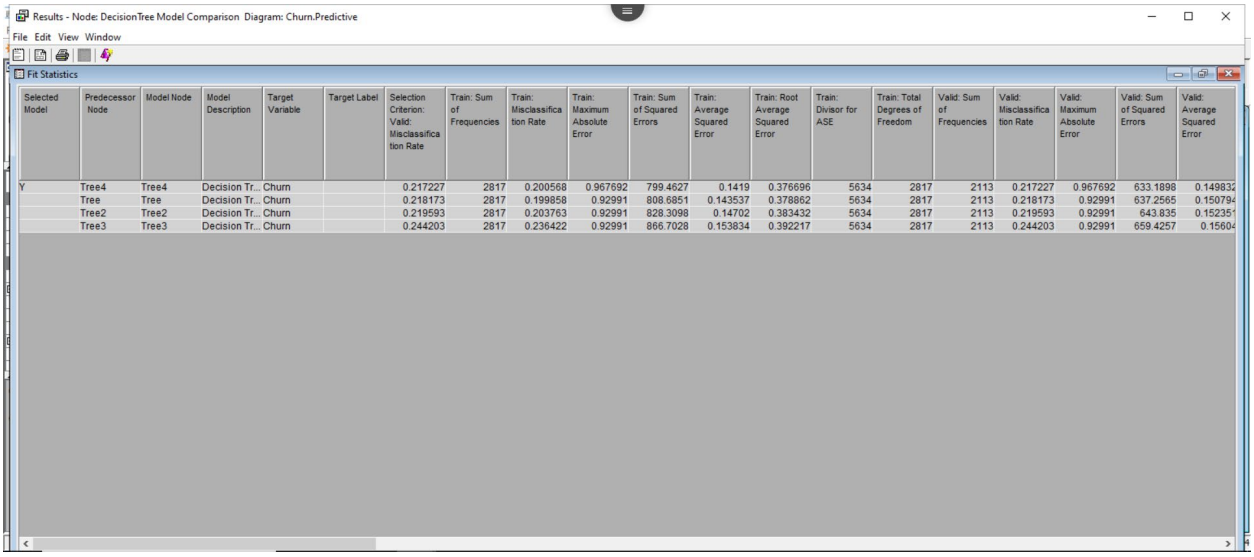
For Decision tree- After analyzing the Results window of B3D6,



