Department of Computer Science and Engineering

Institute of Engineering

JK Lakshmipat University, Jaipur, India

Pravesh Bisaria

praveshbisaria@jklu.edu.in

Abstract

Customer Retention is a very important aspect of any business-like tele-communication, retail, e-commerce, etc. Customer retention rate has a direct effect on the customer's acquisition cost, and understanding the exact value of a potential customer churn would allow the business to develop a good customer relationship. Predictive Modeling can use used to predict the customer churn of any company and help them to make better strategies for serving their customer and retain their loyal customers, The objective of the case study is to implement logistic regression methods to predict customer churn and also to analyze churning and non-churning consumers by using telco customer churn dataset from Kaggle.

Improve customer retention using data mining

A case study on Customer Churn

# CASE SYNOPSIS

Customer attrition or churn is simply calculated as the number or fraction of the losing customers or clients by any company within a particular amount of time subtracting recent additions during that time frame. It's the unit of measurement for how successful or not you've kept your client engaged for any company.

Churn reduction is very important for a company. According to the researchers a high percentage of revenue loss face by the company is because of customer churn. In the telecom business, where churn is important and multiple competitors can provide similar solutions, companies need to differentiate their services from other competitors and retain their customers.

Telecom firms, Internet service providers, insurance agencies and e-commerce companies typically use consumer churn rates as their business growth indicators as the cost of holding an existing client is lower than the cost of getting the new customers. Organizations in these fields also have customer service teams that are working to get back lost customers, so recovered long-term loyal customers can be lot more worth than recently hired customers.

There are two types of churn which are possible in the company- one is voluntary and other is involuntary. Voluntary churn us the result of the customer’s decision to switch the service provider while Involuntary churn occurs due to the situations like transfer to another place, company shut down, etc. The main focus of the company is the voluntary churn and the reason behind that decision. Data Analyst research on the voluntary churn and the company’s client relationship that companies control, including the user interface, after-sales assistance or other company policies.

Machine Learning techniques are very useful to understand the customer behaviour by identifying the factors which are responsible for to make them leave the company. These features are very helpful in predicting the future churn and also help us to take necessary measure before actual churning of the customers. Using Predictive Analytics and ML algorithms company can easily analyze the reason behind the situation and came up with the proper solutions. Telecommunication service operators that believe in the power of predictive models and data-driven analytics are more likely to turn down their customer attrition rates that the ones that do not.

This case study can be used in Predictive Analytics courses. The case is suitable for teaching classification techniques like logistic regression and its application in business problems. The problem mentioned in the case study is of the classification of customer churn using the telco customer dataset for the Kaggle website.

# Learning Objective of the case study

The case may be used in Predictive Analytics Modeler course of Big Data Analytics. The learning objectives are:

* 1. Understand the Customer churn and the problems associated with it.
  2. Learn how Customer attrition rate affects company customer relationship.
  3. Demonstrate the applications of Logistic Regression in solving classification problems.
  4. Identify and apply data pre-processing techniques to be utilise by logistic regression algorithm.
  5. Understand the effect of influential data points.
  6. Understand the challenges associated in developing a logistic regression model with class imbalanced dataset.
  7. Illustrate descriptive analytics to infer conclusions about multi-dimensional dataset.
  8. Develop a multivariate classification model to predict a churn using logistic regression.
  9. Learn different methods to validate a logistic regression model.
  10. Apply the recursive feature elimination to construct the model.
  11. Demonstrate knowledge about the significance of customer churn using machine learning by writing a research paper.

# Study Questions

* 1. How would you define Customer Churn and Customer churn rate?
  2. In the current business scenario, how important is knowing the status of the customer engagement with the company and its services?
  3. What are the indicators what shows signs of a possible churn in the company or organisation?
  4. How can the Machine or Data driven approach help to manage the churn?
  5. What data is needed for churn management analysis?
  6. Identify the features in our dataset that should be used for predicting whether or not a customer will be churned? What pre-processing techniques are used with the variables to use them while building a model?
  7. How different features in the dataset are correlated to target variable i.e. churn/non-churn? Also give the Summary Statistics of the variables.
  8. Discuss the customer attrition in the dataset. How different features are distributed in customer attrition?
  9. The customer churn number is very less than non-churn customers. What kind of modelling process one can expect when cases in one class in binary classification are much lower than the other class? How can one handle datasets with imbalanced classes?
  10. Use the dataset to develop a logistic regression model that can be utilized to predict customer attrition. Comment on the model.
  11. How would you intercept sensitivity, specificity and model accuracy of the model developed in Question No. 10? How it is determined?
  12. Use Recursive Feature Elimination approach to the construct the model developed in question 10 and choose the best or worst features.
  13. What are the limitations of the model developed and how should we handle the limitation?
  14. Recommend some practices that will help lower the customer lifetime value and delight our existing customers.

# Suggested Answer of Study Questions

* 1. How would you define Customer Churn and Customer churn rate?

Answer: Customer churn is the rate at which your current customers stop working with you. In other words, it is the rate at which your existing customers stop using product and service of your company.

Customer churn rate can be calculated using:

*%* ***Customer\_churn\_rate*** *= Customer lost during time period t / Total customer at the start of time period*

***Customer lost during time period t*** *= Total customer at the start of time period t + new customer gained during the time period t – total customer at the end of the time period t*

For e.g., you have 100 clients at the end of month March. You gained 20 new clients by the end of month April. The total number of clients at the end month April stood at 116.

The customer lost during month of April => 100 + 20 – 116 = 4

The customer\_churn\_rate for the month April => 4/100 = 4%

* 1. In the current business scenario, how important is knowing the status of the customer engagement with the company and its services?

Answer: Customer engagement is the continue interaction between company and its customer. It may be in the form of any service provided by company and chosen by customers. It demonstrates that the customer in some sense paying for a relationship with company, which means they must think you have something potential valuable to them.

It gives us huge insight into the requirements that the customer has the exact problem you need to fix and the challenges you will need to overcome to generate relevant revenue.

Customer Engagement can be done in many ways some of them are:

1. **Maintain social media presence:** Many big companies and firms uses social media platforms like Facebook, Instagram, Twitter, etc to engage their customers and connect them virtually with the company and promoting their new services and offers to not only existing customers and also to attract the new ones.
2. **Carry out promotional campaigns:** Another way to maximize customer engagement is to target the right audience. For this different campaign drives and the sessions can be organized with customers to consider their actions and demand and to develop new creative products and services according to their needs.
3. **Personalization:** Personalized emails and newsletters can be sent to specific clients which may be interested in subscribing your services. These things can easily grab their attention and generate interest in your service and product.
4. **Empower your employees:** Empowering the employees with valuable tools and skills can help them to maintain a good customer relationship with the company. Necessary training can be provided to employees to engage more customers in an effective manner.
   1. What are the indicators what shows signs of a possible churn in the company or organisation?

Answer:

The indicators which shows signs of a possible churn in the company or organisation are:

* 1. **Usage level:** Customer churn is often is the result of the decreasing usage levels by the customers. Checking up on user behaviour, such as login, helps us identify high -risk customers when they are about to churn.
  2. **Feature Adoption:** Every product has some features which differentiates it from the other competitors. If certain customers are not using that feature there may be a possibility of churning of customers because of that feature.
  3. **Consumer's KPIs:** If our product is not capable of satisfying customers with their KPIs, then the probability of them churning up is going to increase.It is advised to take regular feedbacks through polls, surveys, or phone calls from the customer about the existing services.
  4. **Financial health:** Financial health: It is measured on a monthly basis, whether payments are made on time or delayed, and whether payment options are valid.
  5. How can the Machine or Data driven approach help to manage the churn?

