CUSTOMER CHURN ANALYSIS

Submitted to: Prof. Olga Baysal

ABSTRACT

The project analyzes data of telecommunication customers based on business idea of Churn

Saad Hasan & Chanpreet Singh

INTRODUCTION TO DATA SCIENCE: DATA-5000

CUSTOMER CHURN ANALYSIS IN TELECOM INDUSTRY: A CLOSE LOOK ON IBM DATA

Saad Hasan
Department of Systems and Computer Engineering
Carleton University
Ottawa, Canada

Chanpreet Singh
Department of Information Technology
Carleton University
Ottawa, Canada

ABSTRACT

The project analyzes data of telecommunication customers based on business idea of Churn. Contributing factors which drive the customers to leave the company and join the other service providers is vital for the service provider to find out these factors to control the customer churn problem. The existing method of algorithms like logistic regression and random forest have also been implemented for comparison to check the accuracy and recall of the new and existed methods based on supervised machine learning. The finding of the study shows the factors contributing the customers to churn and different machine learning algorithms used for the accuracy and recall of the churn. These new existed methods include Gradient Boosted Machines and Artificial Neural Network forward feed. In this project, Unsupervised clustering using K-means is also done based on the importance of the customer for the company. The business model created based on business principles is used for the comparison of customer importance. The findings of the study are helpful for the telecommunication service providers to understand the factors and improving their quality if this problem applied to large private company dataset.

Keywords— Customer Churn prediction, Supervised Machine learning, Artificial Neural Networks, Unsupervised Clustering

I. INTRODUCTION:

During the rapid advancement in the field of Information and technology and area of machine learning and data science. Data science inclusion of the advancement of machine learning has played a vital role in different field like Cyber Security in the form of anomaly detection, health care in terms of prediction of cancer tumors malignant or benign, business for predictive analysis. As the use of internet in Canada as of 2018 is around 90 percent and they are increasing rapidly. In Canada, there is no concept of single Telecom giant, but the market is divided into different service providers so it's vital for the business to bloom only if and if only they have a good market share and less customers churn otherwise it's a nightmare for the business to survive and thrive in this world without customers. Telecom customers churn has mainly two types includes voluntary and involuntary. Involuntary customers are those who are forcefully remove by the telecom company due to overcharges, fraud. Voluntary customers are those who left the company based on many factors that could be poor performance, cost and could be the bad behavior of company operators or customer personal reasons. Voluntary churn has two types as well which include incidental and deliberate churn. incidental churn is happening when customer either moved to different location or due to change in financial conditions. While deliberate churn mainly due to quality, advancement in technology and economic reasons. Most of the telecom service providers try to focus more on deliberate churn rather than other types. [1]

Customer is more important than everything for the company because it is big loss for the company to lose an older customer even get a new customer. For
telecom industry where almost, every company is on the race to compete their rivals in terms of technology, increasing businesses, providing better services to the customer to earn maximum profit.
The main objectives of the paper are as follows:

• Definition and explanation of related terms in churn prediction modeling

• Finding the gaps in the existing literature.

• Applications of machine learning algorithms for Customer Churn prediction

• Implementation of the novel framework that helps in churn prediction

• Evaluation of the techniques used in the churn prediction

• Unsupervised clustering of customers based on their importance and value

• Developing a business model based on customer importance and compare with our clustering.

The paper is organized as follows. Starts with the introduction followed by section II which highlights some background and concepts of the customer churn analysis and the research done previously on this topic. In section iii which shed some lights on the main objectives and goals of the project. In section VI represents theoretical concepts of machine learning algorithms used in this study. Section V and VI represses Dataset details and statistical details respectively. Section VII explains the findings of the project followed by section VIII which gives brief discussion about the result. In the section IX we conclude the result and in the last acknowledgment.

II. BACKGROUND

The main objective of this project is to analyze and find the factors which involving the customers to leave the company and which category should be the most important for customer churn analysis. To get the indistinctive reflection of the customer churn analysis, there is a background and concept discussed in this section.

Customer Churn: It is the concept of business where customers leaving the company and joining the other. This concept is discovered by Berson et al. (2000) which means process of subscribers could be prepaid or post-paid switching from one service provider to other. Churn could have many types which includes active, deliberate, rotational, incidental, passive, involuntary [2]. With managing the customers, it is possible to minimize the churn and increase the profit. It is the concept helping Churn in advance. [3]. It is the novel method which widely used in business and interns increased revenue, higher referrals, increased upsell revenue and higher Customer Lifetime Value which bolster and shine the business.