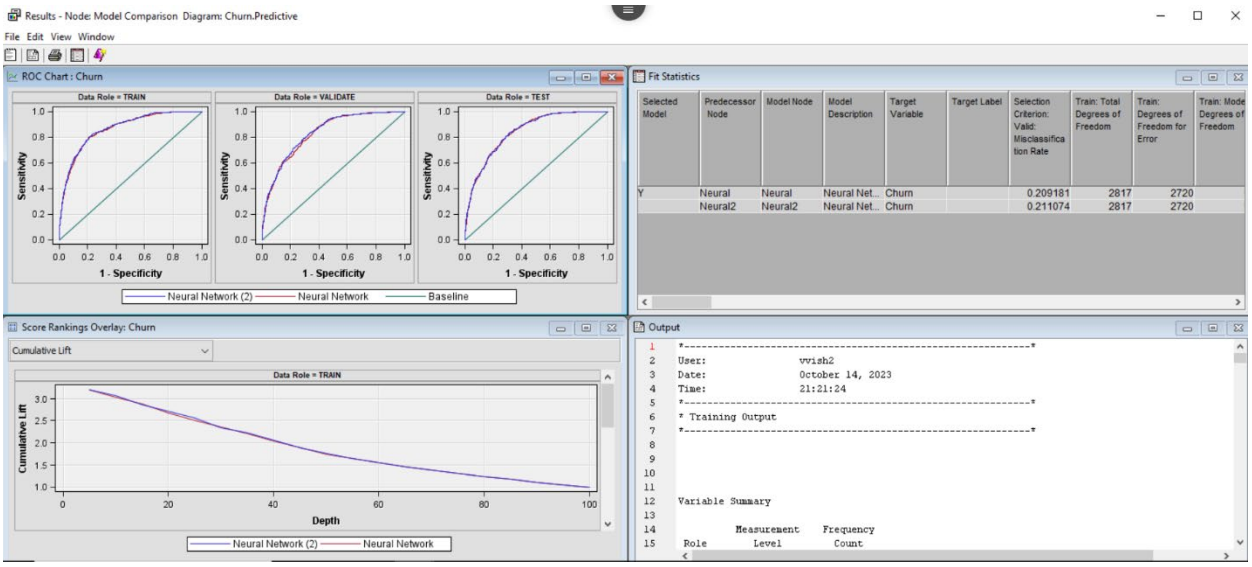


We can say that the best decision tree model is Decision tree B3D6.

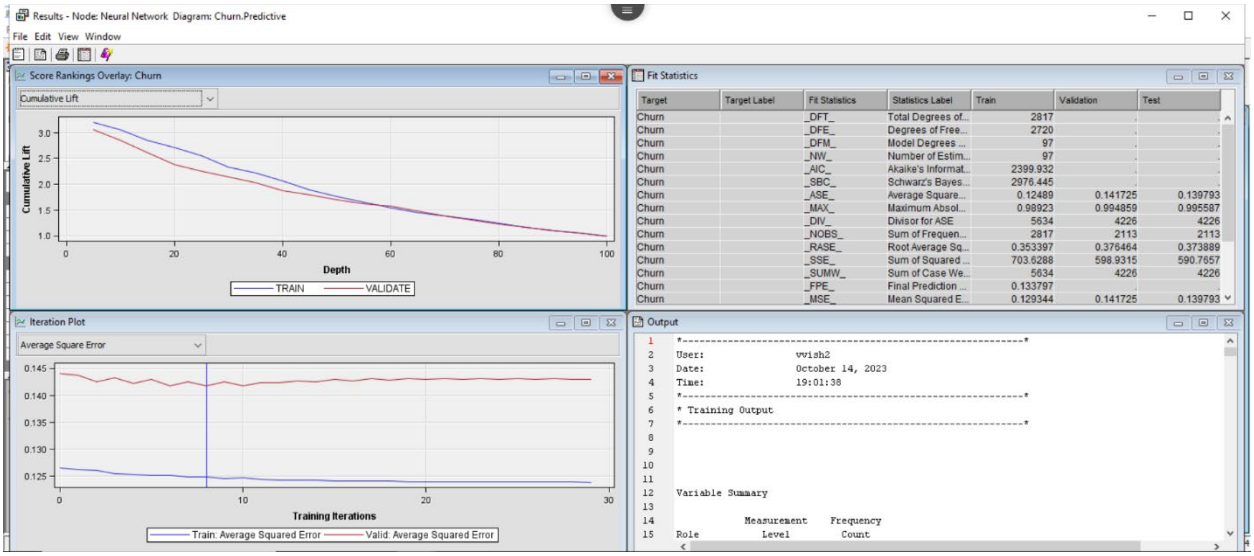

Results - Node: DecisionTree Model Comparison Diagram: Churn.Predictive
File Edit View Window
ROC Chart : Churn
Data Role = TRAIN
Data Role = VALIDATE
Data Role = TEST
Sensitivity
1 - Specificity
Decision Tree B3D6
Decision Tree (B2D6)
Decision Tree B2D2
Baseline
Decision Tree B2D4
Fit Statistics
Selected Model
Predecessor Node
Model Node
Model Description
Target Variable
Target Label
Selection Criterion: Valid: Misclassification Rate
Train: Sum of Frequencies
Train: Misclassification Rate
Train: Maximum Absolute Error
Y Tree4 Tree4 Decision Tr... Churn 0.217227 2817 0.200568 0.9676
Tree Tree Decision Tr... Churn 0.218173 2817 0.199858 0.929
Tree2 Tree2 Decision Tr... Churn 0.219593 2817 0.203763 0.929
Tree3 Tree3 Decision Tr... Churn 0.244203 2817 0.236422 0.929
Score Rankings Overlay: Churn
Cumulative Lift
Data Role = TRAIN
Cumulative Lift
Depth
Decision Tree B3D6
Decision Tree B2D2
Decision Tree B2D4
Decision Tree (B2D6)
Output
User: vvish2
Date: October 14, 2023
Time: 18:06:03
* Training Output