Answer: A good attrition management system should be capable of identifying early signs for possible churners. If we know early that the client is about to leave the service of your company, possible measures and actions can be taken to avoid this from occurring. And that's where data analytics come into play and perform a transformative role. This can be done by:

* + 1. **Identifying the relation among known negative indicators and their impact:** analyzing past data on negative customer interactions and how consumers from different categories have reacted to them will help to create a reliable model that can predict future churn. This can be used to monitor similar triggers experienced by existing clients to evaluate how they are prone to respond.
    2. **Knowing past-based fading behaviour patterns:** Past data can be used to forecast non-trigger-driven churning. Sorting existing clients, based on a 360º comparison of them along with churned traditional customers, will help to predict high-risk clients. These behaviour patterns could include a decrease in the number of segments I am purchasing, a decrease in the frequency of transactions, a decrease in the redemption of my reward points, etc.
    3. **Recognizing high risk segments:** It is important to use something like a decision tree classifier to classify the client population in specific behavioural traits, thereby allowing the identification of segments or groups with high risk of churn. Such groups can then be listed as high-risk groups, and consumers belonging to them can be handled properly.
  1. What data is needed for churn management analysis?

Answer: While handling customer behaviour in the future, more and more information can be utilized, the smarter it would be. Having a 360° view. including all levels of customer engagement with your company would make it easier for you to have a comprehensive approach. The figure below illustrates a few of the key sets of data that can be integrated.

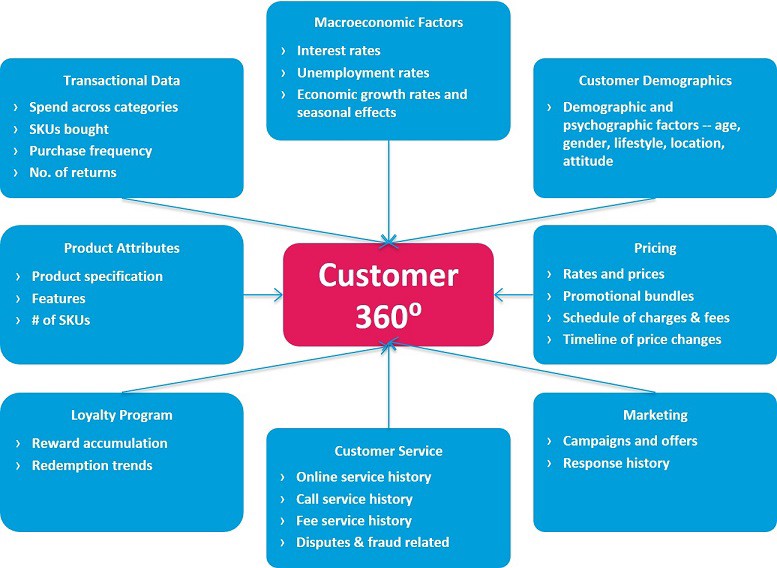


Figure 1 Possible datasets to Build a 360° Customer View

* 1. Identify the features in our dataset that should be used for predicting whether or not a customer will be churned? What pre-processing techniques are used with the variables to use them while building a model?

Answer: Almost all the features present in our dataset in necessary while building a machine learning model to predict the customer churn. The variables are:

|  |  |  |
| --- | --- | --- |
| S No. | Variable | Description |
| 1 | gender | Client is male or a female |
| 2 | SeniorCitizen | Client is a senior citizen or not -- (1, 0) |
| 3 | Partner | Client with a partner or not -- (Yes, No) |
| 4 | Dependents | Client with a dependents or not -- (Yes, No) |
| 5 | tenure | No. of months the client has stayed |
| 6 | PhoneService | Client with phone service or not -- (Yes, No) |
| 7 | MultipleLines | Client with multiple lines or not -- (Yes, No, No phone service) |
| 8 | InternetService | Client’s internet service provider -- (DSL, Fiber optic, No) |
| 9 | OnlineSecurity | Client with online security or not -- (Yes, No, No internet service) |
| 10 | OnlineBackup | Client with online backup or not -- (Yes, No, No internet service) |
| 11 | DeviceProtection | Client with device protection or not -- (Yes, No, No internet service) |
| 12 | TechSupport | Client with tech support or not -- (Yes, No, No internet service) |
| 13 | StreamingTV | Client with streaming TV or not -- (Yes, No, No internet service) |
| 14 | StreamingMovies | Client with streaming movies or not -- (Yes, No, No internet service) |
| 15 | Contract | Client’s Contact tenure -- (Month-to-month, One year, Two year) |
| 16 | PaperlessBilling | Client with paperless billing or not -- (Yes, No) |
| 17 | PaymentMethod | payment method used by client -- (Electronic check, mailed check, Bank transfer (automatic), Credit card (automatic)) |
| 18 | MonthlyCharges | Monthly amount charged to the client |
| 19 | TotalCharges | Total amount charged to the client |
| 20 | Churn | Client is churned or not (Yes or No) |

There are two types of features are present in our data one is categorical other is continuous. For the former Label Encoding can be used to convert each value in the column to a number. LabelEncoder in Python is used to do this task and for the Feature Scaling can be used to standardize the numeric columns in a fixed range, so that the model does not assigns higher weights to greater values and consider value of small values lower. For this task StandardScaler in used in Python to transform the numeric features.

* 1. How different features in the dataset are correlated to target variable i.e. churn/non-churn? Also give the Summary Statistics of the variables.

Answer:

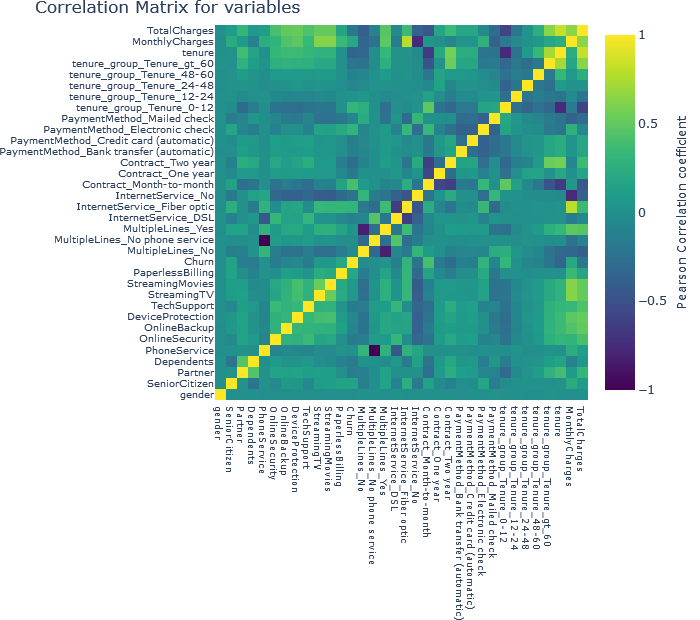


Figure 2 Correlation Heatmap of Variables

The figure shows the correlation matrix of the variables and how they are correlated to each other. The variables like month to month contract, absence of tech support and online security seem to be positively correlated with variable churn. While, variables two-year contract, tenure seem to be negatively correlated with churn. Interestingly, services such as Online security, streaming TV, online backup, tech support, etc. without internet connection seem to be negatively related to churn.