Previous literature on Telecom Customer Churn: Although the concept is old but the application of machine learning techniques in Customer Churn is new. So, there is a brief description of the work done on this topic based on business perspective.

In the existed work, Customer Churn Analysis done in Chain Retail Industry based on the support vector machines and principal component analysis and compare different algorithms. [4]

The advancement of machine learning algorithms like Logistic regression and Decision trees are used for better prediction of Customer Churn in Telecom Industry. [5]. N.Kamalraj and A.Malathi [6] focussed their research on the better understanding of churn prediction using data. This approach used in customer retention activities with respect to customer relationship management.

The existed work on the dataset we are working on are the Logistic Regression and Random Forest.

Customer Importance: Customers in the company can be classify into different levels based on Price, Experience, Performance and Reliability. There is no work done based on the clustering of Telecom customers based on importance.

III. MOTIVATION AND OBJECTIVES

In Canada lots of telecom service providers are keep trying to make more profit by increasing numbers of customers. Giant companies include Rogers communication, BCE Inc., Tellus Corporation and others. By applying different data mining techniques on this IBM dataset will be helpful to understand the churn

prediction in Canadian Telecom industry for the large dataset.

The main objectives of the project are the following,

- First the question arises which factors to consider and are important for the telecom company for churn analysis. In other words, the features contributing the customers to churn.
- In predicting the churn, a comparison of different algorithms to compare based on accuracy, specificity, recall and precision.
- Furthermore, unsupervised learning, Clustering have also been done in this dataset to cluster the telecom customers based on the importance for Telecom company like Highly valuable customer, moderate customers and low valuable customers.

IV. METHODOLOGY

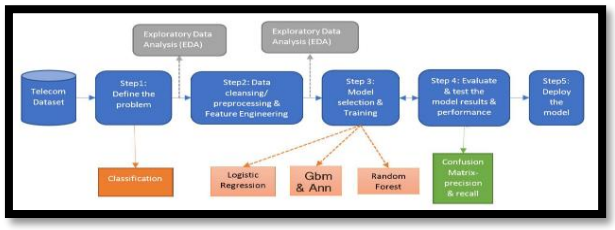

Figure 1: Showing complete cycle of the 1st half of the project

A. LOGISTIC REGRESSION

The first method which is used in our project is Logistic Regression for comparing our model as we need to predict the Churn whether the customer leave the company or not. A logit model Known as Logistic regression is used to model dichotomous outcome variables Under which the log odds of the outcome are modeled as a linear combination of the predictor variables. It helps to explore and evaluate among the relations between one dependent binary variable and one or more nominal, ordinal, interval or ratio-level independent variables. Logistic Regression is a classification algorithm used to predict a binary outcome (1 / 0, Yes / No, True / False) given as a set of variables that are independent.

Logistic regression equation as follows

$$Logit(p) = b_0 + b_1X_1 + b_2X_2 + \dots\dots\dots\dots b_k X_K$$

where p is the probability of presence of the characteristic of interest. The logit transformation is defined as the logged odds:

$$Odds = p/(1-p) = \frac{probability\ of\ presence\ of\ characteristic}{probability\ of\ absence\ of\ characteristic}$$

$$\text{And Logit(p)} = ln\left(\frac{(p)}{(1-p)}\right)$$

Where P = 0 means No and 1 means Yes to not leave and leave the company respectively.

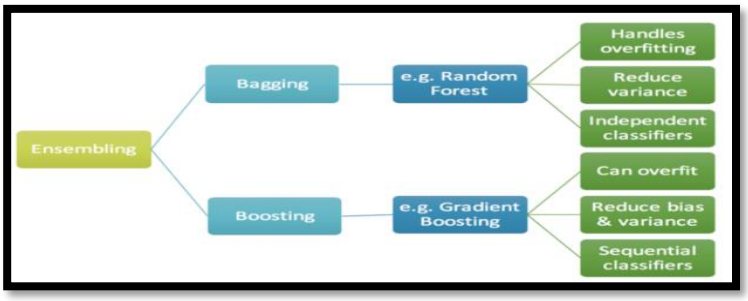

Figure 2: Types of Boosting based on different problems in the dataset

B. RANDOM FOREST

The second existed method is Random Forest which is used in this project, a part of bagging ensemble learning algorithm technique. Random Forest, one of the most Comprehensible and vigorous ensemble methods put-upon today in Machine Learning. Many decision trees are created in the random forest approach where every observation is fed into every decision tree. Random forest is utilized when problems fall under categories:

1. Handles Overfitting
2. Reduce Variance
3. Independent Classifiers

C. GRADIENT BOOSTING MACHINES

Similarly, the third technique used is Gradient Boosting algorithm which is also a part of ensemble learning meaning combination of different algorithms for better prediction. The method used to reduce bias and variance and good for sequential classifiers, but it may overfit.