Results - Node: DecisionTree Model Comparison Diagram: Churn.Predictive
File Edit View Window
Fit Statistics
Selected Model
Predecessor Node
Model Node
Model Description
Target Variable
Target Label
Selection Criterion: Valid: Misclassification Rate
Train: Sum of Frequencies
Train: Misclassification Rate
Train: Maximum Absolute Error
Train: Sum of Squared Errors
Train: Average Squared Error
Train: Root Average Squared Error
Train: Divisor for ASE
Train: Total Degrees of Freedom
Valid: Sum of Frequencies
Valid: Misclassification Rate
Valid: Maximum Absolute Error
Valid: Sum of Squared Errors
Valid: Average Squared Error
Y Tree4 Tree4 Decision Tr... Churn 0.217227 2817 0.200568 0.967692 799.4627 0.1419 0.376696 5634 2817 2113 0.217227 0.967692 633.1898 0.149832
Tree Tree Decision Tr... Churn 0.218173 2817 0.199858 0.92991 808.6851 0.143537 0.378862 5634 2817 2113 0.218173 0.92991 637.2565 0.150794
Tree2 Tree2 Decision Tr... Churn 0.219593 2817 0.203763 0.92991 828.3098 0.14702 0.383432 5634 2817 2113 0.219593 0.92991 643.835 0.152351
Tree3 Tree3 Decision Tr... Churn 0.244203 2817 0.236422 0.92991 866.7028 0.153834 0.392217 5634 2817 2113 0.244203 0.92991 659.4257 0.15604

For Neural network- Best model in neural network is Neural network Backprop.

