Visualization And Model Development

Customer Churn Analysis

1.0 Problem Description

Customer churn is a term used when a customer or subscriber stop using the services of a particular business, also known as customer attrition and customer defection. One industry in which churn rates are particularly useful to is the telecommunications industry, because the cost of retaining an existing customer is far less than acquiring a new one.

This aim of this assessment is to train a machine learning model on the available data that will predict with a high accuracy which customers are about to churn, what is driving the churning of customers. **Exploratory Data Analysis**

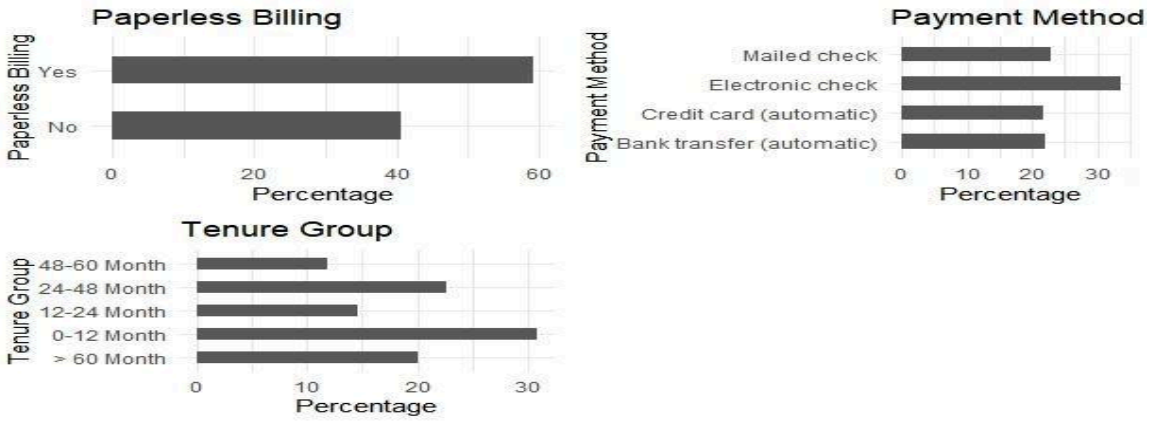


Fig.1.0. shows the Exploratory data analysis of the variables i.e paperless billing, payment method and tenure group.

From figure1.0 above, it could be noticed that the probability of a customer churning is high i.e paperless billing is a significant variable to customer churning.

Also, it could be noticed that customers with tenure group 0-12 month are more likely to churn compare to churn compare to tenure group 12–24 month, tenure group 24-48 month, tenure group 48-60 month and tenure group > 60 month.

In additional, it could be noted that the customers with electronic check are likely to churn compare to mailed check, credit card (automatic), and bank transfer payment method.

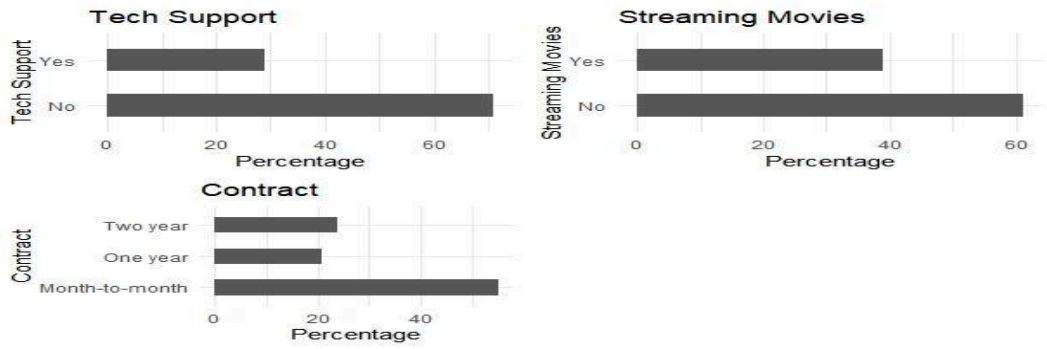


Fig.1.1. shows the Exploratory data analysis of the variables i.e tech support, streaming tv and contract.