D. ARTIFICIAL NEURAL NETWORK

Artificial neural networks (ANN) systems are intelligent computing systems that resemble the biological neural networks in human brains. An ANN consists the networks of connected units or nodes called artificial neurons [3]. In an ANN, a typical artificial neuron receives the signal, process it according to activation function used to program it to send to the next artificial neurons connected to it via a connection between two consecutive nodes.

One of the most popular ANN paradigms is the feed-forward neural network (FNN) and the associated back-propagation (BP) training algorithm. Feedforward Neural Networks are the type of artificial neural networks where the connections between do not form a cycle. Feedforward neural networks were the first type of artificial neural network invented and are simpler than their counterpart, recurrent neural networks. They are called feedforward because information only travels forward in the network (no loops), first through the input nodes, then through the hidden nodes (if present), and finally through the output nodes.

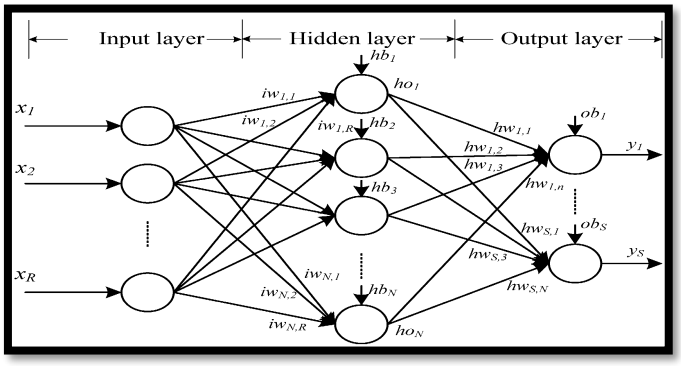

Figure 3: A feedforward Neural Network with information flowing left to right: Courtesy MATLAB

Feedforward neural networks are primarily used for supervised learning in cases where the data to be learned is neither sequential nor time-dependent. there are no feedback connections or loops in the network. It has an input layer, an output layer, and a hidden layer. In general, there can be multiple hidden layers. Each node in the layer is a Neuron, which can be thought of as the basic processing unit of a Neural Network.

In a Neural network, An Artificial Neuron (AN) or a perceptron is the basic unit that does all the decision taking the task in very small level. A schematic diagram of a neuron is given below. An ANN takes inputs with their corresponding weights(*w*) and adds them. After summation it uses an activation function to normalize the sum. There are weights called bias(*b*) associated with each input of a neuron. These are the parameters which the network learns during the training phase.

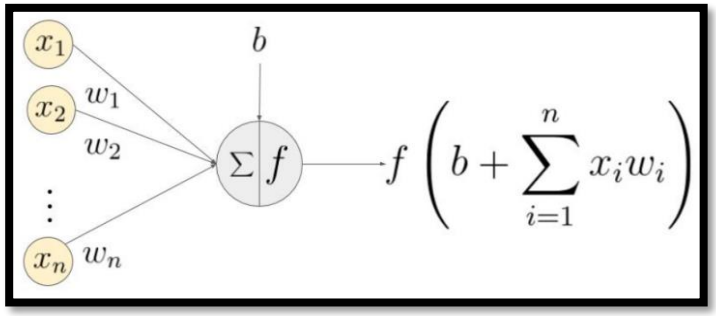

Figure 4: An example of a neuron showing the input (x1 – xn), their corresponding weight (w1 – wn), a bias (b) and the activation function f [2].

Activation Functions:

In a neuron, all the decision about the output is being taken by the activation function. It also helps the neuron learns Linear or Non-linear decision boundaries [1]. Additionally, because of its cascading effect it limits the output values of neurons to become very large after several layers by normalizing the output of the neuron. There are three most widely used activation functions

- Sigmoid
- Tanh
- Rectified Linear Unit (ReLU)

FNNs are usually organized into something called layers. Generally, a typical artificial neural network composed of an input layer, zero or few hidden layers, and an output layer. While there are no hidden layers in a single-layer perceptron. On the other hand, there is at least one hidden layer of multiple perceptions in an FNNs. Each layer importance in an FNN in described below

Input Layer: This is the initial layer of a neural network supply the input data or features to the network

Output Layer: This is the final layer which gives out the predictions. For a regression problem, where the output is not a predefined category, we can simply use a linear unit.