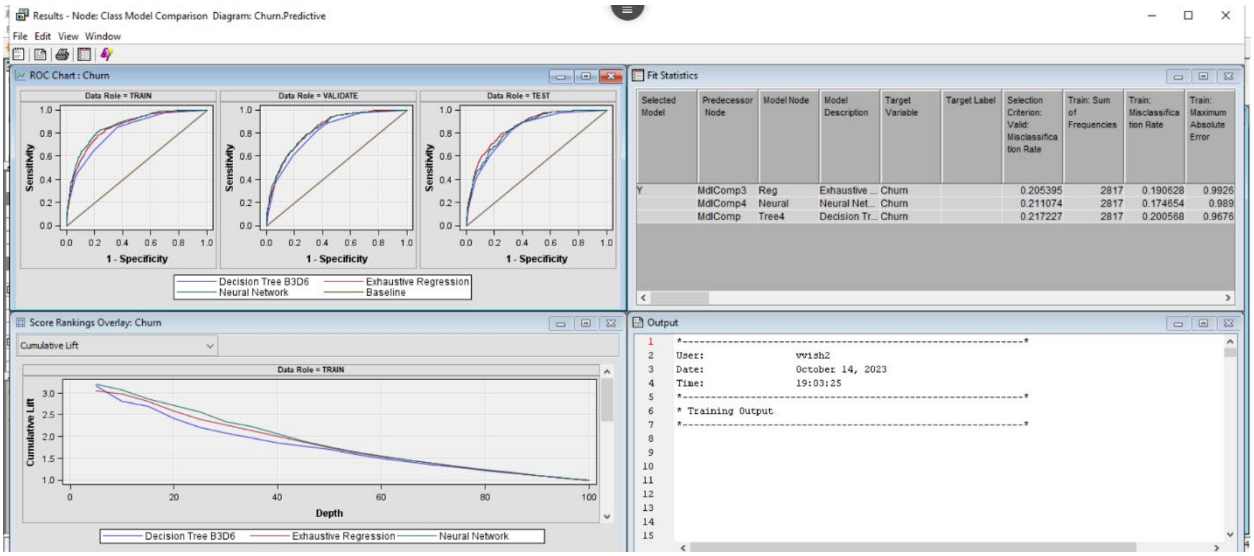


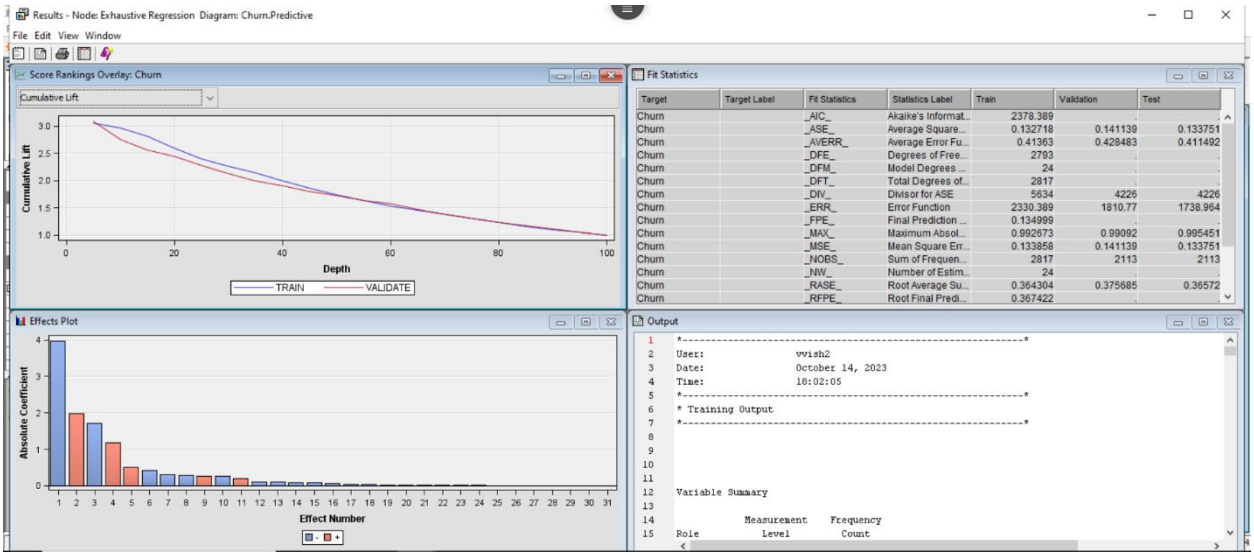
Result window of the best neural network Backprop-



