Table of Contents

1. Introduction

The project objective is to predict subscribers or customers of a telecom operator who have a high chance of churning. Identifying these subscribers is essential as enticements can be sent their way that they can be retained. By conducting some data analysis and constructing an artificial neural network we could predict consumers who are likely to churn with a 75% accuracy. Additionally recommendations were made to enable the business take cost saving measures and increase profitability through the reduction of churn.

2. Data Understanding and Business Objective

2.1 Business Problem

The key business issue is low profitability in the telecommunications industry owing to market saturation. Current subscribers are lured to depart by competitors' reduced prices or special offers. As a result, average revenue per subscriber is down and churn is rising.

2.2 Business Objective

To resolve the business problem two key objectives must be met. A robust model must be built to predict which consumer is likely to churn. This will be achieved by building a neural network and decision tree model. Secondly, actionable recommendations would be provided to guide business decision making and elevate profits.

2.3 Data Exploration

The data is composed of numbers and characters. Some of these characters are categorical and would be converted into an appropriate state. The dataset has 7001 observations and 21 columns.

	customerID	gender	SeniorCitizen	Partner	Dependents	tenure	PhoneService	MultipleLines	InternetService	OnlineSecurity
1	7590-VHVEG	Female	0	Yes	No	1	No	No phone service	DSL	No
2	5575-GNVDE	Male	0	No	No	34	Yes	No	DSL	Yes
3	3668-QPYBK	Male	0	No	No	2	Yes	No	DSL	Yes
4	7795-CFOCW	Male	0	No	No	45	No	No phone service	DSL	Yes
5	9237-HQITU	Female	0	No	No	2	Yes	No	Fiber optic	No
6	9305-CDSKC	Female	0	No	No	8	Yes	Yes	Fiber optic	No

	OnlineBackup	DeviceProtection	TechSupport	StreamingTV	StreamingMovies	Contract	PaperlessBilling
1	Yes	No	No	No	No	Month-to-month	Yes
2	No	Yes	No	No	No	One year	No
3	Yes	No	No	No	No	Month-to-month	Yes
4	No	Yes	Yes	No	No	One year	No
5	No	No	No	No	No	Month-to-month	Yes
6	No	Yes	No	Yes	Yes	Month-to-month	Yes

	PaymentMethod	MonthlyCharges	TotalCharges	Churn
1	Electronic check	29.85	29.85	No
2	Mailed check	56.95	1889.50	No
3	Mailed check	53.85	108.15	Yes
4	Bank transfer (automatic)	42.30	1840.75	No
5	Electronic check	70.70	151.65	Yes
6	Electronic check	99.65	820.50	Yes

2.4 Assumptions

There are some important assumptions underlying this dataset.

- All currency amounts are in USD ($)
- Average cost to acquire a new subscriber $750
- TotalCharges is the amount of revenue earned by the telecoms business over a period which is the lifetime billing of the subscriber, i.e. the total billed to the subscriber since they first became a subscriber of the company. Use this field for each subscriber as the amount lost by the business if they were to leave.
- MonthlyCharges are the average (or maybe just the last monthly charge)“fixed” charges made per month to the subscriber.
- The average cost of enticements for the retention of subscribers is 10% of their lifetime billing of the subscriber with a minimum cost of $25.
- The dataset that has been provided that has been randomly sampled from a larger natural dataset and represents the same class distribution as the natural dataset.

2.5 Approach

To meet the business objectives, we will first prepare that dataset and unveil some insights through data analysis. There this we will develop a neural network and decision tree model and have them evaluated. Finally a cost benefit analysis would be made and final recommendations given. This follows the CRISP-DM Data Mining framework which has 6 facets; Business Understanding, Data Understanding, Data Preparation, Modelling, Evaluation, and Deployment.

3. Data Preparation and Preprocessing

3.1 Missing Values

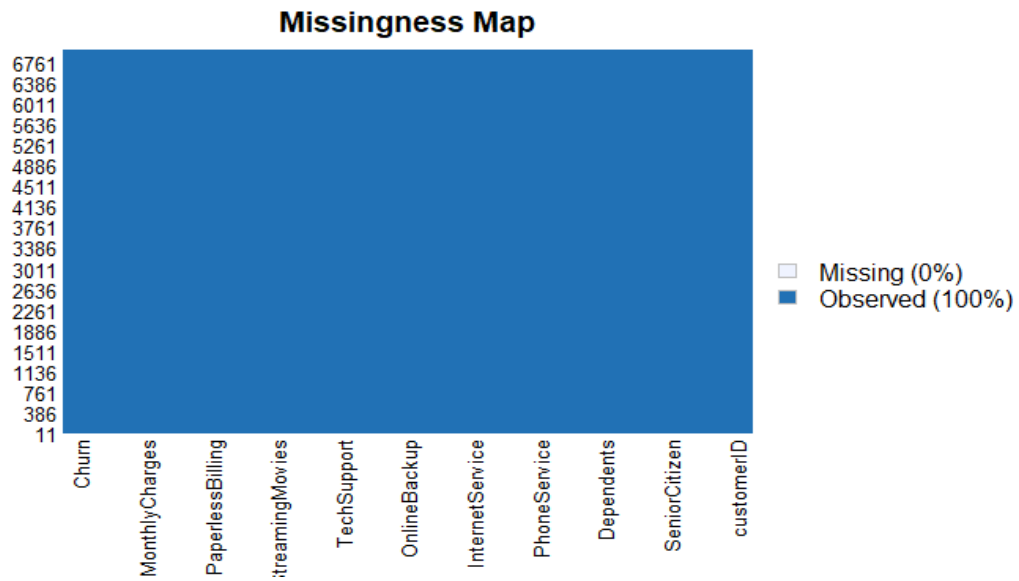
A look at the data shows that only the TotalCharge variable had eleven missing data points.