Hidden layer: A feedforward network applies a series of functions to the input. By having multiple hidden layers, we can compute complex functions by cascading simpler functions. The number of hidden layers is termed as the depth of the neural network.

Working of Artificial feed forward network:

The input nodes simply pass on the input vectors x_i. The nodes in the hidden layer and output layer are processing units. Each processing node has an activation function which is commonly chosen to be the sigmoid function, where is a constant controlling the slope of the function. The net input to a processing unit j is given by

$$f(x) = b + \sum xi\ w_i \qquad (i)$$

where x_i's are the outputs from the previous layer, w_i is the weight (connection strength) of the link connecting unit i to unit j, and b is the bias, which determines the location of the sigmoid function on the x axis.

A feed-forward neural network works by training the network with known examples. A random sample $(x_p\ y_p)$ is drawn from the training set and x_p is fed into the network through the input layer. The network computes an output vector op based on the hidden layer output. o_p is compared against the training target y_p. A performance criterion function is defined based on the difference between op and y_p.

The error computed from the output layer is backpropagated through the network, and weights (w_i) are modified according to their contribution to the error function. where is called learning rate, which determines the step size of the weight updating.

E. K-MEANS ALGORITHM

K-means clustering is used to clusters the data based on the importance of customers. K means measures the distance of each point with a specific cluster and the cluster which has less distance with the point joins the same cluster using Euclidean distance.

V. DATASET AND LIMITATION

The dataset contains 7022 rows and 21 columns. Features include total Charges, monthly charges, tenure, services availed by customers like Internet Services, Phone Services, type of connections, contract etc. The dependent variable is Churn, whether the customer leaves the company or not. The data set was collected from IBM website which was published on April 2018.

VI. DESCRIPTIVE STATISTICS & IMPLEMENTATION DETAILS

In this section, the descriptive statistics of our project which includes all the details.

Data Cleaning: Data cleaning takes most of the time for the data science project. As most of the data in our data set are categorical with string and false values so, it's replaced with numerical identity and correction of values based on other columns and domain knowledge. The most important step in data analysis is to find the correlation between variables based on positive or negative correlation. After omitting the missing values and replacing the tenure zero with multiple condition based on total charges, visualize the categorical and continues variable. In figure.6 in internet service graph, which is showing most of the customers who churn are the ones using Fiber optics.

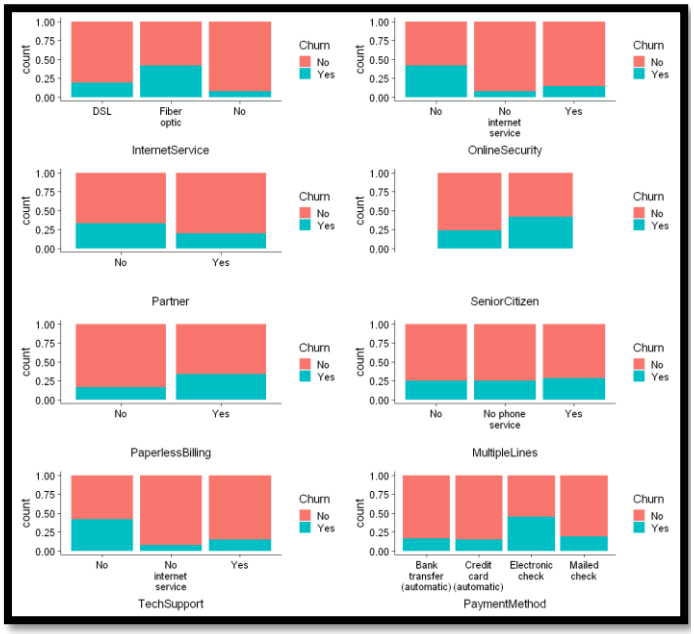

Figure 5: Showing the categorical variable based on dependent variable Churn

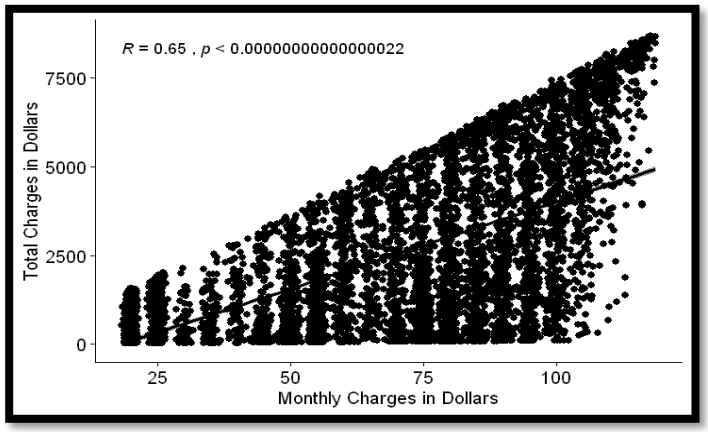

Figure 6: Showing relation between continuous variables, a strong relation between them