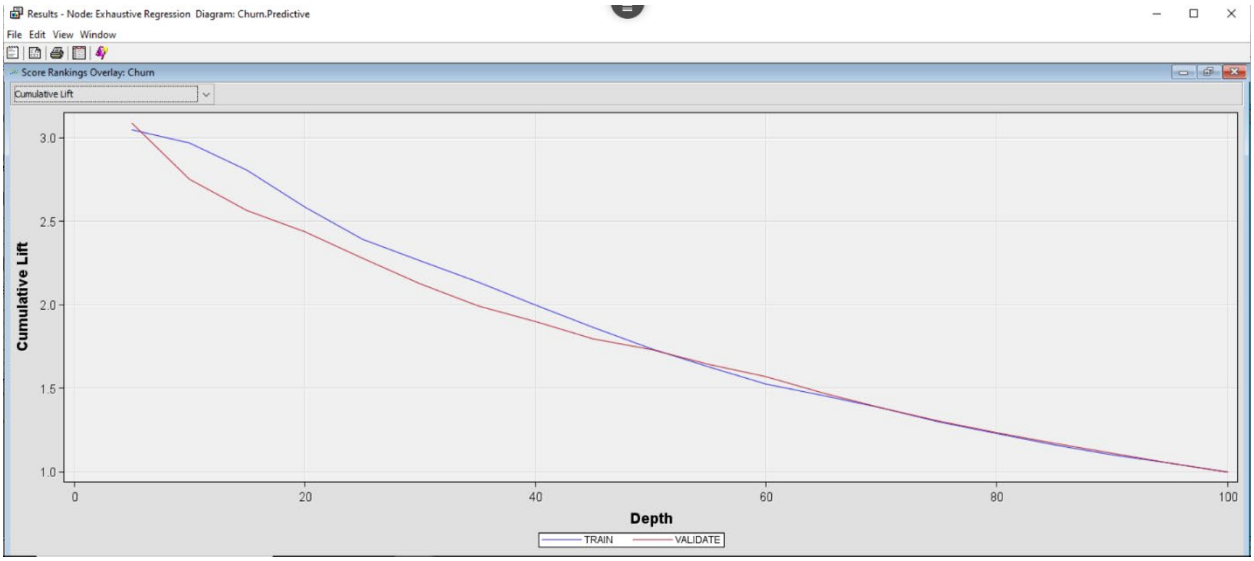
Final comparison of the models-

After comparing all the models we can say that Exhaustive regression is the best model for data analysis.





The graph for score rankings overlay-

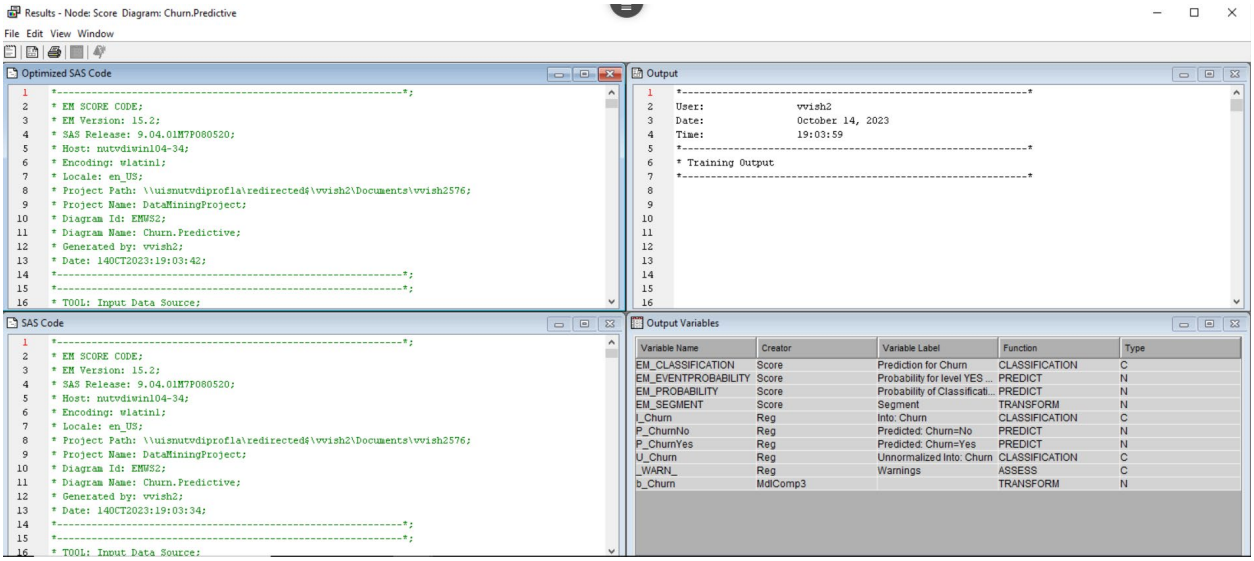


Score Data: The Target's measuring level is evident. The churn prediction variable was built using the binary and Em_classification indices. The score node's results window is shown below.