From the figure1.1 above, it could be noticed that customer that subscribe for tech support are not likely to churn. i.e they are likely to continue in the company.

Also, it could be noticed that customers with contract month-month are more likely to churn compare to custom.

In additional, it could be noted that the customers with streaming movie subscription are not likely to churn. i.e they are likely to continue with the company



Fig 1.2. shows the Exploratory data analysis of the variables i.e multiple lines, online backup and device protection.

From the figure1.2 above, it could be noticed that customers with multiple lines are not likely to churn. i.e customers that subscribe for multiple line has higher chances of staying.

Also, it could be noticed that customers that subscribe for device protection have higher chances of not churning. i.e they are likely to stay with the company.

In additional, it could be noticed that the customers that subscribe for online backup are not likely to churn.

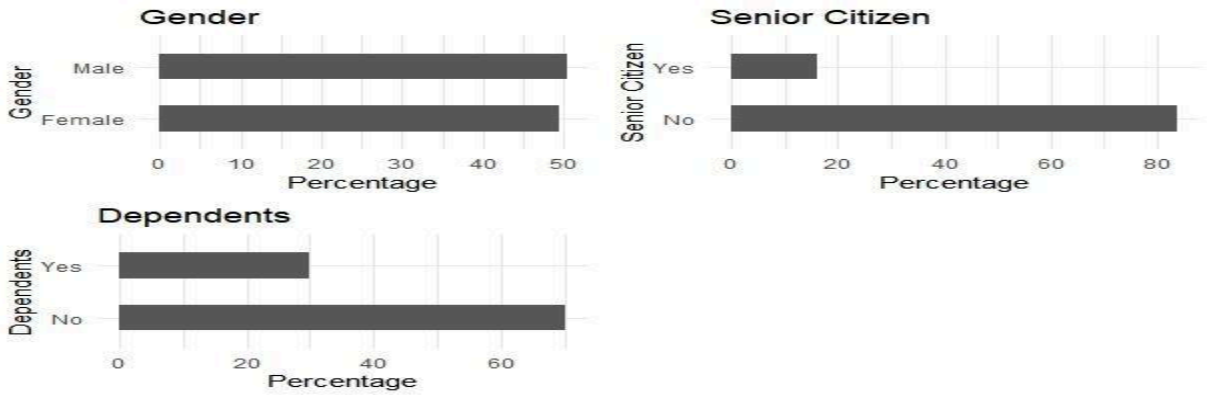


Fig 1.3. shows the Exploratory data analysis of the variables i.e gender, senior citizen and dependent

From the figure1.3 above, it could be noticed that the probability of a customer churning is the same for both males and female. i.e gender is not contributing to the probability of churning.

Also, it could be noticed that customers that subscribe for dependents are not likely to churn.

In additional, it could be noticed that the customers with senior citizen are not likely to churn.

HANDLING MISSING VALUES.

The handling of missing data is very important during the preprocessing of the dataset as many machine learning algorithms do not support missing values.

When dealing with missing data, data_scientists can use two primary methods to solve the error: **imputation or the removal of data**.

The imputation method develops reasonable guesses for missing data. It's most useful when the percentage of missing data is low. If the portion of missing data is too high, the results lack natural variation that could result in an effective model.

The other option is to remove data. When dealing with data that is missing at random, related data can be deleted to reduce bias. Removing data may not be the best option if there are not enough observations to result in a reliable analysis. In some situations, observation of specific events or factors may be required.

The method used in the dataset is a removal method (deletion), this is because the I needed to reduce the biased

Deletion

There are two primary methods for deleting data when dealing with missing data: listwise and dropping variables.

Listwise

In this method, all data for an observation that has one or more missing values was deleted. The analysis was run only on observations that have a complete set of data. Since the data set is not small, it would be the most efficient method to eliminate those cases from the analysis. However, in most cases, the data are not missing completely at random (MCAR).