customerID	gender	SeniorCitizen	Partner	Dependents	tenure	PhoneService
0	0	0	0	0	0	0
MultipleLines	InternetService	OnlineSecurity	OnlineBackup	DeviceProtection	TechSupport	StreamingTV
0	0	0	0	0	0	0
StreamingMovies	Contract	PaperlessBilling	PaymentMethod	MonthlyCharges	TotalCharges	Churn
0	0	0	0	0	11	0

3.2 Data Imputation with Multivariate Imputation by Chained Equation (MICE)

Imputation is the process of replacing missing data with substituted values. There are several ways to achieve this such as using the mean values, the most frequent values and a KNN model. For this project we will use multivariate imputation by chained equation to replace the missing data points. MI (multiple imputation) is a more advanced method for dealing with missing information. MI imputes missing values to construct a number of datasets (denoted by m). That is, in the original dataset, one missing value is replaced by m reasonable imputed values. These figures take into account imputation uncertainty. Each dataset's statistics of relevance are estimated and then integrated into a final one. While single imputation has been criticized for its bias (e.g., overestimation of precision) and ignorance of uncertainty in missing value estimates, MI, when done correctly, can provide an accurate estimate of the true result.

The MI process substitutes numerous potential values for each missing value. This approach, as opposed to single imputation, takes into consideration the uncertainty associated with missing value estimation. Several datasets are generated as a result of the approach, from which parameters of interest can be calculated. If you're looking for a coefficient for a covariate in a multivariable model, for example, the coefficients will be calculated from each dataset, yielding m coefficients. Finally, these coefficients are added together to produce a coefficient estimate that accounts for uncertainty in missing value estimation. This leads the variance of the coefficient estimated this manner is less likely to be underestimated.



By applying the imputation method we can see there are no more missing data points.

3.3 Creating a new variables and recoding words

A new variable called tenure_bin was created. This converts the tenure variable which was in months into years. Additionally certain redundant words like “No internet service” was converted into “No” to reduce the amount of levels we would be working with in some variables.

3.4 Converting variables into factors and numbers

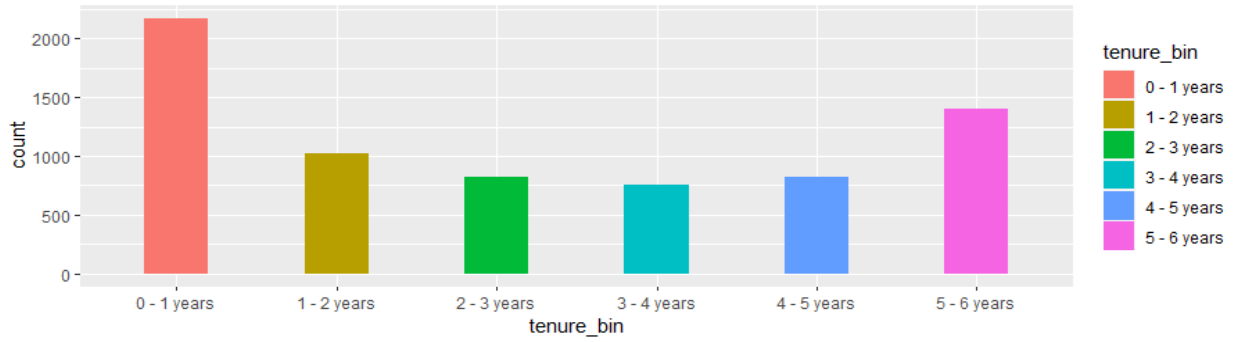
Finally since some variables were wrongly classified as characters in the dataset, a conversion was required. Several character variables were converted to factors and numbers in order to facilitate modeling and data analysis.