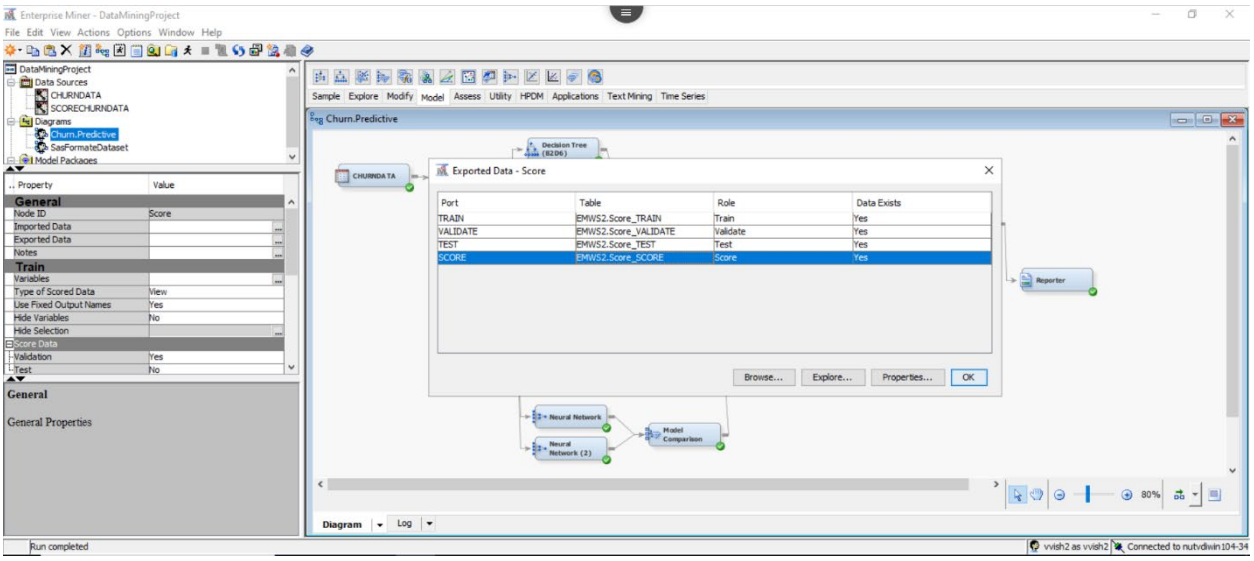
A crucial stage in assessing the prediction model's performance is scoring the model on fresh data. Using a fresh dataset and the predictive model, the scoring node generates predictions for each observation. The accuracy of the model is then evaluated by comparing these predictions to the observed values.

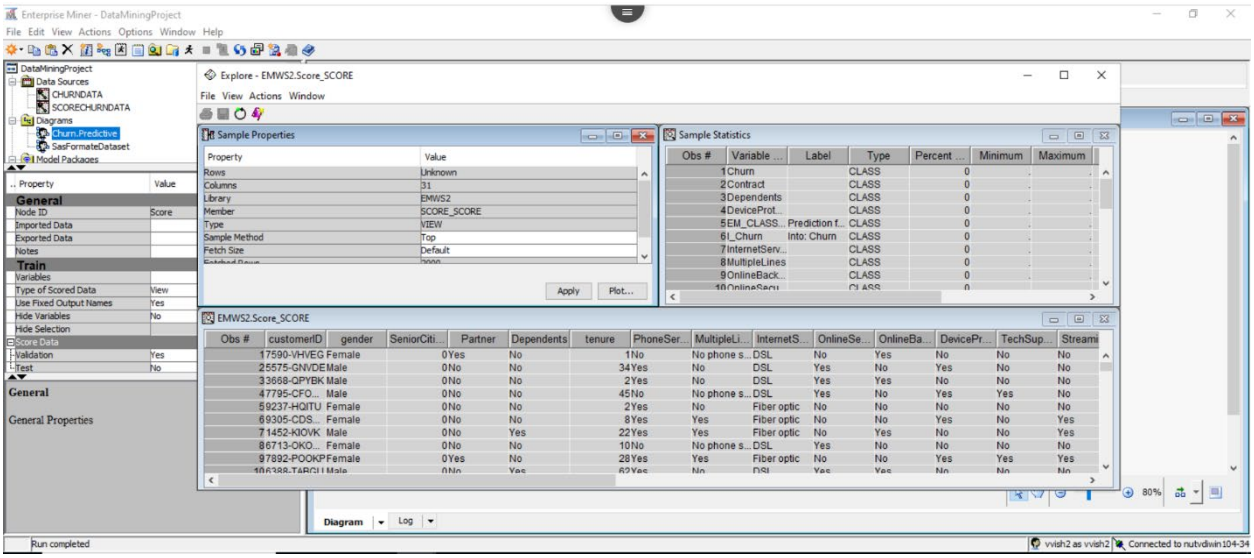
We can use the scoring node to assess the predictive model's performance on fresh data. The scoring outcomes reveal details about how well the model predicted the intended variable.