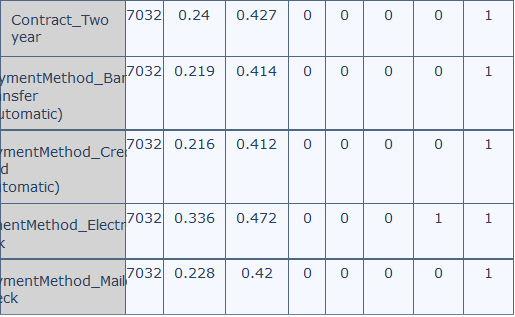
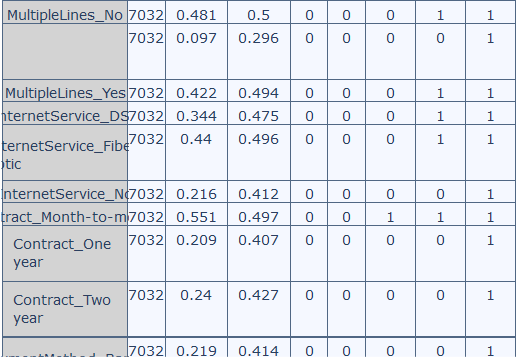
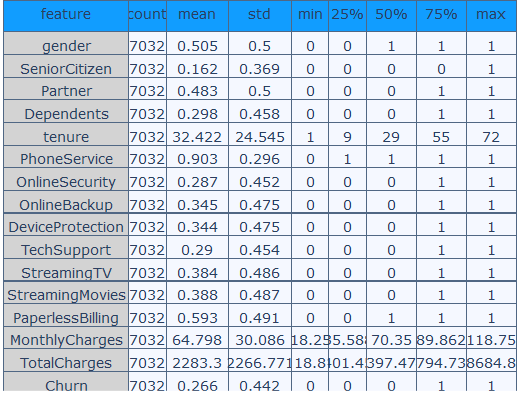


Figure 3 Summary Statistics of the Variables

The above figure shows the Statistics Summary of the Variables to be used in making the model to predict the churn.

* 1. Discuss the customer attrition in the dataset. How different features are distributed in customer attrition?

Answer: The below figure shows the customer attrition i.e. churn in the dataset.

1: represent the percentage of churn customers, while

0: represent the percentage of non-churn customers.

It is observed that the percentage of non-churn customers is far higher than that of churn customers. The percentage of churn customers is 26.2 per cent, while the number of non-churn customers is 73.4 per cent.

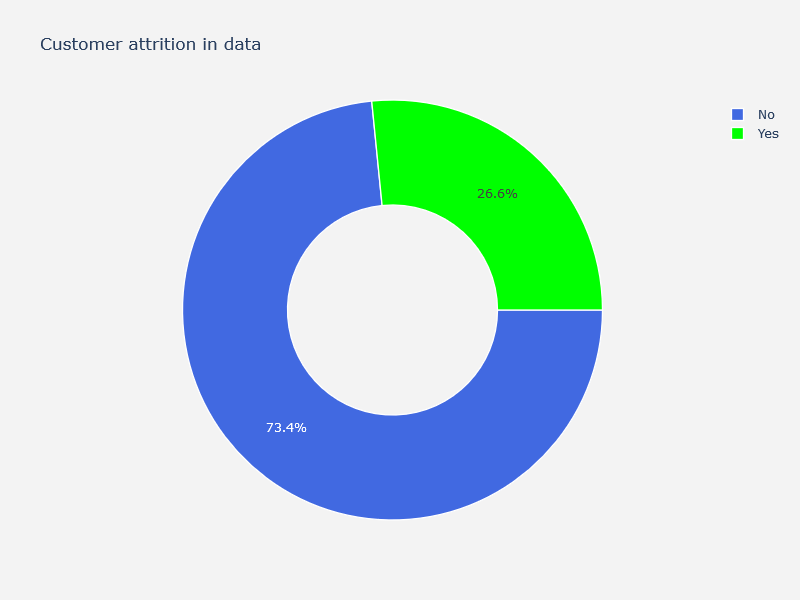


Figure 4 Customer Attrition in the dataset

The different features distribution in customer attrition can easily be analyzed by plotting them with customer attrition.