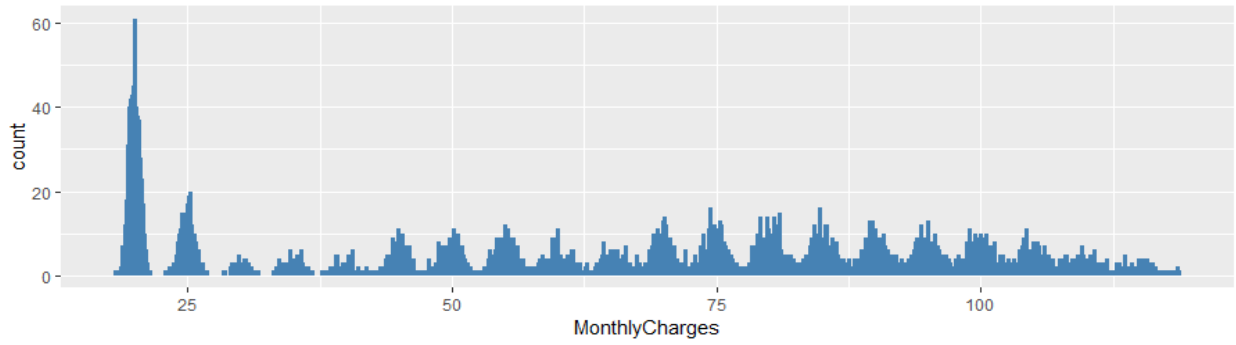
4. Data Exploration

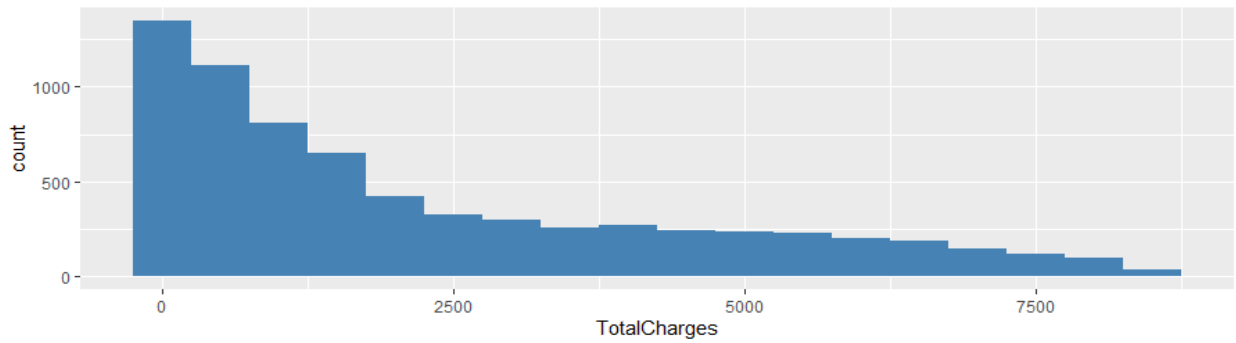
4.1 Exploration of the tenure_bin, MonthlyCharges, TotalCharges Variables

The tenure_bin variable contains the amount of years each subscriber has been a customers to the telecom operator. We can see that most customers have been subscribed to the opertator within the span of a year, this counts up to more than 2000 customers. The next highest tenure group is that of the 5 to 6 year group with a count of about 1200. The tenure group with the least count that is 750 belongs to the 3 to 4 year tenure group.

With regards to the monthly charges we see that most customers are charged a low of approximately $20.00. This increases to about $135.00 for the lowest segment of the subscribers. A similar pattern can be observed for the total charge chart with most of the customers having paying the minimum total charge of approximately $18.00. This also increases to about $8,600.00 for lowest count of customers.

