Project Report

Customer churn for a telecom provider

*ANISHA GOYAL*

*April 2020*

**Contents**

**1. Introduction**

1.1 Problem Statement . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 3

1.2 Data . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 3

1.3 Exploratory Data Analysis . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 4

**2. Methodology**

2.1 Pre Processing . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 6

2.1.1 Missing Values . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 7

2.1.2 Outlier Analysis . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 7

2.1.3 Feature Selection . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 8

2.1.4 Feature Scaling . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .9

2.2 Modeling . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 7

2.2.1 Logistic Regression . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 9

2.2.2 KNN . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 9

2.2.3 SVM . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .. . . . . . . . . 10

2.2.4 Naïve Bayes . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 11

2.2.5 Decision Tree. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 11

2.2.6 Random Forest . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .12

2.2.7 Artifical Neural Networks . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 12

**3. Conclusion**

3.1 Model Evaluation . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 13

3.2 Model Selection . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 13

**Appendix**

Extra Figures . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 14

**Chapter 1**

**Introduction**

* 1. **Problem Statement**

Churn is one of the largest problems facing most businesses. Accordingly **it costs between 5 times and 25 times as much to find a new customer than to retain an existing one**. In other words, your existing customers are worth their weight in gold!

* 1. **Data**

There are 44 variables in our data in which 43 are independent variables and 1 (canc\_flag) is dependent variable. These data can be viewed as classification tasks.

**Variables Information:**

1. **Customer Churn**
2. **Gender**
3. **Year of birth**
4. **State of the customer**
5. **Locality Type**
6. **Average Living area**
7. **Service Revenue Average**
8. **Average Revenue**
9. **Average discount**
10. **Average installment**
11. **Service Revenue Standard**
12. **Standard Revenue**
13. **Standard Voucher**
14. **Standard Discount**
15. **Standard Installment**
16. **Subscriber Handset Indicator**
17. **Tariff Plan Name**
18. **L3 Tariff Description**
19. **L4 Tariff Description**
20. **L5 Tariff Description**
21. **Tariff Group**
22. **Device Type**
23. **Months Since Activation**
24. **Star Rating**
25. **Number of Contract Cycle**
26. **Number of Contract Months Remaining**
27. **Months Since Contract Start**
28. **Months Since Last Contract**
29. **Month Till Next Contract**
30. **Average Voice Minutes For Wireline**
31. **Average Voice Minutes Onnet**
32. **Average Voice Minutes Offnet**
33. **Average Voice Minutes Roaming**
34. **Average of Total Data Volume**
35. **Average of Total Data Volume Roaming**
36. **Minimum Voice International Standard**
37. **Minimum Voice Onnet Standard**
38. **Minimum Voice Offnet Standard**
39. **Minimum Voice Roaming Standard**
40. **Total Data Volume Standard**
41. **Total Data Volume Standard Roaming**
42. **Total Calls**
43. **Dropped Calls**
44. **Average Days Of Usage**

**1.3 Exploratory Data Analysis**

Exploratory Data Analysis (EDA) is an approach to analyzing data sets to summarize their main characteristics. In the given data set there are 44 variables. There are 1001 observations and 44 columns in our data set. Missing Values are present in our data.

