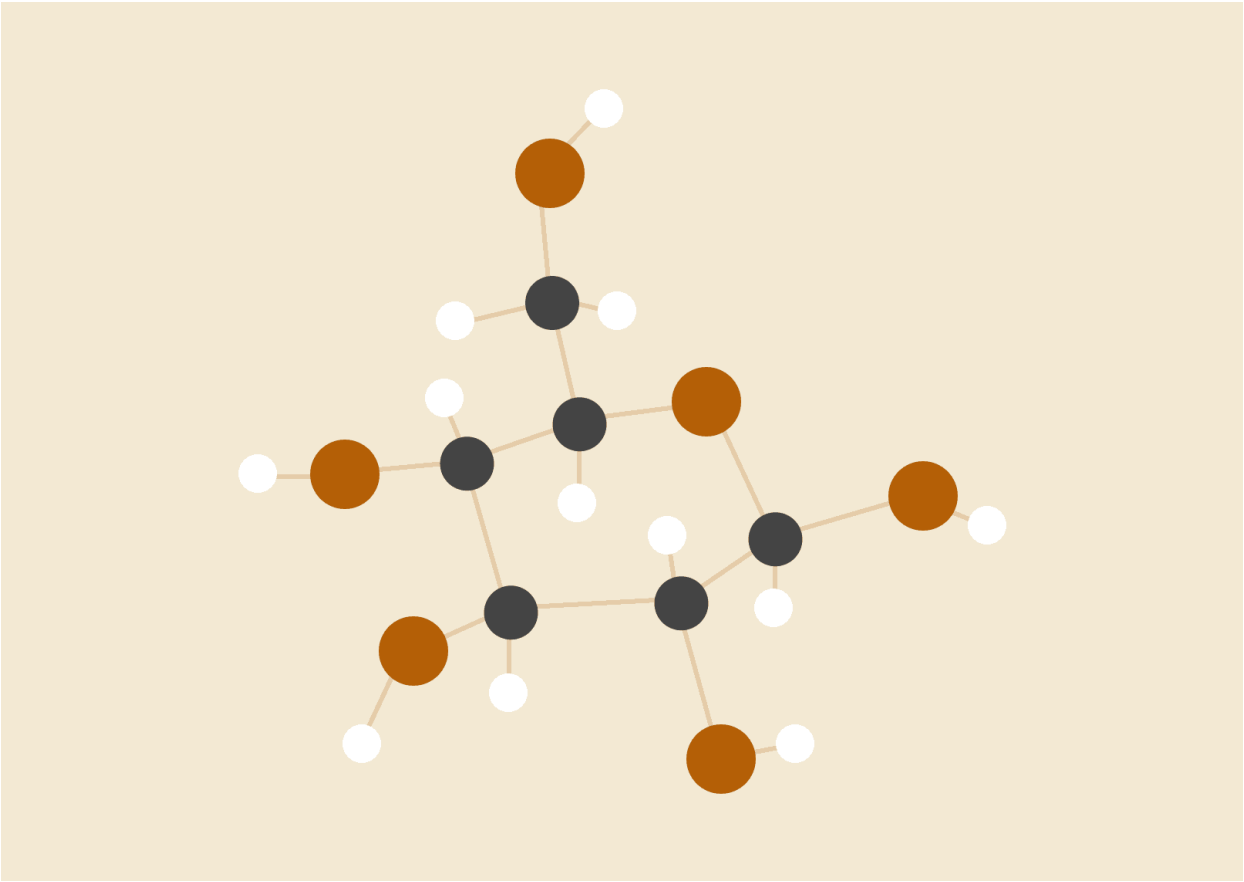
THE TELCO INDUSTRY DEEP LEARNING AND NEURAL NETWORK MODEL TO PREDICT CUSTOMER CHURN REPORT



PREPARED BY THE FOLLOWING DATA SCIENTISTS

Janet Cheserem

Colleta Nandutu

Joseph Leshakwet

TABLE OF CONTENTS

A BRIEF LOOK AT CUSTOMER CHURN RATE IN THE TELCO INDUSTRY

INTRODUCTION

The main aim of the project is to examine the Teleco Dataset obtained from Kaggle, by performing analysis on the industry customers and seek to understand behavioural changes and what is influencing it. The main objective of the study is to understand the metrics such as the percentage of the customers who discontinued their subscription within a particular time frame. High churn rate means either dissatisfaction with the industry's services or other underlying issues. This may potentially be problematic for the Telco industry and its services.

It is crucial to understand and keep track of the company's customer behavior by looking out for pattern and trends especially in those customers with high rate of churning. This will enable decision makers to develope strategies to curb or at least reduce churn rate and potentially retain customers.

METRICS OF SUCCESS

This research or analysis will be considered a success when it is able to predict the number of customers churning services of the Telco industry after a few months.

Also if it can give answers as to why the customers are churning, age of the churners, demographic, number of customers churning and gender.

RESEARCH DESIGN

The research focuses on the following steps to comprehensively analyze the Telco Dataset and extract relevant information from it.

1. Business Understanding
2. Data Understanding
3. Data Preparation
4. Analysis
5. Exploratory Analysis
6. Implementation
7. Challenging The Solution
8. Conclusions
9. Recommendations

Business Understanding

Connectivity is the most important part of any business and having a well connected customer base gives any business an hedge against potential rivals. The Telco industry thrives in having more subscribers to its services especially long term ones who would guarantee long term revenue streams.

Project Objectives

The research seeks to answer the following questions:

1. How long as an individual be a customer?
2. Does he/she have a contract?
3. What payment methods is most preferred?
4. Do they have partners or no?
5. What age and gender?