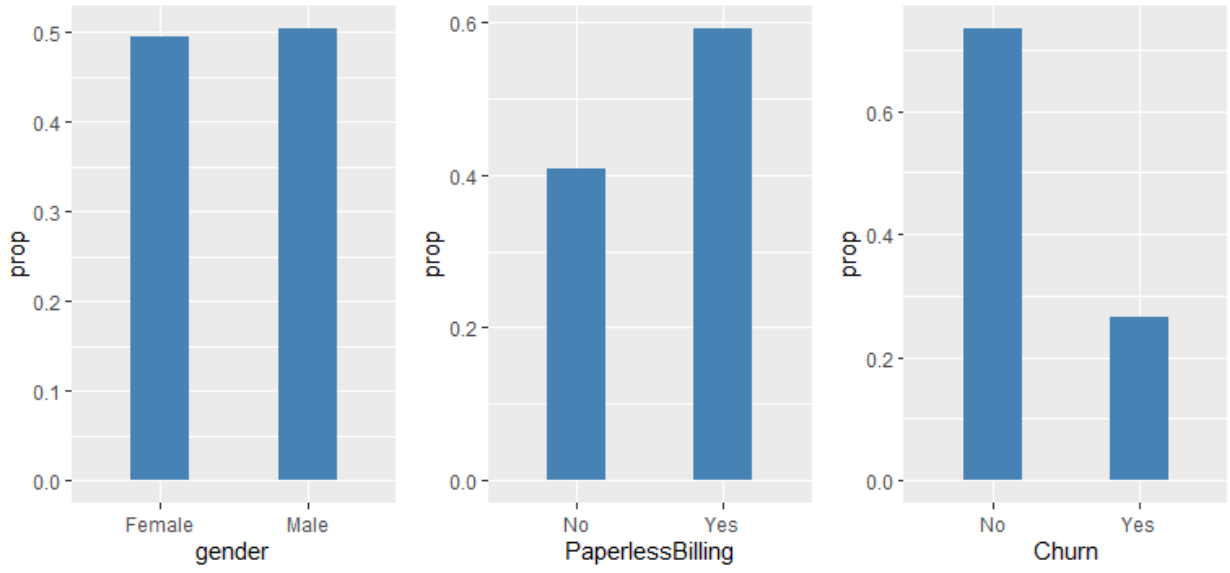






4.2 Exploration of the Gender, PaperlessBilling and Churn Variables

With regards to gender the subscriber base portrays, a relative balance. The male and females seems to have an approximate proportion of about 50%. Again about 60% of the customers are billed through the paperless option whiles about 40% are not. Fortunately most customers in the database do not churn, these subscribers represent roughly 73% of all customers in the dataset, however those who churn represent about 27% only.



4.3 TotalCharges vs MonthlyCharges

A plot of total charges vs monthly charges generally show a linear relationship, as expected. However a deeper dive into the data shows that the people who churn have a less linear relationship than those who do not churn. This implies that a proportional change in monthly charges leads to a less than proportional change in total charges. A reason for this may be that customers who churn don't complete their monthly payment on the month they churn, leading to less total charges.

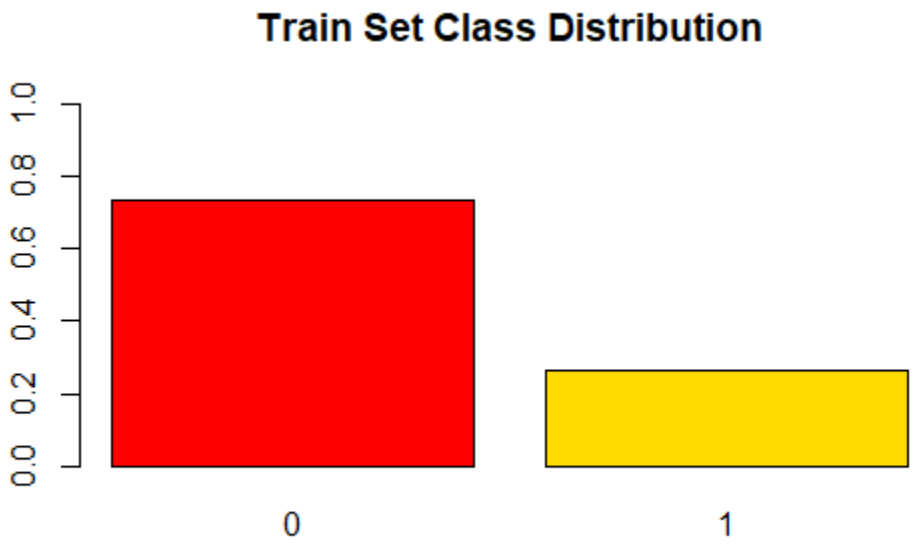
5. Data Modeling

5.1 Data Preparation for Neural Network

The first algorithm we will use is the artificial neural network. A key requirement of this algorithm is that all values out of the 0 and 1 range have to be normalized. As such dummy variables where create for the categorical variables and the numeric variables were normalized. The customer ID and tenure variables were also removed as customer ID was a character variable and had no place in our models. The tenure_bin variable already captured information from tenure and so maintaining it would be redundant.

5.2 Data Splitting and Class Imbalance Correction

A train and a test set of data were created. The train set accounted for 70% of the data, whereas the test set accounted for 30%. Some class imbalance was detected in the train set. “Yes” was converted to 1 and “No” to 0 as part of the preparation for the neural network. The 1 class constitute about 26% of the data whiles the 0 class represented 0.



This was corrected through Downsampling. Downsampling is a mechanism that reduces the count of training samples falling under the majority class. As it helps to even up the counts of target categories and enhance model performance.

After down sampling can class balance was achieved. We will now move to the construction of the artificial neural network.