INTERPRETATION OF CHURN ANALYSIS

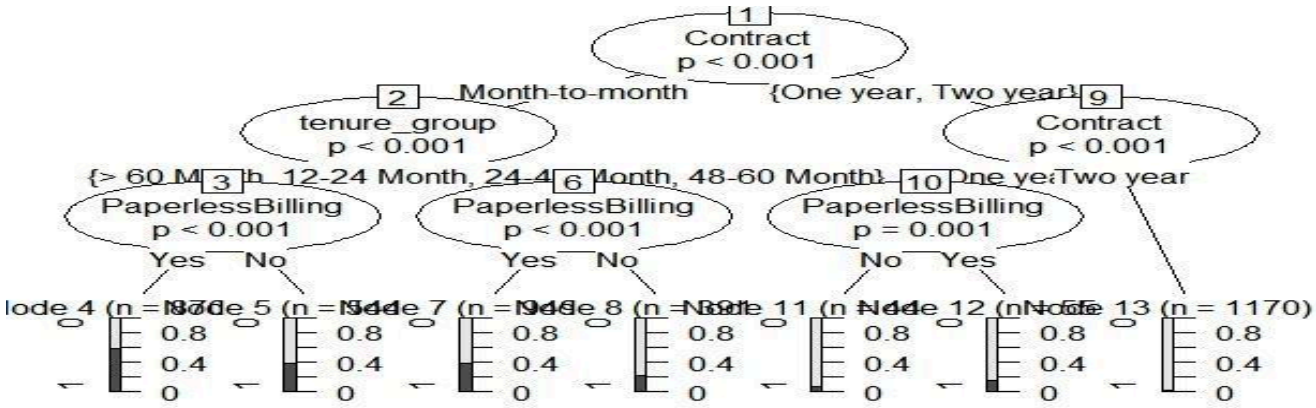


Fig 1.4. Decision Tree of Telco company

From fig 1.4 it could be noticed that the contract is most important variable here because it predicts the churn rate of every particular customer in the company.

It explained that customers with contract within 12 months (month-month) are more likely to churn compare irrespective of the service they are subscribed to compared to customers with contract of one year, two years who are very less likely to churn irrespective of the services they subscribed to.

Table 1.0 Confusion Matrix for Decision Tree

	Actual	
Predicted	0	1
0	1414	332
1	134	228

Table 1. 0 shows the confusion matrix that illustrate the actual and predicted number of customers that are likely and less likely to churn.

It can be seen that 1414 customers are less likely to churn while 228 customers are likely to churn.

Table 1.1 Decision Tree Accuracy Table

Decision Tree Accuracy = 0.778937381404175

Table 1.1 illustrate accuracy of the decision tree model with about 77.89% accurate to predict the churn analysis of telecommunication customers.

Questions

1. The effectiveness of your churn analysis: What was the percentage of time at which your analysis was able to correctly identify the churn? Can this be considered satisfactory outcome? Explain why or why not;

Ans: The percentage was about 77.89% to confirm that my analysis correctly identifies the churn.

Yes, it can be considered satisfactory because the accuracy was approximately 80% with 95% confidence interval.

2. Who is churning: Describe the attributes of the customers who are churning and explain what is driving the churn?

Ans: customers with contract within 12 months (month-month) are churning at high rate.

The churn is driven by the contract because its most important variable of all. It's the contract that determine the churn rate of customers.

3. Improving the accuracy of your churn analysis: Describe the effects that your previous steps, model development and handling of missing values had on the outcome of your churn analysis and how the accuracy of your churn analysis could be improved.

Ans: It has greatly help in cleaning and waggling of data in other have a good insight about the data. It has help in standardizing the data in other to avoid spurious result or wrong prediction.

It could be improved if we had carried out other approach in solving churn analysis like logistic regression and Random Forest method then compare the accuracy of each approach to pick best prediction.