**List of columns and their number of unique values** -

Customer Churn 2

Gender 2

Year of birth 65

State of the customer 17

Locality Type 4

Average Living area 782

Service Revenue Average 943

Average Revenue 944

Average discount 605

Average installment 13

Service Revenue Standard 896

Standard Revenue 897

Standard Voucher 72

Standard Discount 555

Standard Installment 10

Subscriber Handset Indicator 2

Tariff Plan Name 113

L3 Tariff Description 19

L4 Tariff Description 36

L5 Tariff Description 111

Tariff Group 12

Device Type 4

Months Since Activation 234

Star Rating 6

Number of Contract Cycle 16

Number of Contract Months Remaining 23

Months Since Contract Start 116

Months Since Last Contract 114

Month Till Next Contract 109

Average Voice Minutes For Wireline 869

Average Voice Minutes Onnet 931

Average Voice Minutes Offnet 872

Average Voice Minutes Roaming 238

Average of Total Data Volume 804

Average of Total Data Volume Roaming 109

Minimum Voice International Standard 169

Minimum Voice Onnet Standard 935

Minimum Voice Offnet Standard 873

Minimum Voice Roaming Standard 248

Total Data Volume Standard 778

Total Data Volume Standard Roaming 114

Total Calls 158

Dropped Calls 7

Average Days Of Usage 443

**From EDA we have concluded that there are 32 continuous variable and 12 categorical variable in nature.**

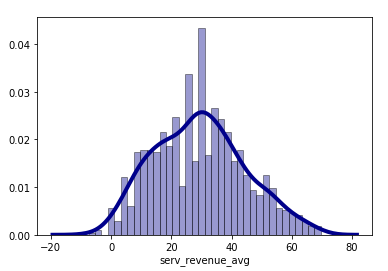
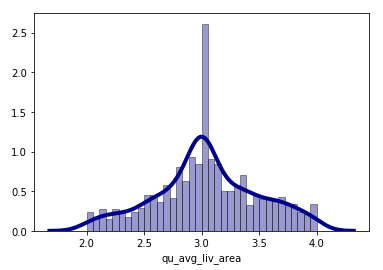
**Chapter 2**

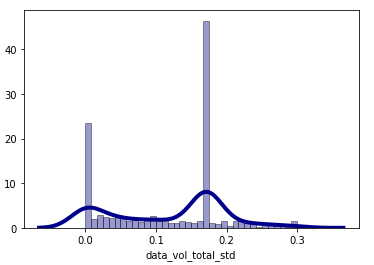
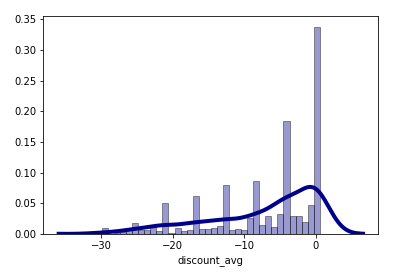
**Methodology**

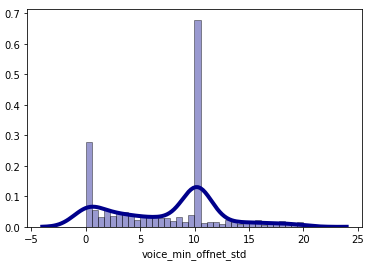
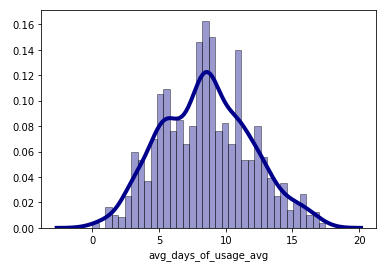
Before fitting the data to the model we need to clean the data and convert it to a proper format. It is the most crucial part of data science project we spend almost 80% of time in it.

**2.1 Pre Processing**

Any predictive modeling requires that we look at the data before we start modeling. However, in data mining terms looking at data refers to so much more than just looking. Looking at data refers to exploring the data, cleaning the data as well as visualizing the data through graphs and plots. This is often called as Exploratory Data Analysis.







**2.2.1 Missing Value Analysis**

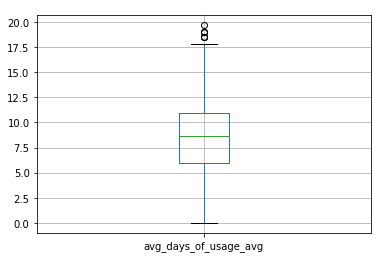
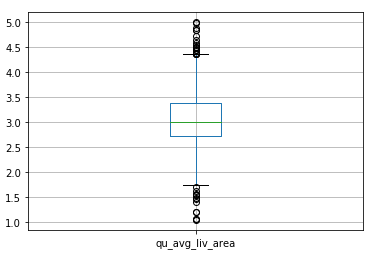
In statistics, missing data, or missing values, occur when no data value is stored for the variable in an observation. Missing data are a common occurrence and can have a significant effect on the conclusions that can be drawn from the data. If a columns has more than 30% of data as missing value either we ignore the entire column or we ignore those observationsSo we will compute missing value for all the columns.