5.3 Artificial Neural Network

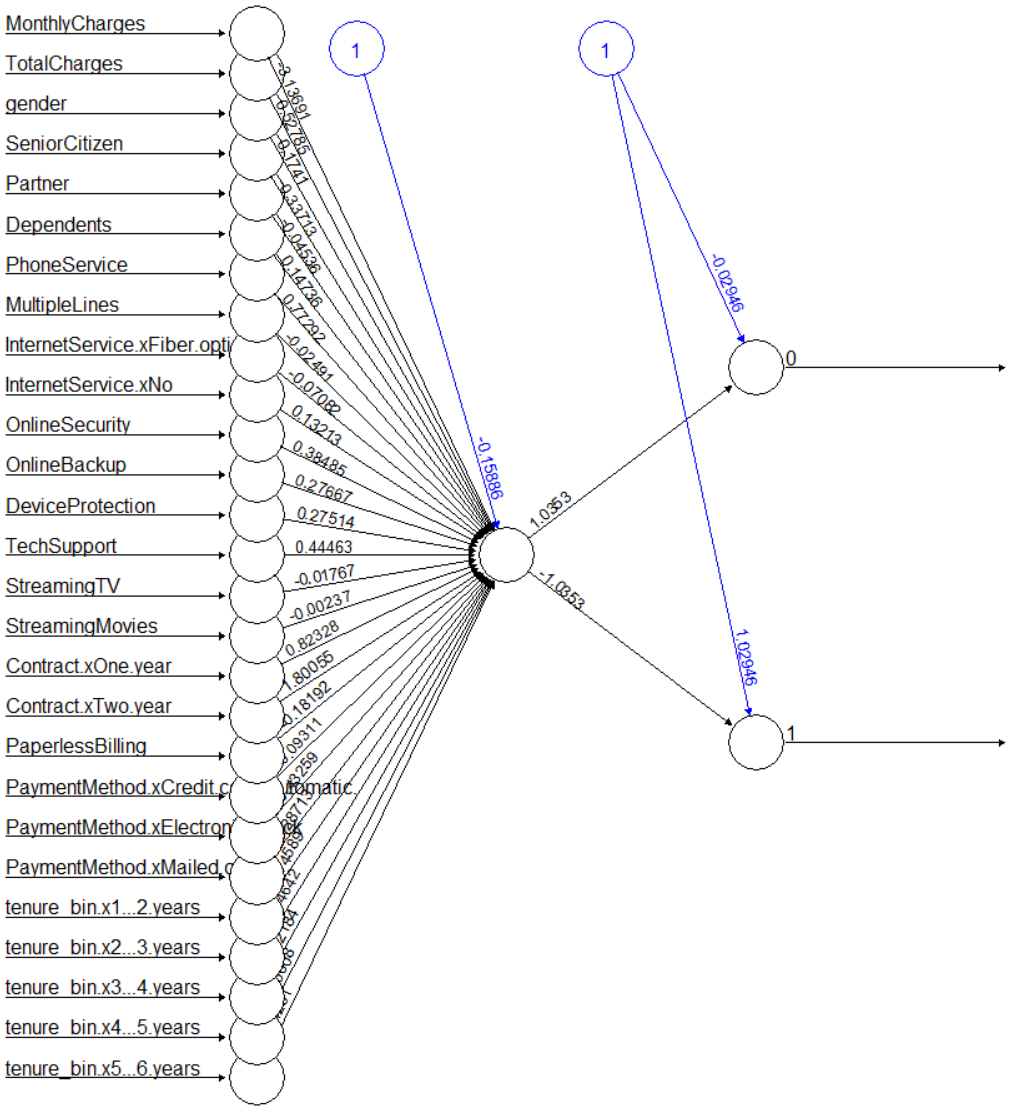
The first algorithm we will use is the artificial neural network. Artificial neural networks are a type of machine-learning algorithm that is based on the structure of the human brain. They can solve issues by trial and error, just like other types of machine--learning algorithms, without being explicitly programmed with rules to follow. In the 1950s, neural networks were created to test theories about how interconnected neurons in the human brain retain information and respond to input data. The output of an artificial neural network, like that of the brain, is determined by the strength of the connections between its virtual neurons — except that in this case, the "neurons" are not actual cells, but rather connected modules of a computer program. Deep learning occurs when the virtual neurons are coupled in multiple layers. Through trial and error, a learning process fine-tunes these connection strengths in order to maximize the neural network's performance in solving a problem. The purpose could be to match input data and make predictions about fresh data the network hasn't seen before (supervised learning), or it could be to maximize a "reward" function to find innovative solutions to a problem (unsupervised learning) (reinforcement learning).

The number and configuration of neurons in a neural network, as well as the division of labor between specialized sub-modules, is frequently suited to each problem. They excel at finding subtle trends and matching patterns in extremely multidimensional data. They also make progress toward their goal even if the programmer does not know how to tackle the challenge in advance. This is useful for problems with complex or poorly understood solutions. For example, in image recognition, a programmer may not be able to put down all of the rules for detecting whether a given image contains a cat, but given enough samples, a neural network can figure out what the relevant elements are on its own.

The layers of neurons are input, hidden, and output. The input layer is made up of record values that are used as inputs to the next layer of neurons, rather than entire neurons. The hidden layer is the next layer. A neural network can have multiple hidden layers. The output layer is the final layer, with one node for each class. The record is assigned to the class node with the

highest value after a single sweep forward across the network assigns a value to each output node.

As explained above, all the variables on the left are the input layers, the ones in blue are the hidden layers and finally the output layers can be found on the far left. All the input layers are multiplied by their respective weights in addition to the hidden layer which serves as impute to the next node.