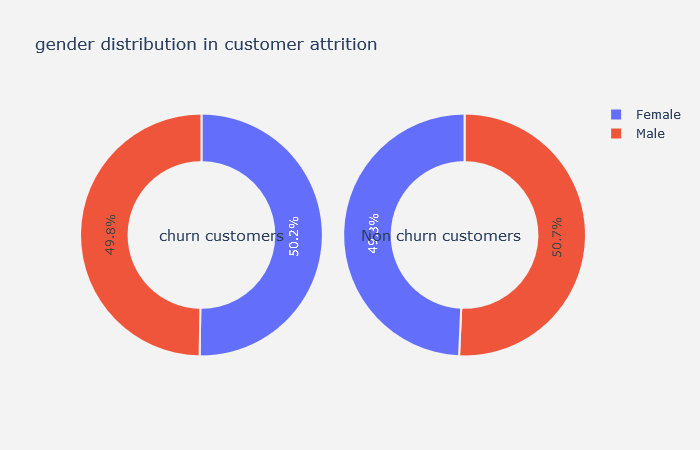


Figure 5 Gender distribution in Customer Attrition

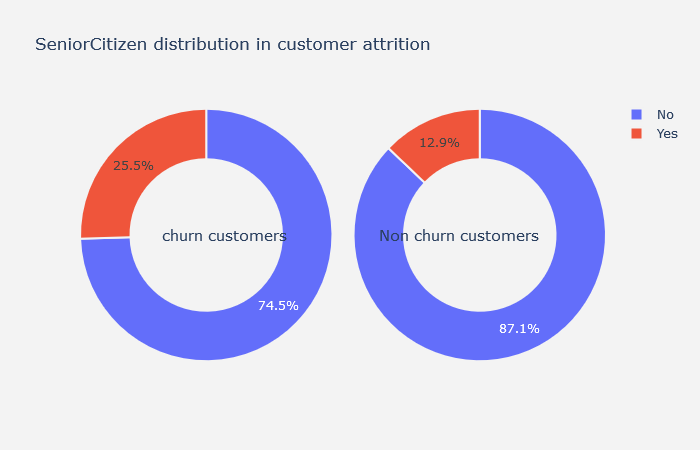


Figure 6 SeniorCitizen Distribution in Customer Attrition

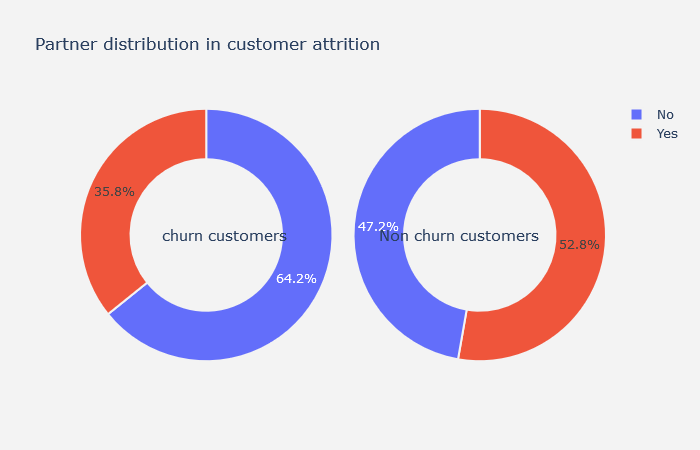


Figure 7 Partners distribution in Customer Attrition

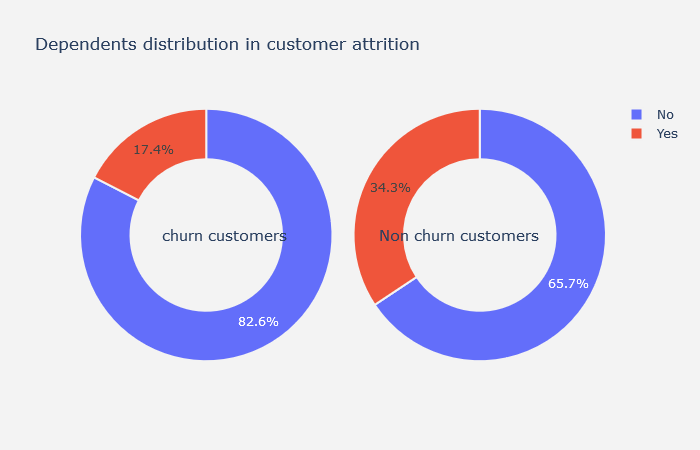


Figure 8 Dependents distribution in Customer Attrition

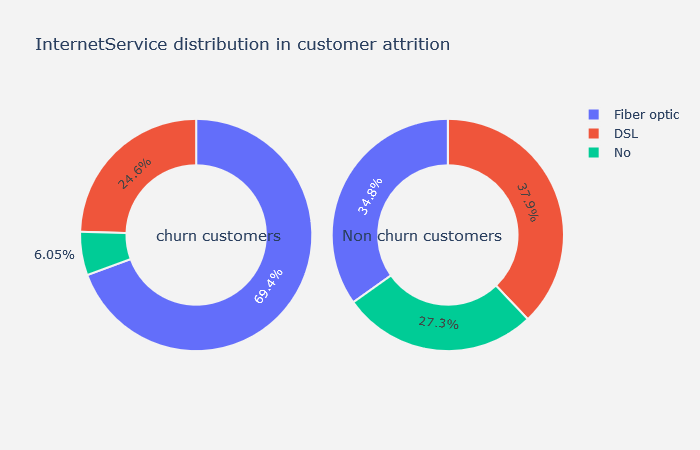


Figure 9 InternetService Distribution in Customer Attrition

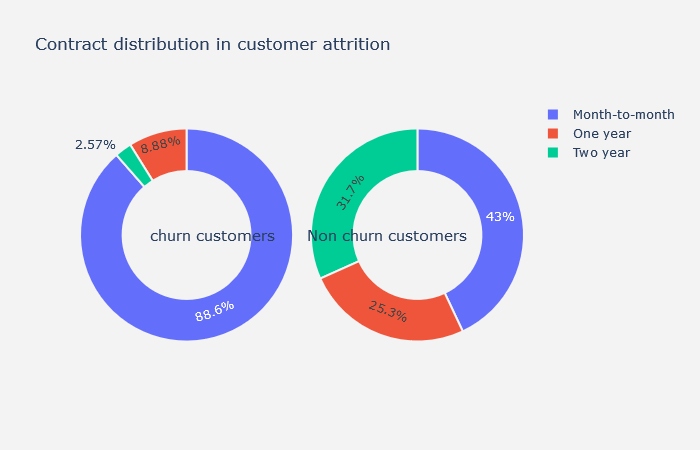


Figure 10 Contract Distribution in Customer Attrition

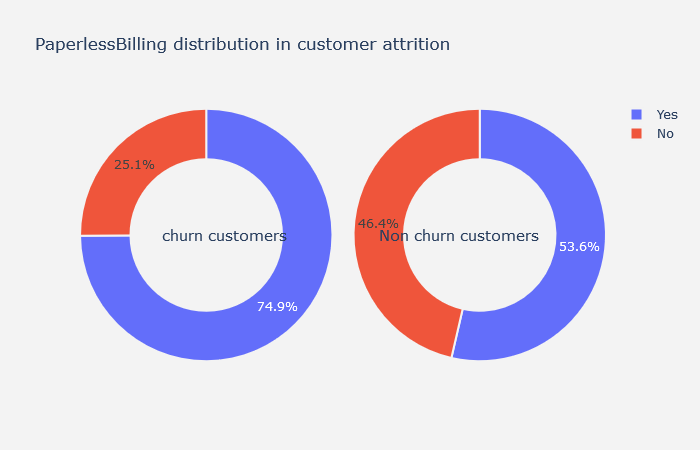


Figure 11 PaperlessBilling Distribution in Customer Attrition

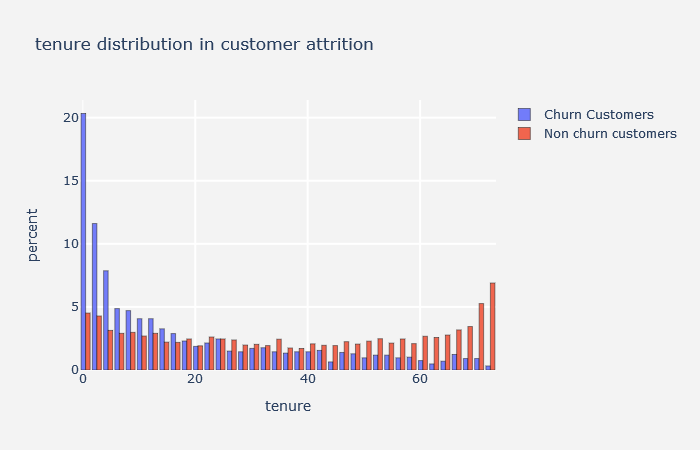


Figure 12 Tenure Distribution in Customer Attrition

The following insights can be drawn from all the graphs:

1. Gender-In the case of males and females, the churn percentage is almost similar.
2. In case of Senior Citizen, the churn percentage is quite high.
3. Customers with Partners and Dependents shows low churn rate then those who don't have Partners and Dependents.
4. Churn rate is even higher for Fiber Optical InternetServices.
5. Large percentage of clients with a monthly subscription have been left out compared to clients with a one-or two-year contract.
6. Customers with paperless billing option have high churn percentage.
7. Churn percentage is higher during the starting of the tenure.
   1. The customer churn number is very less than non-churn customers. What kind of modelling process one can expect when cases in one class in binary classification are much lower than the other class? How can one handle datasets with imbalanced classes?

Answer: In most of the real-life scenario when we analyse classification problems, we may have imbalanced dataset. In imbalanced dataset the one of the binary class is more than the other in dataset, like in our dataset the number of churn customers are very less than the number of non-churn customers. This can create problem while model development. The model may become bias towards the dominant class.

When we have class imbalanced dataset, following approaches can be used:

1. **Random oversampling of the minority class:** Random oversampling simply reproduces the data points of the minority class randomly. Random oversampling is known to increase the probability of model over-fitting.
2. **Random under-sampling of the majority class**: a basic under-sampling method is often used to under-sample the majority class randomly and uniformly. It might possibly result in the loss of information.
   1. Use the dataset to develop a logistic regression model that can be utilized to predict customer attrition. Comment on the model.

Answer: This question requires to develop a logistic regression model with Churn as the response variable (dependent variable) while the other variables as explanatory variables (independent variables).