In Figure 7: also showing the correlation between all variables visually based on the correlation coefficients if they are correlated or not based on the scale of -1 and 1 while nearer to 1 supports stronger relation and the

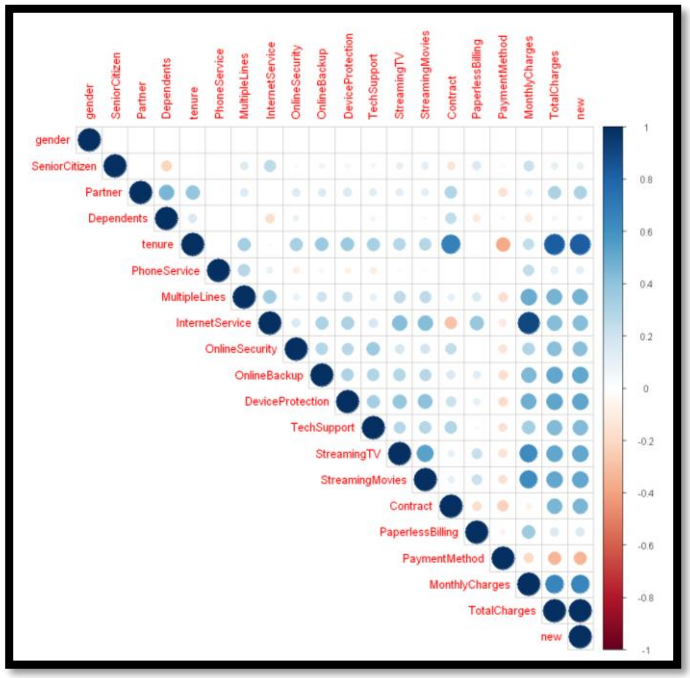

Figure 7: Showing the correlation between all variables

The parameters selected based on glm method are Internet service, Monthly Charges, Total Charges and tenure. In simple words, it predicts the probability of occurrence of an event by fitting data to a logit function. The most important features based on less p-value and

```
Call:
glm(formula = Churn ~ tenure + +TotalCharges + Contract + InternetService,
    family = "binomial", data = traindata)

Deviance Residuals:
       Min          1Q      Median          3Q         Max
-1.5459065  -0.7026787  -0.3044734   0.8515745   3.5016446

Coefficients:
                    Estimate     Std. Error   z value              Pr(>|z|)
(Intercept)     -0.33118367993  0.14738363072  -2.24709              0.024635
tenure          -0.05991519703  0.00651590885  -9.19522 < 0.000000000000000222
TotalCharges     0.00031873798  0.00006874361   4.63662           0.0000035415
Contract        -0.87096141906  0.08375180090 -10.39932 < 0.000000000000000222
InternetService  1.03292000198  0.06678236896  15.46696 < 0.000000000000000222

(Intercept)     *
tenure          ***
TotalCharges    ***
Contract        ***
InternetService ***
```

Figure 8: Showing the features selected by glm and their p-values and other details

more correlation which in R programming, there is a method which is known as gbm used for feature selection as well. The five features which selected by the gbm are internet charges, monthly charges, total charges, contract and tenure. These same features are also used in random forest and neural network as well. The most important feature is contract whether the contract is monthly, yearly or two-year contract, followed by Internet service, Total charges, monthly charges and tenure. The threshold selected for all the algorithms like logistic regression, gradient boosting machine and random forest is 0.5 i-e if it is greater than that it will be 1 otherwise 0

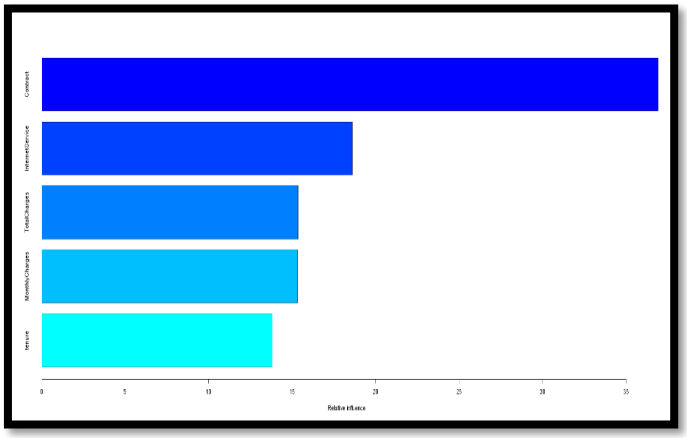

Figure 9: Features selected by GBM method showing importance graphically

	var	rel.inf
Contract	Contract	36.93163895
InternetService	InternetService	18.60420304
TotalCharges	TotalCharges	15.35249884
MonthlyCharges	MonthlyCharges	15.30416170
tenure	tenure	13.80749747