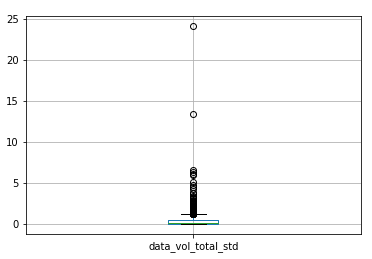
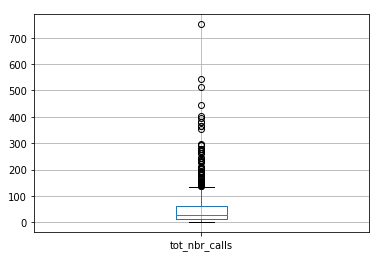
**In this project we have used forward fil method to impute missing value**.

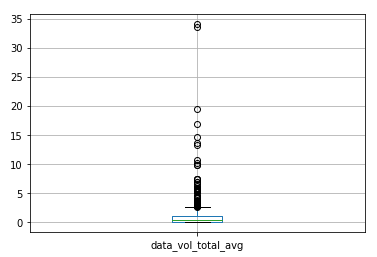
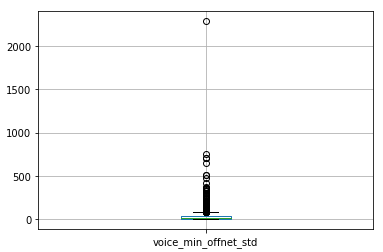
**2.1.2 Outlier Analysis**

The skew in these distributions can be most likely explained by the presence of outliers and extreme values in the data. One of the other steps of pre-processing apart from checking for normality is the presence of outliers. In this case we use a classic approach of removing outliers. We visualize the outliers using boxplots.

In figure we have plotted the boxplots of the variables. A lot of useful inferences can be made from these plots. First as you can see, we have a lot of outliers and extreme values in each of the data set.







From the boxplot almost all the variables consists of outliers. We have replaced the outliers with median.

**2.1.3 Feature Selection**

Before performing any type of modeling we need to assess the importance of each predictor variable in our analysis. There is a possibility that many variables in our analysis are not important at all to the problem of class prediction. Selecting subset of relevant columns for the model construction is known as Feature Selection. We cannot use all the features because some features may be carrying the same information or irrelevant information which can increase overhead. To reduce overhead we adopt feature selection technique to extract meaningful features out of data. In this project we have selected **Correlation Analysis** for numerical variable(i.e Pearsons and Kendall test) and **chi-square test** for categorical variable.

**2.2.4 Feature Scaling**

**Feature scaling** is a method used to standardize the range of independent variables or features of data. In data processing, it is also known as data normalization and is generally performed during the data preprocessing step. Since the range of values of raw data varies widely, in some machine learning algorithms, objective functions will not work properly without normalization. For example, the majority of classifiers calculate the distance between two points by the Euclidean distance. If one of the features has a broad range of values, the distance will be governed by this particular feature. Therefore, the range of all features should be normalized so that each feature contributes approximately proportionately to the final distance. Since our data is not uniformly distributed we will use **Normalization** as Feature Scaling Method.

**2.2 Modeling**

After a thorough preprocessing we will be using some regression models on our processed data to predict the target variable. Following are the models which we have built –

**2.2.1 Logistic Regresssion**

**Logistic regression** is a statistical **model** that in its basic form uses a **logistic** function to **model** a binary dependent variable, although many more complex extensions exist. In **regression** analysis, **logistic regression** (or **logit regression**) is estimating the parameters of a **logistic model** (a form of binary **regression**).The confusion matrix and the accuracy is given by

|  |  |  |
| --- | --- | --- |
|  | Predicted Value | |
| Actual value | 183 | 0 |
| 17 | 0 |

Accuracy is **91.5** percent

**2.2.2 KNN**

K-nearest neighbors (**KNN**) algorithm is a type of supervised ML algorithm which can be used for both classification as well as regression predictive problems. However, it is mainly used for classification predictive problems in industry.

|  |  |  |
| --- | --- | --- |
|  | Predicted Value | |
| Actual value | 181 | 2 |
| 16 | 1 |

Accuracy is **91** percent

**2.2.3 SVM**

A support vector machine (**SVM**) is a supervised machine learning model that uses classification algorithms for two-group classification problems. After giving an **SVM** model sets of labeled training data for each category, they're able to categorize new text.