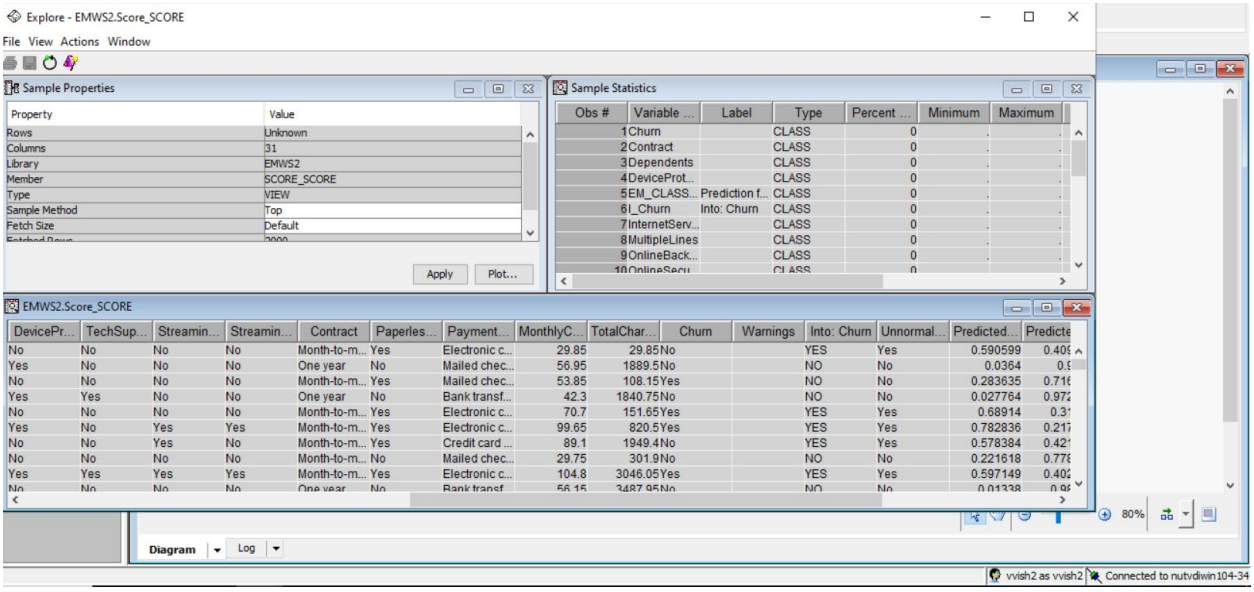
We may assess the findings and evaluate the model's performance after scoring the model on fresh data. The output of the Score node in SAS Enterprise Miner normally contains various performance indicators for the predictive model on the new data, as well as the predicted values for the target variable(s) in the new dataset, based on the results.