Figure 10: Features selected by GBM representing in terms of variances

```
# create model
model <- keras_model_sequential()
# define and compile the model
model %>%
  layer_dense(units = 100, activation = 'relu', input_shape = c(9)) %>%
  layer_dropout(rate = 0.5) %>%
  layer_dense(units = 264, activation = 'relu') %>%
  layer_dropout(rate = 0.5) %>%
  layer_dense(units = 18, activation = 'relu') %>%
  layer_dropout(rate = 0.5) %>%
  layer_dense(units = 1, activation = 'sigmoid') %>%
  compile(
    loss = 'binary_crossentropy',
    optimizer = 'adam',
    metrics = c('accuracy')
  )
summary(model)
# training model
annmodel <- model %>% fit(train_x,
                train_y,
                epochs = 500,
                batch_size = 100,
                validation_split = 0.10)
```

Figure 11: Showing the parameters for the Artificial Neural Networks

In parameters of neural network, the optimizer used is Atom which is one of the good optimizers. Loss function is binary cross entropy because our problem is classification while other details showed in the figure.12. As it is implemented in the python library Keras which uses TensorFlow in its backend, so the model is used is sequential, as the model could be used as functional with more powerful neural network and more elasticity. There are two hidden layers with 264 neurons and 18 neurons respectively. Input layers contains 9 inputs as we have 5 inputs but two input variables like Contract and Internet Service is hot encoded meaning the creation of dummy variables as we have three 3 categories in each, so it will be 6 variables including other 3 will make it 9 variables

as an input. The output layer has one output which is the prediction of probability which intern change to 1 or 0. The activation function we used in input layer and hidden layer is Rectifier Linear Unit while for the output layer we used sigmoid function as it's the binary classification problem.

VII. RESULTS

This section discusses the result and comparison of the different algorithms based on confusion matrix. If we go deeper into the results the accuracies of all the algorithms are all most equal with little differences as shown in the table-1. If we investigate the previous existed algorithms like Random forest and Logistic regression although the accuracy is more than 75 percent, but the recall is too less in both algorithms. In Logistic regression it's 0.44 and in Random Forest it's 0.33 which is not even greater than 50 percent of the recall value. As the recall value is so poor for the existed algorithm as every customer who is leaving the company must be identified so, we focused on different machine learning techniques to get good recall value. First gradient boosting machines algorithm used to predict Churn. So, the recall is increased to 0.50 which is 50 percent, but this recall is still too less. We focused on Neural Network for implementing neural network in KERAS we got promising result which is still not good as the recall still less than 75 percent but it's good as compared to others although accuracy is improved to 84 percent, but our focus is recall which is 73 percent refer to table-1.

	Logistic Regression	Gradient Boost Machine	Neural Network	Random Forest
Accuracy	0.77	0.80	0.84	0.78
Precision	0.62	0.65	0.50	0.72
Recall	0.44	0.50	0.73	0.33
Specificity	0.90	0.90	0.84	0.95

Table 1: Showing the comparison of different algorithms based on confusion matrix

Unsupervised clustering: The second part of the project is unsupervised clustering. The most common method is K-means algorithm, so it's implemented based on two important continuous variables monthly charges and tenure. In figure.12 which shows the three clusters. Cluster 1 in pink represents least important customers for the company as these customers have less tenure and less monthly charges. While cluster 2 in blue represents moderate importance customers for the company as they have more monthly charges but less tenure. The most important customers are in cluster 3 which represents in orange color as shown in the figure 11 because these are the prioritized customers as if the company lost these customers it is a big loss for them. figure.12: Showing clusters of customers based on importance using k-means algorithm.