SVM using kerenal as linear function then the confusion matrix and accuracy is given by

|  |  |  |
| --- | --- | --- |
|  | Predicted Value | |
| Actual value | 183 | 0 |
| 17 | 0 |

Accuracy is **91.5** percent

SVM using kerenal as radial base function then the confusion matrix and accuracy is given by

|  |  |  |
| --- | --- | --- |
|  | Predicted Value | |
| Actual value | 183 | 0 |
| 17 | 0 |

Accuracy is **91.5** percent

SVM using kerenal as polymonial function then the confusion matrix and accuracy is given by

|  |  |  |
| --- | --- | --- |
|  | Predicted Value | |
| Actual value | 183 | 0 |
| 17 | 0 |

Accuracy is **91.5** percent

SVM using kerenal as sigmoid function then the confusion matrix and accuracy is given by

|  |  |  |
| --- | --- | --- |
|  | Predicted Value | |
| Actual value | 178 | 5 |
| 16 | 1 |

Accuracy is **89.5** percent

**2.2.4 Naïve Bayes**

**Naive Bayes** is a simple, yet effective and commonly-used, machine learning classifier. It is a probabilistic classifier that makes classifications using the Maximum A Posteriori decision rule in a **Bayesian** setting. It can also be represented using a very simple **Bayesian** network .below is confusion matrix and accuracy of the model.

|  |  |  |
| --- | --- | --- |
|  | Predicted Value | |
| Actual value | 179 | 4 |
| 17 | 0 |

Accuracy is **89.5** percent

**2.2.5 Decision Tree**

A decision tree is a decision support tool that uses a tree-like graph or model of decisions and their possible consequences, including chance event outcomes, resource costs, and utility. It can be used for classification and regression. It is a supervised machine learning algorithm. Accept continuous and categorical variables as independent variables. Extremely easy to understand by the business users.

|  |  |  |
| --- | --- | --- |
|  | Predicted Value | |
| Actual valve | 156 | 27 |
| 14 | 3 |

Accuracy is **79.5** percent

**2.2.6 Random Forest**

Random Forest is an ensemble technique that consists of many decision trees. The idea behind Random Forest is to build n number of trees to have more accuracy in dataset. It is called random forest as we are building n no. of trees randomly. In other words, to build the decision trees it selects randomly n no of variables and n no of observations to build each decision tree. It means to build each decision tree on random forest we are not going to use the same data.

|  |  |  |
| --- | --- | --- |
|  | Predicted Value | |
| Actual valve | 183 | 0 |
| 17 | 0 |

Accuracy is 91.5 percent

**2.2.7 ANN** 

An artificial neural network (**ANN**) is the piece of a computing system designed to simulate the way the human brain analyzes and processes information. It is the foundation of artificial intelligence (AI) and solves problems that would prove impossible or difficult by human or statistical standards.

|  |  |  |
| --- | --- | --- |
|  | Predicted Value | |
| Actual valve | 157 | 26 |
| 16 | 1 |

Accuracy is **79** percent

**Chapter 3**

**Conclusion**

In this chapter we are going to evaluate our models, select the best model for our dataset and try to get answers of the asked questions.

**3.1 Model Evaluation**

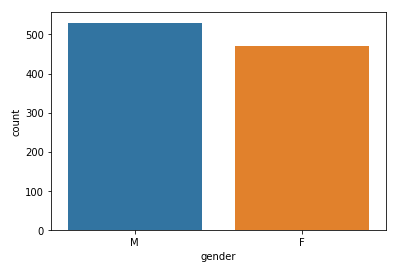
In the previous chapter we have seen the **Confusion matrix** and **Accuracy** Value of different models. A **confusion matrix** is a table that is often used to describe the performance of a classification model (or "classifier") on a set of test data for which the true values are known. The **confusion matrix** gives you a lot of information, but sometimes you may prefer a more concise metric. TP is the number of true positives, and FP is the number of false positives. A trivial way to have perfect precision is to make one single positive prediction and ensure it is correct (precision = 1/1 = 100%).Higher is the accuracy better is the model.

**3.2 Model Selection**

From the observation of all **confusion matrix** and **Accuracy** Value we have concluded that **ANN**  has minimum value of accuracy (i.e 79%).

**Appendix**

**Extra Figures**

**** 