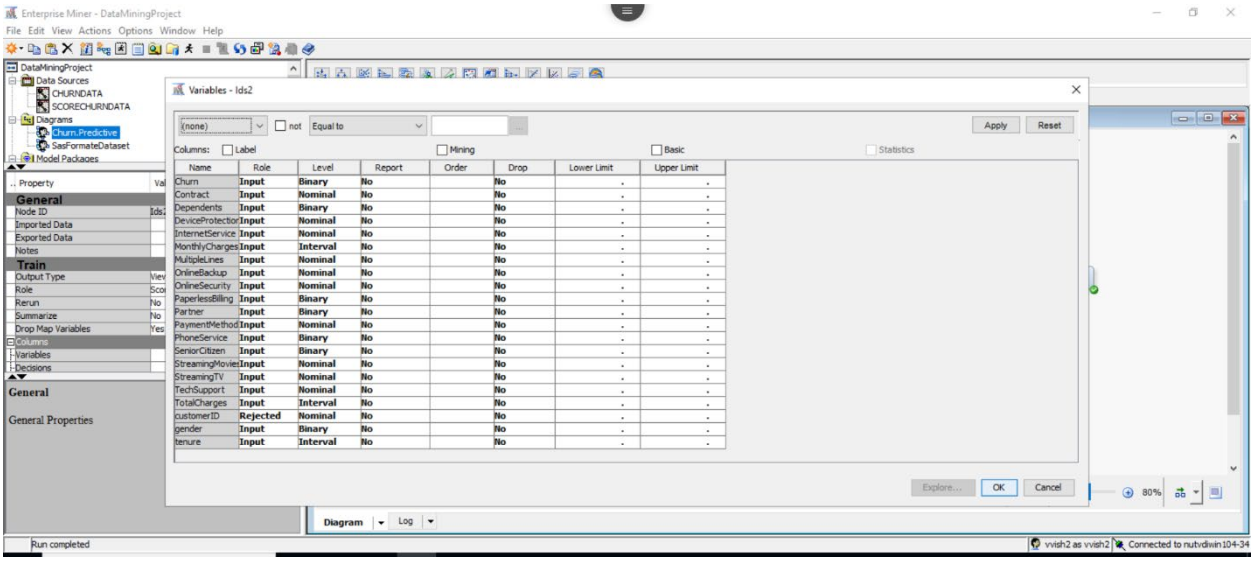


Below we explored the score data









Conclusion: -

The conclusion for the telecom churn dataset would depend on the specific results acquired from the predictive analysis by comparing with the new score data, we discovered after performing the predictive analysis for score data.

But, the goal of predictive analysis is to create a model that can correctly identify which consumers are most likely to leave. The target clients can then be identified using the model to implement retention initiatives, potentially lowering churn rates and boosting customer loyalty.

It can be inferred that the telecom corporation can utilize this model to enhance customer retention efforts and lower churn rates because the predictive analysis is successful and the model can accurately anticipate customer attrition.

References:

Mukherjee, A., & Nath, P. (2013). A study on factors affecting customer churn and proposed a predictive model. International Journal of Advanced Research in Computer Science and Software Engineering, 3(3), 220-227.

Mukherjee, S., & Nath, P. (2013). A decision tree-based approach to predict customer churn in cellular network services. International Journal of Advanced Research in Computer Science and Software Engineering, 3(4), 262-267.

Sheikh, Z. A., & Khattak, H. A. (2014). Analysis of customer churn in telecom industry of Pakistan. International Journal of Management Science and Engineering Management, 9(1), 1-8.

Shankar, R., & Venkatesh, G. (2017). Predicting customer churn in telecom industry using decision tree analysis. International Journal of Applied Engineering Research, 12(21), 10785-10791.

Kasi, V. R., & Raja, G. N. (2015). Analyzing customer churn behavior in telecom sector: A predictive approach. Procedia Computer Science, 48, 372-378.

Hassan, S., &Parves, S. (2019). Analysis of customer churn prediction in telecom sector using machine learning algorithms. International Journal of Advanced Computer Science and Applications, 10(2), 466-473.

- Hassan, S., &Parves, S. (2019). Investigating customer churn prediction in the telecommunication industry: A machine learning approach. Journal of Information & Knowledge Management, 18(3), 1950010. doi: 10.1142/S0219649219500109

Team Contribution-

Day 1-2: Define the problem statement, project goals, and scope

Day 3-4: Obtain the dataset from Kaggle and explore its features and data types

Day 5-6: Conduct preliminary data cleaning and preprocessing

Day 7: Select predictor variables for exploratory and predictive analysis

Day 8-9: Conduct exploratory analysis and generate insights

Day 10-11: Conduct feature engineering and variable selection/dimension reduction

Day 12-13: Build predictive models (e.g., logistic regression, decision tree, random forest) and evaluate their performance

Day 14: Write up the project progress report and refine the project plan for the next 2 weeks