The first step is to create a training and testing dataset. In order to do so using Python, one has to use train\_test\_split () function from the sklearn.model\_selection library of python. The test size is set 0.25 of the entire datasets. The train, 75% data will be used for model building. The remaining 25% data is used for model validation. The logistic regression function of sklearn.linear\_model is used to develop the data and predict the values using test data and then the determine the accuracy of the model. The Classification report of the Logistic Regression model developed in python is below:

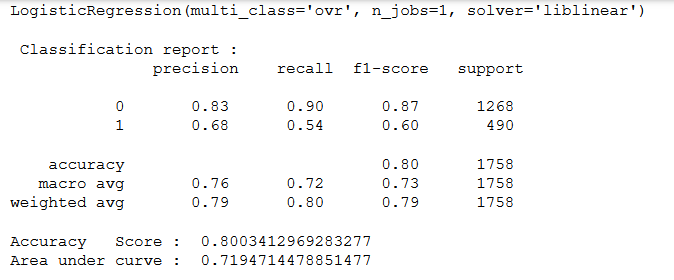


Figure 13 Classification Report of Logistic Regression Model

The Feature Importance with their weights can also be calculated and plotted using python.

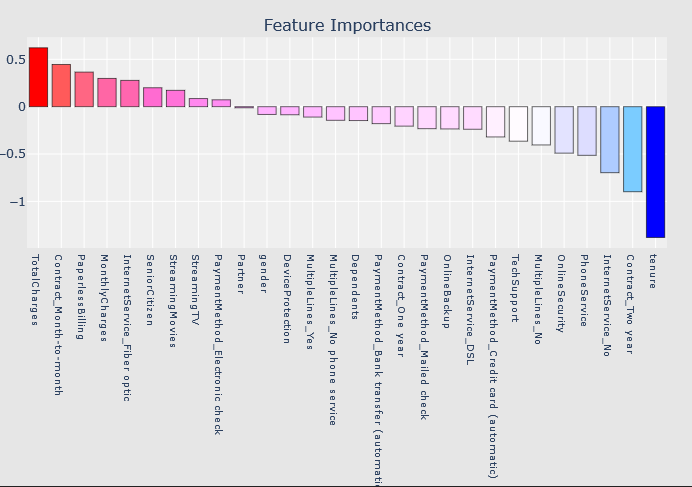


Figure 14 Feature Importance

* 1. How would you intercept sensitivity, specificity and model accuracy of the model developed in Question No. 10? How it is determined?

Answer: The classification table (also known as confusion matrix) from the above logistic regression model developed is used to determine the accuracy parameters.

Confusion Matrix and ROC Curve of the Logistic Regression Model is below:

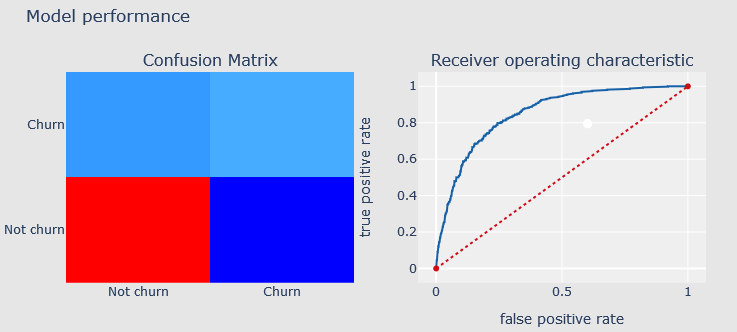
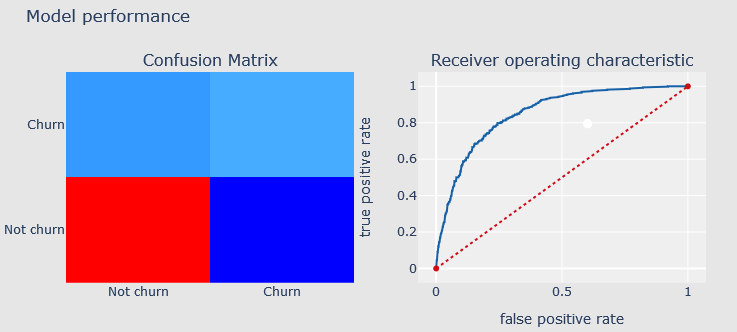
 

Figure 15 Confusion Matrix Figure 16 ROC Curve

Sensitivity, specificity and model accuarcy is calculated on the test dataset, i.e. 25% dataset, which was used to test the model.

* 1. Sensitivity is defined as the probability that predicted class 1 when observed is 1.

**Sensitivity** = = = 32%

* 1. Specificity is defined as probability that the precited class is 0 when the observed class is also 0.

**Specificity** = = = 85%

* 1. Model Accuracy is defined as correct classification of the observed class using the developed model.

**Model accuracy** = = = 78%

Sensitivity shows the true positive rate and the specificity shows the true negative rate. That is sensitivity is the model’s ability to correctly classify positive given the observation is positive, whereas specificity is the model’s ability to correctly classify negative given the observation is negative.

* 1. Use Recursive Feature Elimination approach to the construct the model developed in question 10 and choose the best or worst features.

Answer:

Recursive Feature Elimination (RFE) technique again and again develop a model and picks out the best or worst performing variables, keeping aside the other variables and repeats the same process with all the variables. The process is carried out until the whole dataset is finished. RFE’s goal is to select variables/features by recursively using small sets of variables

RFE module from sklearn.feature\_selection is used for recursive feature elimination. The number of features used is 10 and the model is logistic regression which is constructed recursively.