Neural networks have a number of advantages, some of which are listed below:

1) Save data across the entire network: Similarly to how information is saved on the network rather than in a database in traditional programming. When a few pieces of information vanish from one location, the network as a whole continues to function.

2) The ability to work with incomplete or insufficient knowledge: After an ANN has been trained, the data output can be partial or insufficient. The significance of the missing data influences the poor performance.

3) High fault tolerance: If one or more artificial neural network cells are corrupted, the output generation is unaffected. This improves the networks' ability to tolerate malfunctions.

4) Distributed memory: To train an artificial neural network to learn, it is important to outline the instances and teach it according to the intended output by showing the network those examples. The number of instances chosen has a direct relationship with the network's growth.

5) Gradual Corruption: A network does, in fact, degrade and slow down over time. However, the network is not immediately corroded.

6) Machine-learning capability: ANNs learn from events and make decisions based on comparable events.

7) Parallel processing capability: These networks have numerical strength, allowing them to accomplish multiple functions at the same time.

Some disadvantages of ANN are

1) Artificial neural networks are hardware-dependent due to their structure, which necessitates processors with parallel processing power. As a result, the equipment's manifestation is contingent.
2) The most serious issue with ANN is the network's unexplained behavior. When ANN generates a probing solution, it doesn't explain why or how. The network's trust is eroded as a result of this.
3) Establishing the right network structure: There is no set rule for determining how artificial neural networks should be structured. Experience and trial and error are used to create an appropriate network structure.
4) Difficulty in presenting the problem to the network: ANNs are capable of working with numerical data. Before introducing ANN to a problem, it must be transformed into numerical values. The display mechanism chosen here will have a direct impact on the network's performance. This is dependent on the user's skill level.

5) The network's duration is unknown: When the network's error on the sample is decreased to a specific value, the training is said to be complete. This value does not get the best outcomes.

5.4 Decision Tree

A decision tree is a diagram that depicts the various outcomes of a set of related options. It enables a person or organization to compare and contrast several options based on their prices, probabilities, and advantages. They can be used to spark informal debate or to create an algorithm that mathematically predicts the optimal option. A decision tree usually begins with a single node and branches out into different outcomes. Each of those results leads to new nodes, each of which leads to new possibilities. It takes on a tree-like shape as a result of this.

Some advantages of decision tree are;

1) How simple they are to comprehend
2) They can be used with or without hard data, and they only require minimal preparation for any data.
3) Existing trees can have new options added to them.
4) Their worth in selecting the best selection from a list of numerous
5) The ease with which they can be combined with other decision-making tools

Decision trees, on the other hand, might become overly complicated. A more compact influence diagram can be a useful choice in certain situations. Influence diagrams focus attention on key decisions, inputs, and outcomes.

A decision tree can also be used to aid in the development of automated predictive models for use in machine learning, data mining, and statistics. This method, also known as decision tree learning, uses observations about an item to forecast its worth. Nodes represent data rather than judgments in these decision trees. A classification tree is another name for this sort of tree. Each branch has a collection of properties, or classification rules, that are associated with a specific class label found at the branch's conclusion. These rules, also known as decision rules, can be expressed in an if-then clause, with each decision or data value forming a clause, such as "if conditions 1, 2, and 3 are met, then outcome x will be the result with y certainty," for example. Each new piece of information improves the model's ability to predict which of a finite set of values the subject in question falls into. This data can then be included into a bigger decision-making model. A real value, such as a price, is sometimes used as the predicted variable. Regression trees are decision trees that have an endless number of alternative outcomes.

Multiple trees are sometimes employed in ensemble approaches for greater accuracy:

• Bagging resamples the underlying data to generate numerous trees, which then vote to establish a consensus.

• A Random Forest classifier is made up of many trees that are designed to improve classification accuracy.

• Boosted trees, which can be utilized for both regression and classification.

• PCA (principal component analysis) on a random sample of the data is used to train the trees in a Rotation Forest.

When a decision tree represents the most data with the fewest levels or questions, it is said to be optimal. CART, ASSISTANT, CLS, and ID3/4/5 are all algorithms for creating efficient decision trees. Building association rules and setting the target variable on the right can also be used to create a decision tree. Each technique must find the optimum approach for splitting the data at each level. Measuring the Gini impurity, gaining knowledge, and reducing variance are all common strategies for doing so.

There are various advantages of using decision trees in machine learning:

• With each additional data point, the cost of using the tree to predict data reduces.