Business Model: we created a business model to compare our soft k-means clustering. The function our business model as shown in the figure.14 which based on the conditions that if the customer's duration is less than one year and the monthly charges less than 50 dollars then we put him in the first type which is less important customers. If the customer duration is less than one year and monthly charges are greater than 50 dollars or if the charges are less than 50 dollars, but their duration starts somewhere in second year then the customer we put him in the category 2 which is the category of moderate customers. If the customers duration is greater than two years, then we call them the important customers as they are prioritized customers and we put them in the category 3.

```
ClusteringModel <- function(x){
  a <- x[1]
  b <- x[2]
  c <- x[3]
  d <- x[4]
  e <- x[5]
  value <- if(a > 0 & a < 12 & b > 18 & b < 50) c
      else if(a > 0 & a < 12 & b > 50) d
      else if(a > 12 & a < 24 & b > 18 & b < 50) d
      else if(a > 12 & a < 24 & b > 50) e
      else if(a > 24 & b > 18 & b < 50) e
      else if(a > 24) e
  return(value)
}
```

Figure.13: Showing the function of our business model

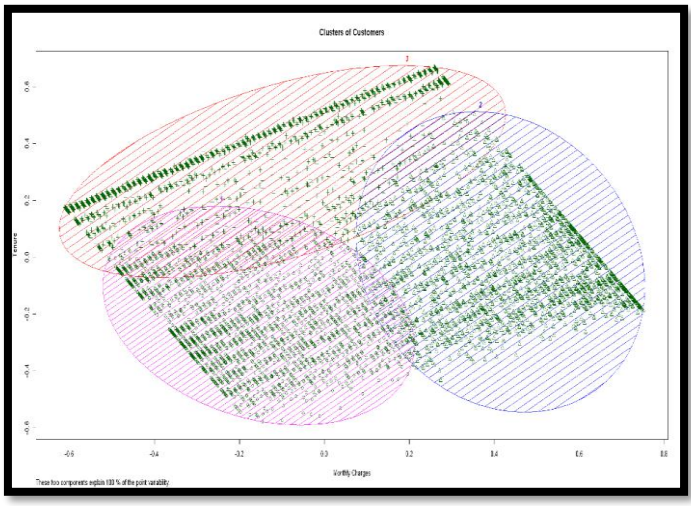

Figure 12: Showing clusters of customers based on importance using k-means algorithm

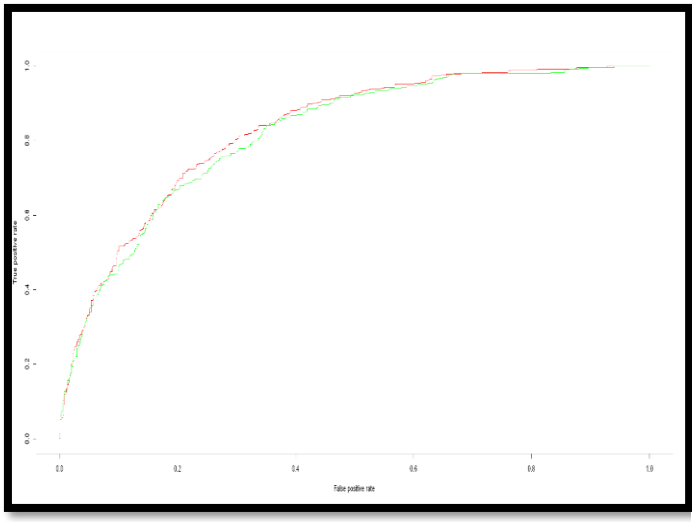

figure.14: Showing the ROC curve for Gradient boosting machines and logistic regression

Performance using ROC Curve: There is a comparison of different algorithms based on ROC curve which plots the graph based on false positive and true positive rate. We compared only logistic regression and gradient boosting machines as shown in the figure.13 to show they are almost similar because of low difference in their accuracies. That's why we didn't show all the ROC curves on one plot.

VIII. DISCUSION

In this section, there is a discussion based on the implications of the results we achieved in the previous section. we compare the tool which is better in terms of other whether the accuracy, precision, recall and specificity in this section.

The first classifier algorithm is Logistic regression which is one of the most common algorithms for classification type of problems in machine learning and basic algorithm in terms of complexity. But this algorithm has very poor recall 0.44 but it's better than Random forest because Random Forest has recall 0.33 lesser than Logistic Regression although random forest has good accuracy than Logistic Regression.

The first new method we used is Gradient Boosting Machine which is the advanced version of decision trees called it under the category of Ensemble Learning which is the combination of different algorithm for better result. So, this algorithm improves a little bit improvement in the accuracy which is 80 percent, but this algorithm has a problem of bad recall which is 50 percent although this algorithm is one of the most powerful algorithms it's not only used in classification problems but also in regression problems. It's also a powerful method for feature selection. The features selected by the gbm method are Contract, Monthly Charges, Total Charges, Tenure and Contract. So, company can put their focus on these categories for better performance in the area also increase great share of their market. E.g. if the customers are leaving because of internet which shows that they need to improve their internet service or decrease the cost of internet service or might give better rewards to attract new customer and prevents their own customer to leave the company.