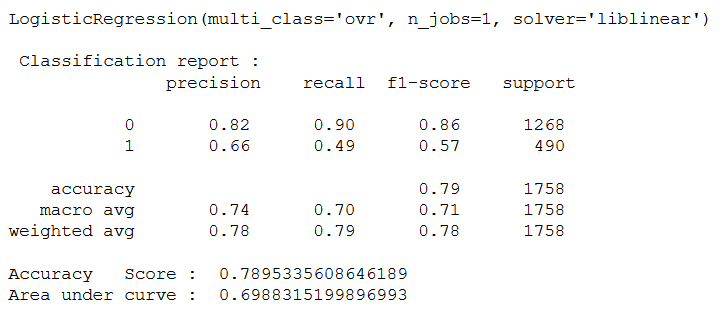


Figure Classification Report after RFE

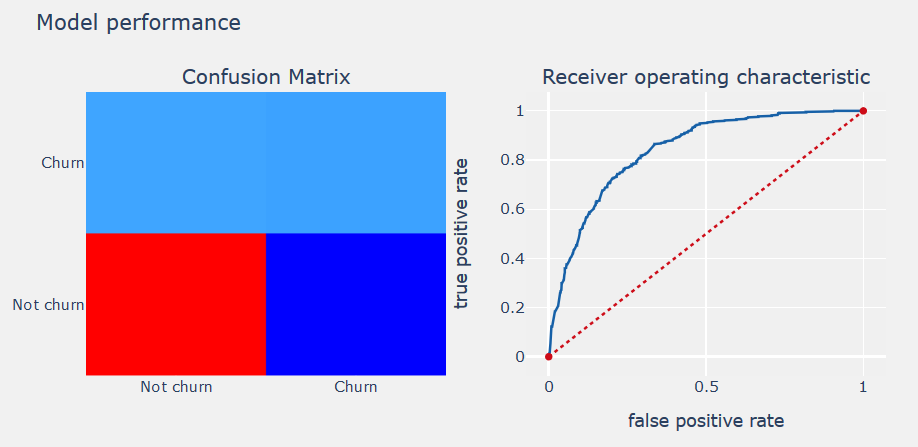


Figure Model Performance after RFE

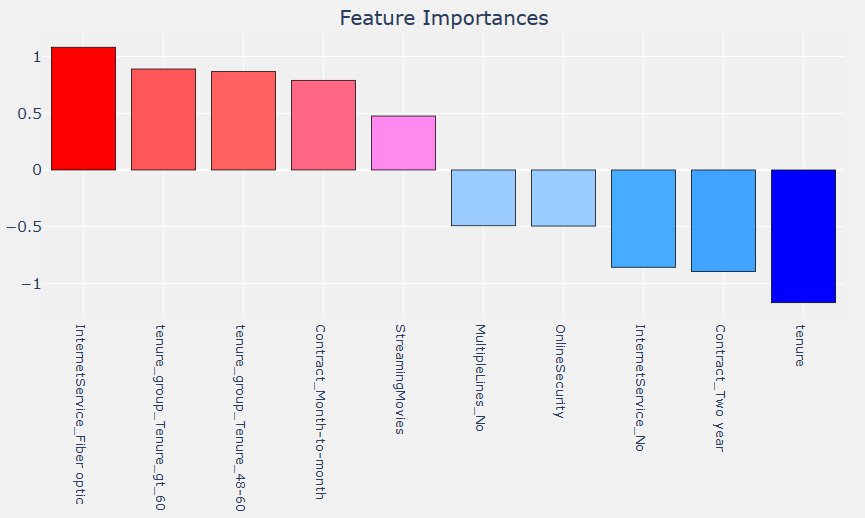


Figure Feature Importance

The 10 best or worst features which are used during RFE method and their importance in the model development. The coefficients of the features are the shown below:

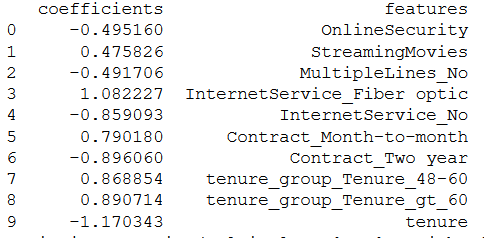


Figure Feature and their Coefficients

* 1. What are the limitations of the model and how should we handle the limitation?

Answer: There are some limitations of the model those are need to be discussed and also the way to make model more robust. They are:

* + 1. In the model development the dataset is imbalanced, i.e. having data points of one class is more than the other class. So, there may be possibility of the model to become more bias toward the dominating class. So, to overcome this problem further techniques like oversampling and under-sampling can be used to develop the model.
    2. Also, in this approach I considered the dataset only for the last month and considered them all as equal. However, events happening closer to the churn event might have a greater impact on our prediction. Splitting the variables according to the events(time-wise) might help in identifying the customers who are more likely to churn.
  1. Recommend some practices that will help lower the customer lifetime value and delight our existing customers.

Answer: Here are some practices which can be taken to reduce customer lifetime value and churn and delight our existing customers:

* + 1. **Listening and Fixing:** Listening is the key to understanding and getting a chance to fix the customer's problem. Daily conversation in the product design and development cycle to address the customer's issues when you get feedback will go a long way towards minimizing mistakes.
    2. **Chat or Talk Live:** Automation is very useful in talking to the customer, but talking them personally gives them a sense of their value to the company.
    3. **Patience and Empathy:** Hire more employees who have patience and empathy to deal with the problems facing by your customer.
    4. **Solution Oriented:** Solution-oriented organizations do not leave customer support to sales and service, but they are more involved in solving the customer's problems.
    5. **Regular Feedbacks**: Feedbacks are really necessary in knowing the mood of the customer towards the service of the company. Feedbacks can be taken through emails, surveys or personal calls.
    6. **Eye on the Competitors:** If you need to retain your existing customers it is necessary to make your product more advanced and customer friendly than of your competitors. It is important to make the perception in the eye of the customer is your product is one among the best in the market’

# References

* 1. ALAN AGRESTI, Categorical Data Analysis, 2nd Edition, John Wiley and Sons, 2002
  2. Hosmer, David W. and Lemeshow, Stanley, Applied Logistic Regression, 2nd Edition, John Wiley and Sons, 2000
  3. Goran Klepac, Robert Kopal and Leo Mri, Developing Churn Models Using Data Mining Techniques and Social Network Analysis, 1st Edition, 2014

# Appendix A: Logistic Regression – Algorithm Documentation

# Introduction

Regression techniques are one of the popular methods of doing any data analysis problem concerned with explaining a relationship between a response variable and one or more explanatory variables. The logistic regression model is one of the frequently used regression models for the analysis of these data.

Before going further explanation of the Logistic Regression model it is important to understand that the goal of the model is same as the other regression models used in statistics, that is, to find the best fit line which can describe a relationship between an outcome(dependent or response) variable and a set of independent(predictor or explanatory) variables.

The difference between logistic regression and a linear regression is that the response variable in logistic regression is *binary or dichotomous.* This difference is reflected both in the form of the model and its assumptions.

Example of logistic regression problems:

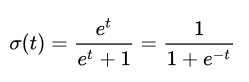
|  |  |  |
| --- | --- | --- |
|  | 1 or Positive Class | 0 or Negative Class |
| Loan | Approved | Not Approved |
| Email | Spam | Not Spam |
| Passenger | Survived | Not Survived |

# Logistic function, odds, odds ratio, and logit

### 

### Definition of the logistic function

The logistic function is a sigmoid function which takes any real input t, (t R), and outputs a value between zero and one, for the logit, this is interpreted as input log-odds and having output probability. The standard logistic function :R(0, 1) is defined as follows:

****

A graph of the logistic function on the t-interval (-6, 6) is shown in Figure 1.

Assuming t is a linear function of a single explanatory variable x (the case where t is a linear combination of multiple explanatory variables is treated similarly). t can be expressed as:



And the general logistic function p: R(0, 1) can now be written as:

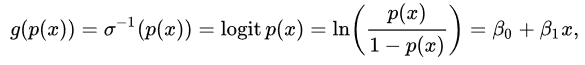


In the logistic model, p (x) is interpreted as the probability of the dependent variable Y equaling a success/case rather than a failure/non-case.

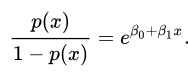
                                          

### Definition of the inverse of the logistic function

We can now define the logit (log odds) function as the inverse  of the standard logistic function. It is as follows:

****

 and equivalent, after exponentiating both sides we have the odds:



Interpretation of these terms:

In the above equation, the terms are as follows:

* g is the logit function. The equation g( p(x) ) is equivalent to the linear regression expression.
* ln denotes the natural logarithm.
* p (x) is the probability of the dependent variable is equal to the value of the logistic function of the linear regression expression.p (x) ranges from 0 to 1.
* 0is the intercept from the linear regression equation (the value where the predictor value is equal to zero).
* 1x is the regression coefficient multiplied by some value of the predictor.
* Base e denotes the exponential function.

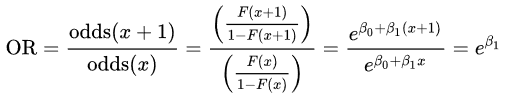
### Definition of the odds

Odds describes the ratio of success to ratio of failure.The odds of the dependent variable equaling a case (given some linear combination x of the predictors) is equivalent to the exponential function of the linear regression expression as follows:



### The odds ratio

For a continuous independent variable the odds ratio can be defined as:



This exponential relationship provides an interpretation for 1: The odds multiply by e1 for every 1-unit increase in x.