• Can model problems with many outputs • Uses a white box model • Can work with category or numerical data (making results easy to explain)

• A tree's accuracy can be checked and quantified • A tree's reliability can be tested and quantified • A tree's accuracy tends to be correct regardless of whether it violates the assumptions of source data

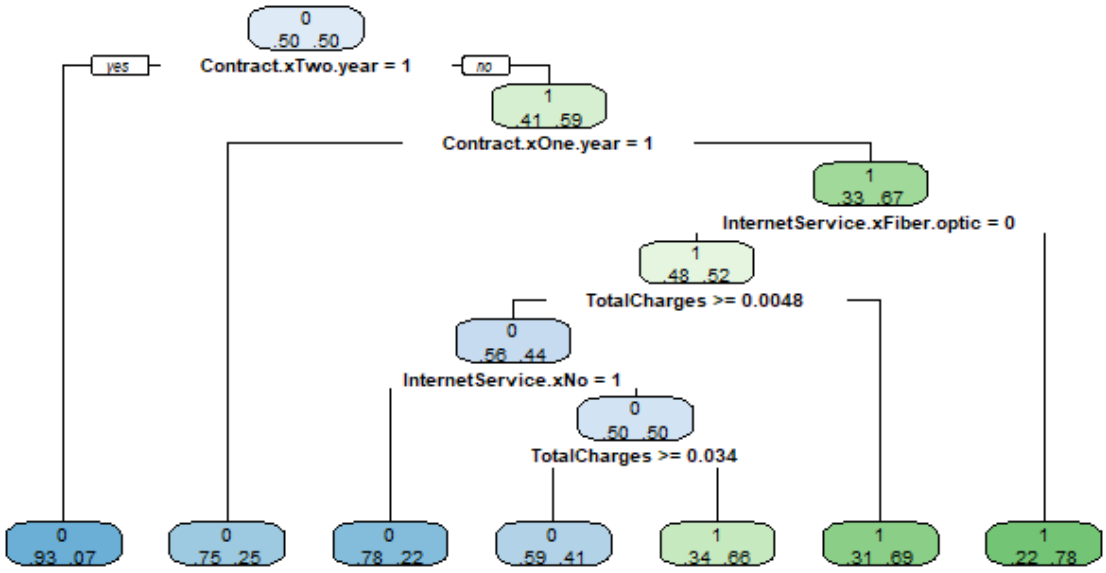
They do, however, have a few drawbacks:

• When working with categorical data that has numerous levels, the information gain is skewed toward the qualities with the most levels.

• When dealing with uncertainty and a large number of interconnected outcomes, calculations can become complicated.

• In decision graphs, nodes can only be linked by AND, whereas in AND graphs, nodes can be linked by OR

With regards to its application in our dataset. We determine that the two year contract was the most important variable. Customers who had a two year contract were less likely to churn. On the other hand customers who don't have a two or one year contract (month to month contract) and lack internet service fiber optics where most likely to churn.



6. Results

6.1 Confusion Matrix

6.1.1 Neural Network Confusion Matrix

The neural network produced a confusion matrix with an overall **accuracy of 0.7534**, sensitivity of 0.7436, and specificity of 0.7810 and a balanced accuracy of 0.7623.

```
          Reference
Prediction    0    1
         0 1157  122
         1  399  435

               Accuracy : 0.7534
                 95% CI : (0.7345, 0.7717)
    No Information Rate : 0.7364
    P-Value [Acc > NIR] : 0.0391

                  Kappa : 0.4523

 Mcnemar's Test P-Value : <2e-16

            Sensitivity : 0.7436
            Specificity : 0.7810
         Pos Pred Value : 0.9046
         Neg Pred Value : 0.5216
             Prevalence : 0.7364
         Detection Rate : 0.5476
   Detection Prevalence : 0.6053
      Balanced Accuracy : 0.7623

       'Positive' Class : 0
```

6.1.2 Decision Tree Confusion Matrix

The decision tree model produced a confusion matrix with an **overall accuracy of 0.7482**, sensitivity of 0.7391, and specificity of 0.7738 and a balanced accuracy of 0.7564.

```
          Reference
Prediction    0    1
         0 1150  126
         1  406  431

               Accuracy : 0.7482
                 95% CI : (0.7291, 0.7666)
    No Information Rate : 0.7364
    P-Value [Acc > NIR] : 0.1128

                  Kappa : 0.4416

 Mcnemar's Test P-Value : <2e-16

            Sensitivity : 0.7391
            Specificity : 0.7738
         Pos Pred Value : 0.9013
         Neg Pred Value : 0.5149
             Prevalence : 0.7364
         Detection Rate : 0.5442
   Detection Prevalence : 0.6039
      Balanced Accuracy : 0.7564

       'Positive' Class : 0
```

In conclusion the neural network was the best performing model and would be used for our ensuing analysis.