Data Mining Obectives

To interpret and analyze the data, the following steps are taken:

1. Exploratory Data Analysis
2. Univariate, Bivariate and Multivariate Analysis
3. Implementation of the solution
4. Challenging the solution

Tools For Implementation

To successfully work on the project, the data scientist would require the following tools:

1. Relevant Datasets
2. Importing Libraries
3. Using relevant Algorithms

Data Understanding

Before working on the data,it is important to understand the contents of the dataset provided and to establish the relevance of the information contained in the dataset to the research.This therefore involved the exploration of data

Data Collection

The Telco Dataset was downloaded from Kaggle.

Data Description

By checking the dataset description, it is found to contain 7043 rows and 21 columns

Data Exploration

Exploration is important in any research as it helps identify errors, checking for any unique values which helps in the preparation of the data for data cleaning.

Data Quality

After previewing and exploring the data, data cleaning processes are required to make sure that the data being used is of quality and one that will give the most accurate results possible.

Data Preparation

This is the process of making sure the data achieves and maintains quality for the final use in the analysis.

Data Cleaning

There are key steps taken to ensure that the data being worked on is clean and they are as follows:

1. **Null values**
2. **Data in the wrong format**
3. **Wrong data**
4. **Duplicates**

False output shows that there are no missing values; but the TotalCharges column has been changed into the right datatype

There are 11 missing values in the TotalCharges column

The missing rows can be dropped

The TotalCharges column can be converted to float from object

There are null values in the TotalCharges column

The output is zero; meaning there are no missing values

The customerID column was dropped because it was unnecessary in the analysis

The numeric values of the Senior Citizen variable was converted into categorical

'No' represents 0 and 'Yes' represents 1

Exploratory Analysis

EDA is used to investigate the dataset and summarize the key insights. It gives the basic understanding of the data, it's distribution, and relationship between variables. Grapghs and python functions are used to explore the data. There will be two type of analysis; univariate and bivariate analysis.

Univariate Analysis

The following observations were made when the analysis was conducted:

The data is higly imbalanced; the ratio is 73:27

The churners are 27% while the non-churners are 73%

There is imbalance in the target variable

The numbers of customers who churn is low as compared to those who don't churn

There are no outliers in the TotalCharges, MonthlyCharges and tenure columns

Majority of the customers were charged between 0-2000

The distribution shows skeweness to the right(positive skewed)

The Monthly charges lie towards high values showing a positive skew in the variable.

Most customers pay a monthly charge of between 18 and 30.

Majority of the customers are young.

33.6% of the customers use electronic check for payments

22.8% of the customers use mailed checks

21.9% and 21.6% of the customers use bank transfers and credit cards respectively

Bivariate Analysis

The following was observed after performing the Bivariate analysis of the data:

The churn rate is low in senior citizens.

The churn rate of both male and female is the same; gender has not effect on churn rate

Churn rate among those who have no partners is higher compared to those with partners.

Those who have no dependents are likely to churn services as compared to those with dependents.

Customers subscribed to phone services are more likely to churn as compared to those who are not subscribed.

The churn rate is not affected by the number of lines owned by a customer.

Churn rate is high in customers who use Fibre Optic compared to those who use DSLand those without any internet service.

Customers having no online security tend to have a higher churn rate.

Customers without online backup services are likely to churn.

Customers without device protection are likely to churn.

Customers are likely to churn where there is no Tech Support.

Customers with monthly subscription tend to churn more as compared those with 1-2 year contract.

Customers who use paperless billing tend to churn more than those using paper billing.

Churn rate among those customers using Electronic checks as a payment method is higher.

Customers with higher Monthly Charges are also more likely to churn

There is a strong positive correlation between Total Charges and Tenure.The longer a customer stays the more they pay.

There is a weak positive correlation between tenure and Monthly charges.

Implementing the Solution

Modeling in this phase will include selecting a deep learning technique, designing the test, building the model and assessing model.

1. Data Pre-Processing

The previous dataframe was transformed into numerical variables

2. Modeling

After importing the required Libraries, the dataset was split into X and y. y-churn and X are the other independent variables.

The observations made are as follows:

- Sequential model was created and hidden layers were added.
- Used relu activation function in the hidden layers and sigmoid on the output
- The model expects rows of data with 19 variables
- The first hidden layer has 15 nodes; the second hidden layer has 10 nodes and the output has just one node.

- Some additional properties were specified as they were required for training the network.
- The loss function (binary_crossentropy),the optimizer(adam) was specified and the classification accuracy defined by the metrics argument
- Having been defined and compliled, the model is ready for computaion as well as ready to be fitted on the dataset.
- Use the .fit() function to fit the dataset. The epoch was set to 15, and set batches to 10 ie the number of samples considered by the model within an epoch before weights are updated.
- Trained the neural network on the entire dataset
- Used the .evaluate() function.
- The function returned two variables

1. Loss of the model on the dataset (1.6105)
2. Accuracy of the model on the dataset (79.17%) using 'option' optimizer

CONCLUSION

Customer churn is negatively impacting the company's profitability hence some strategies might be considered to curb the rate of churn. The following points are worth noting:

Churn rate is with those customers who use fibre optic compared to those who use DSL and those without internet services.

Customers without online security tend to have high rate churners.

Customers without only backup services are more likely to churn.

Customers churn rate is high if there is no Tech support.

Customers with monthly subscription tend to churn compared to those with 1-2 years contracts.

Customers who use paperless billing tend to churn more than those using paper billing.

RECOMMENDATIONS

The Telco industry should consider the recommendations to reduce customer churn rate:

1. Improve on fibre optic speed and connectivity
2. Enhance the security measures
3. Offer encrypted data storages services
4. Ensure timely responses to the customer queries
5. Offer affordable monthly packages