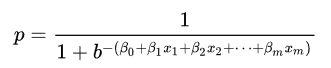
For a binary independent variable, the odds ratio is defined as  where a, b, c and d are cells in a 2x2 contingency table.

Multiple explanatory variables

If there are multiple explanatory variables, the above expression 0+1x can be revised to 0+1x1 +2x2 +.............mxm = i = 1mixi . So, the equation becomes:



and



        where usually b =e.

# Fitting a logistic regression model

Fitting of a logistic regression involves the process to optimize so the model gives the best possible results on training set labels.

This can be done using:

* By numerical approximation of maximum likelihood estimation.
* On large datasets, using stochastic gradient descent.

Just as ordinary least square regression is the method used to estimate coefficients for the best fit line in linear regression, logistic regression uses maximum likelihood estimation (MLE) to obtain the model coefficients that relate predictors to the target.

Maximum-likelihood estimation is a common learning algorithm used by a variety of machine learning algorithms, although it does make assumptions about the distribution of your data (more on this when we talk about preparing your data).

The best coefficients would result in a model that would predict a value very close to 1 (e.g. true class) for the default class and a value very close to 0 (e.g. false class) for the other class. The intuition for maximum-likelihood for logistic regression is that a search procedure seeks values for the coefficients (values) that minimize the error in the probabilities predicted by the model to those in the data (e.g. probability of 1 if the data is the primary class).

# Advantages and Disadvantages

### **Advantages:**

* Makes no assumptions about distributions of classes in feature space.
* Easily extended to multiple classes (multinomial regression).
* Natural probabilistic view of class predictions.
* Quick to train.
* Very fast at classifying unknown records.
* Good accuracy for many simple data sets.
* Resistant to overfitting.
* Can interpret model coefficients as indicators of feature importance.

### **Disadvantages:**

* Linear decision boundary.
* Not good for small datasets.
* Can only be used to predict discrete functions.

# Appendix B: CRISP-DM Model Steps

## B.1 Business Understanding:

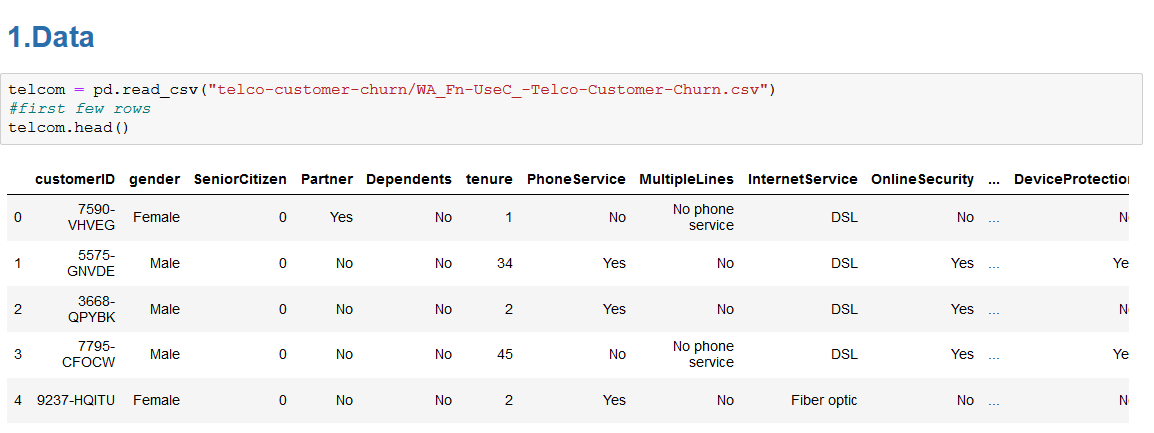
Step one is actually understanding the business or use case with the desired outcome. Only by understanding the final objective we can build a model that is actually of use. In our case the objective is reducing customer churn by identifying potential churn candidates beforehand, and take proactive actions to make them stay.

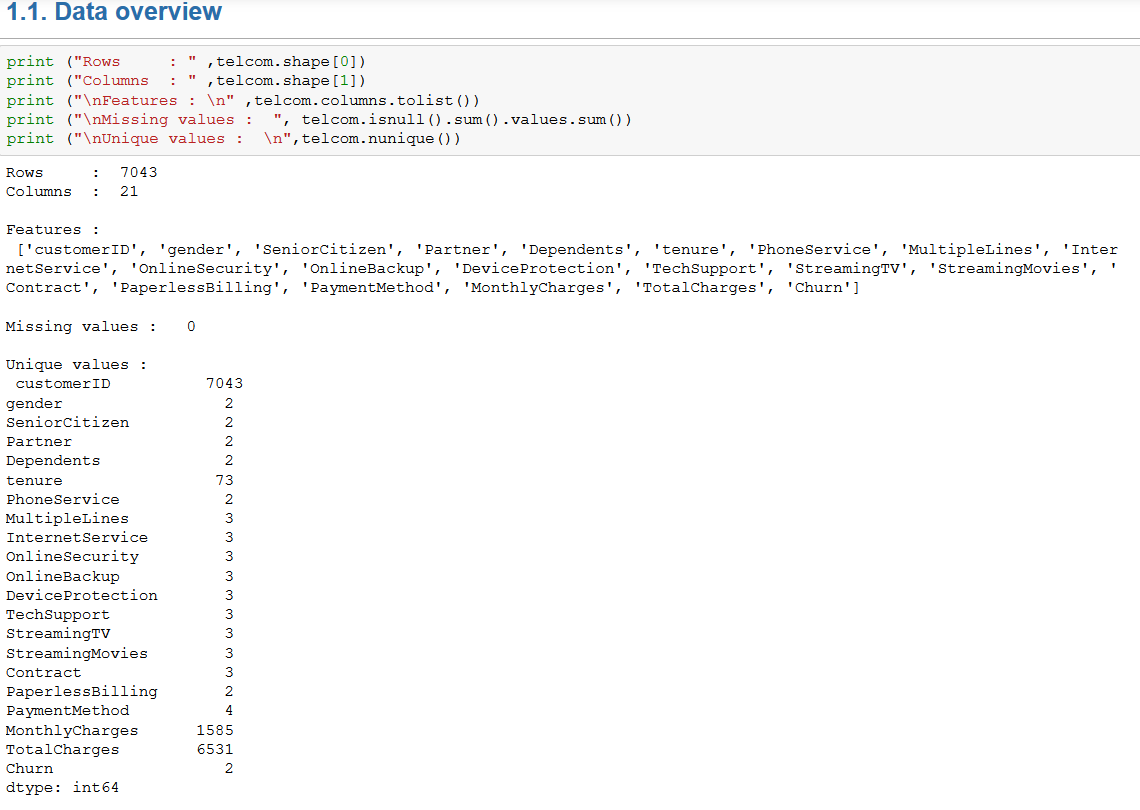
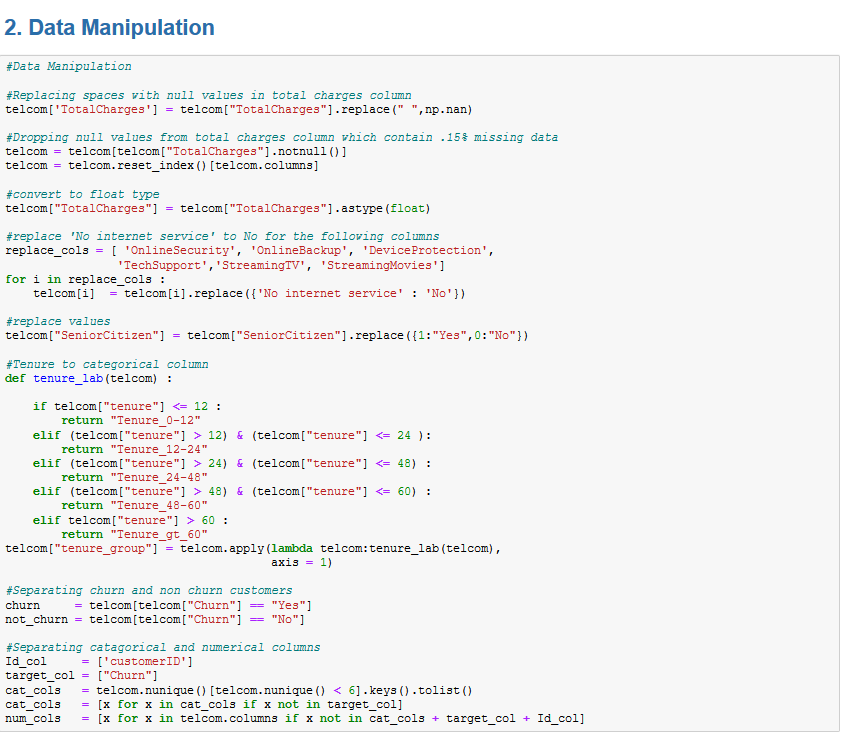
## B.2 Data Understanding