The new novel approach we used for prediction of Customer Churn is implementing Artificial Neural Network. It's so far, the best model for this data set because it has greater recall which 0.73 as compared to others although it has better accuracy as well which is 84 percent. Although we tried to improve Neural Network as well but failed even increase the greater number of iterations by increasing the number of epochs. In the figure.15 which shows the number of epochs although it's increased to 500 but as the loss function is not decreasing in the upper half of the figure and accuracy is not increasing in the lower half of the graph.

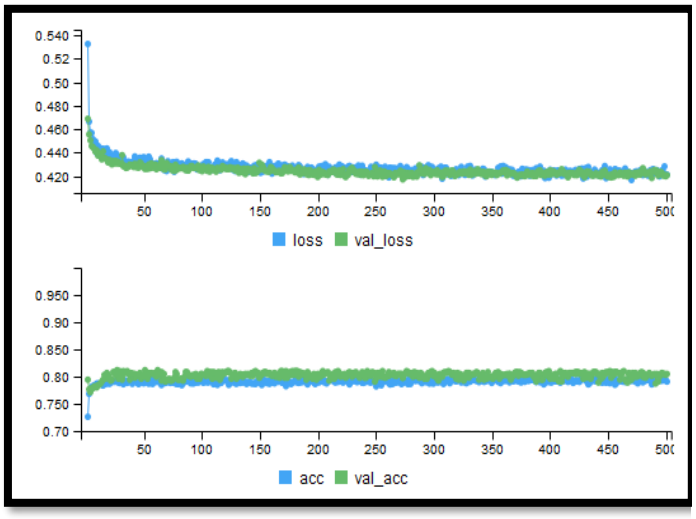

Figure.15: Neural Network Learning loss function vs accuracy

The second half of the project which is unsupervised clustering, so we compare the clustering with our business model as it's around 71 percent similar as the business model could change from different views and schemes.

IX. CONCLUSION & FUTURE WORK

While there has been work on the Customer churn analysis but still too less. The idea of customer churn is an important for business perspective. As the rise of job market in data science specially related to business intelligence, this project is a good way to start. Our model could be used in the industry for large and private dataset as the data for the telecom companies is private and secure.

The second important thing in the project is importance of customers based on tenure and monthly charges although this idea is unethical with respect to the humanitarian perspective but it's profitable for the companies. The limitations in the dataset includes the text so textual mining would help to identify the reason for the churn, also there is not time line available so understand the effect of churn on loss and profit of the company as both are also missing.

This Customer Churn analysis helps to understand the concept of churn and customer importance in telecommunication business.

X. ACKNOWLEDGEMENT

We would like to thank our supervisors Olga Baysal and Elio Velazquez for their valuable guidance and motivation which was essential for the progress and completion of our project.

REFERENCES

[1] Manpreet Kaur and Dr. Prerna Mahajan, "Churn Prediction in Telecom Industry Using R," International Journal of Engineering and Technical Research (IJETR), vol. 3, Issue.5, May. 2015

[2] Shin-Yuan Hung, David C. Yen, H. Wang, "Applying data mining to telecom, Expert Systems with Applications", 31 (2006) 515–524, Elseiver.

[3] V. Umayaparvathi, K. Iyakutti, "Applications of Data Mining Techniques in Telecom Churn Prediction", International Journal of Computer Applications (0975 – 8887) Volume 42– No.20, March 2012.

[4] Chunhua Ju, Feipeng Guo, "Research and Application of Customer Churn Analysis in Chain Retail Industry", International Symposium on Electronic Commerce and Security, IEEE, PP. 670-674, 2008

[5] Preeti K. Dalvi, Siddhi K. Khandge, Ashish Deomore, Aditya Bankar, Prof. V. A. Kanade, "Analysis of Customer Churn Prediction in Telecom Industry using Decision Trees and Logistic Regression", Symposium on Colossal Data Analysis and Networking (CDAN), IEEE, 2016

[6] N. Kamalraj, .A.Malathi, Applying Data Mining Techniques in Telecom Churn Prediction, in proc. International Journal of Advanced Research in Computer Science and Software Engineering, 10, October 2013

[7] IBM data analysis on customer support retrieved from https://www.ibm.com/communities/analytics/watson-analytics-blog/