6.2 ROC curve

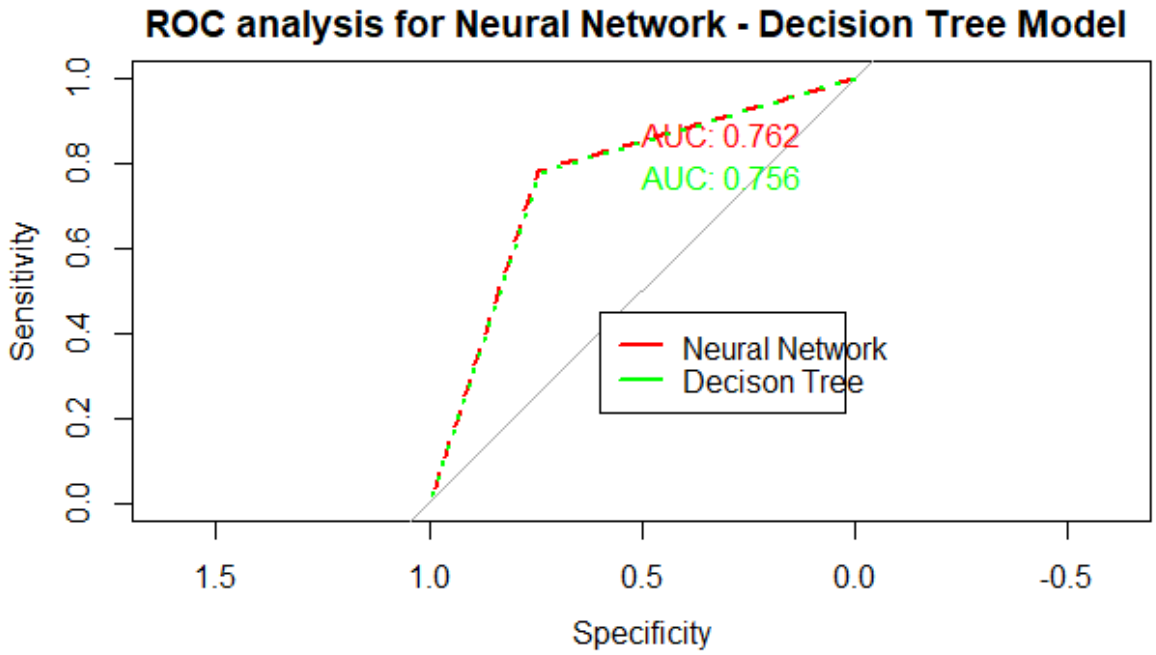
A receiver operating characteristic curve (ROC curve) is a graph that shows how well a classification model performs across all categorization levels. Two parameters are plotted on this curve:

- Sensitivity (True Positive Rate)
- Specificity (False Positive Rate)

The sensitivity of a test is also called the true positive rate (True Positive Rate) and is the proportion of samples that are genuinely positive that give a positive result using the test in question. While specificity (False Positive Rate) is the proportion of samples that are genuinely negatives that gave a negative result using the test in question.

AUC represents the degree or measure of separability.

The ROC graph below confirms our initial evaluation. The Neural Network has a higher AUC of 0.762 making it our model of choice.



6.3 Profitability Analysis

To realize the true value of our model a cost/profit analysis is pivotal. Before commencing that cost/profit analysis certain key values must be gathered. These are

TP= A subscriber was expected to churn and was correctly classified

FP= A subscriber was classified as churn but was actually loyal

FN=A subscriber was classified as loyal but was actually churn

TN= A subscriber was expected as loyal and was correctly classified

Measure	Values
TP	1157
FP	122
FN	399
TN	435
FPR	0.219
TPR	0.7436

Additionally we will use the mean total charge value ($2,279.627) for our calculations

Uplift		
net revenue from subscribers we retain by spending the 10%.	spend: 1157*10%*2,279.627 revenue: 1157* 2,279.627 net revenue: = 2,637,528.44 -263,752.84	$2,373,775.60
wrongly enticed subscribers	(-122*10%*2,279.627)	-$27,811.45
lost revenue due to missed subscribers that we did not entice and so lost	(-399*2,279.627)	-$909,571.17
Cost to replace the lost subscribers	(-399*$750)	-$299,250.00
	Sum for Total uplift from using the model with a test dataset =	$1,137,142.98

You might consider presenting this as a comparison to “if no model existed”. The cost to the business would have been all those subscribers who are known to have churned in the dataset, P=TP+FN. If we assume that the same subscriber revenue lost due to churn is exactly replaced when these are substituted with new subscribers by spending $750 to acquire them, then we get:

nomodel= (P*$750), where P is the number of subscribers that churned

P=TP+FN

P= 1157+399 = 1556

nomodel= (1556 *$750) = $1,167,000.00

nomodel-[model] uplift,

So $1,167,000.00 - $1,137,142.98 = $**29,857.02**

This leaves a profit of $**29,857.02**

7. Conclusion

The neural network algorithm was the best model overall with an accuracy of 0.7534. The model can determine probably subscribers who are likely to churn, enabling the telecom operator offer necessary incentives to ensure their stay. For our data exploration and decision tree model, some key insights were drawn that would improve profitability of the business if applied. These are;

A. Incentives must be given to subscribers on the month to month contract that would convince them into a year or preferably a two year contract. This was the most important factor in our decision tree and would reduce churn if applied.
B. Additionally measure must be taken to give the operators subscriber base access to fiber optic internet. Good internet speed encourages customers to stick with the operator.
C. Subscribers with a tenure of 0 to 1 year must also be targeted and incentivized as they form a bulk of the subscriber base. Prioritizing this segment will increase the company's customer base and profitability.