With understanding the context, it is possible to identify the right data sources, cleansing the data sets and preparing for feature selection or engineering. It sounds quite simple, but this is likely the hardest part. The predicting model is only as good as the data source. The dataset is taken from the Kaggle i.e. Telco Customer Churn dataset.

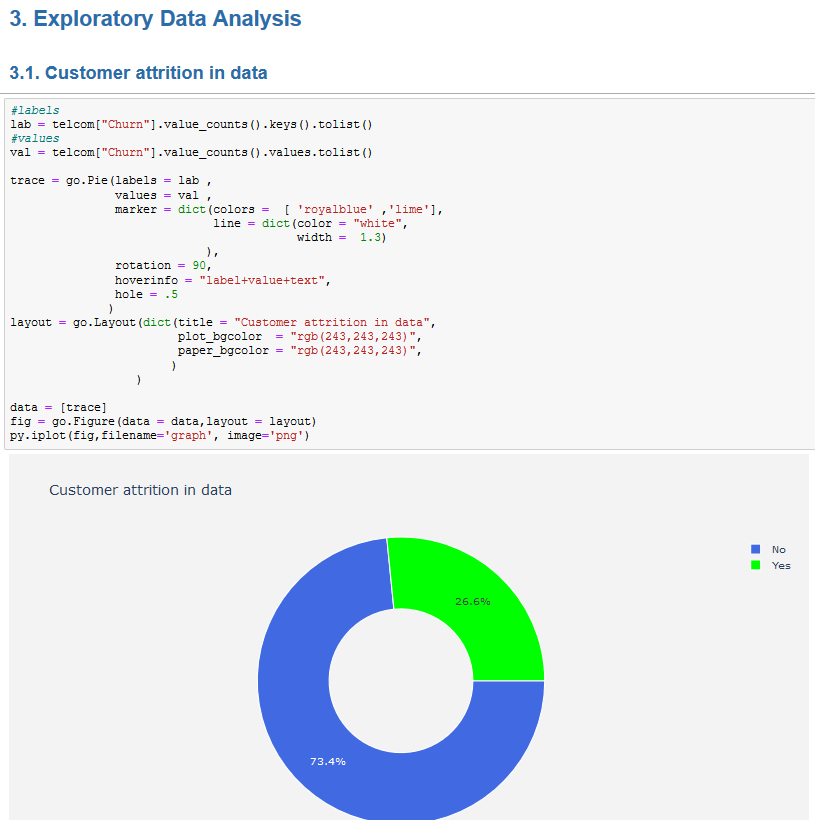
Loading the dataset in the python a then looking at the different variables, unique values and missing values.

Some data manipulation is also done to understand the data in better way.



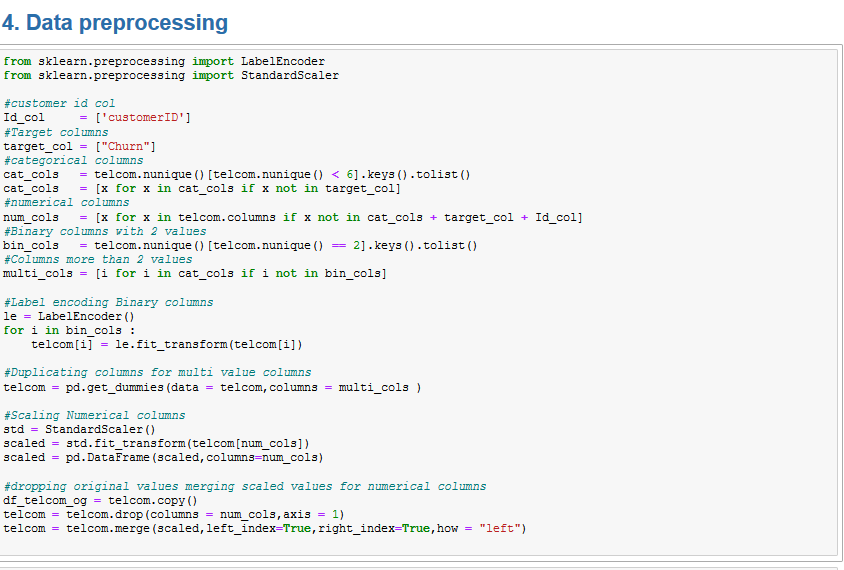
 

Exploratory Data Analysis is also done by plotting different graphs and plots in python.



## B.3 Data Preparation

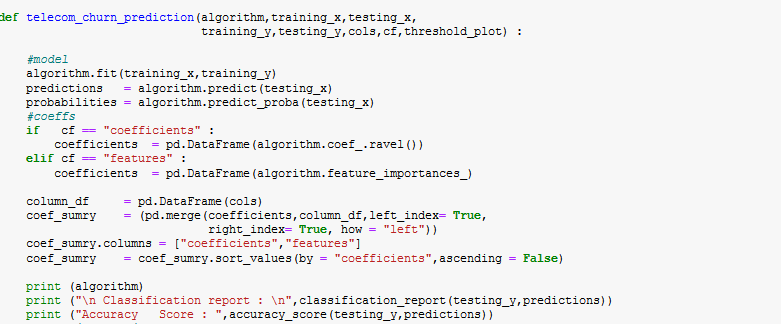
To Prepare the data for the model, data pre-processing techniques are used. Label Encoding is used to encode categorical values into numbers and Standard Scaling is used to standardize the numerical value variables in the dataset.



## B.4 Modelling

For Building the model I have used Logistic Regression Algorithm and build the model after splitting the dataset into test and train dataset.





## B.5 Evaluation

For Evaluation of the model ROC curve and Confusion Matrix is made using python. Feature Importance in the model is also plotted.

