**Churn reduction Report**

**Prajwal R**

**08-12-2018**

**Contents**

**1. Introduction**

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

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

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

**2. Methodology**

2.1 Pre Processing . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 5

2.1.1 Missing Value Analysis . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .5

2.1.2 Outlier Analysis . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 6

2.1.3 Feature Selection . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 7

2.1.4 Feature Scaling . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 8

2.2 Modeling . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .9

2.2.1 Decision Tree. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .9

2.2.2 Random Forest . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 9

2.2.3 Logistic Regression . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .10

2.2.4 KNN. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .10

2.2.5 Naïve Bayes. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 10

**3. Conclusion**

3.1 Model Evaluation . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 11

3.2 Model Selection . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 11

**Appendix**

Extra Figures . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 12

**References**

**Chapter 1**

**Introduction**

* 1. **Problem Statement**

Churn (loss of customers to competition) is a problem for companies because it is more expensive to acquire a new customer than to keep your existing one from leaving. This problem statement is targeted at enabling churn reduction using analytics concepts.

* 1. **Data**

There are 21 variables in our data in which 20 are independent variables and 1 (Churn) is dependent variable. Since our target variable is categorical in nature, this is a classification problem.

**Variables Information:**

1. state
2. account length
3. area code
4. phone number
5. international plan
6. voicemail plan
7. number of voicemail messages
8. total day minutes used
9. day calls made
10. total day charge
11. total evening minutes
12. total evening calls
13. total evening charge
14. total night minutes
15. total night calls
16. total night charge
17. total international minutes used
18. total international calls made
19. total international charge
20. number of customer service calls made
21. Churn (Target variable)

**1.3 Exploratory Data Analysis**

Exploratory Data Analysis (EDA) is an approach to analyze data sets to summarize their main characteristics. In the given data set there are 21 variables and data types of all variables are float64, int64 and object. There are 3333 observations and 21 columns in our data set.

RangeIndex: 3333 entries, 0 to 3332

Data columns (total 21 columns):

state 3333 non-null object

account length 3333 non-null int64

area code 3333 non-null int64

phone number 3333 non-null object

international plan 3333 non-null object

voice mail plan 3333 non-null object

number vmail messages 3333 non-null int64

total day minutes 3333 non-null float64

total day calls 3333 non-null int64

total day charge 3333 non-null float64

total eve minutes 3333 non-null float64

total eve calls 3333 non-null int64

total eve charge 3333 non-null float64

total night minutes 3333 non-null float64

total night calls 3333 non-null int64

total night charge 3333 non-null float64

total intl minutes 3333 non-null float64

total intl calls 3333 non-null int64

total intl charge 3333 non-null float64

number customer service calls 3333 non-null int64

Churn 3333 non-null object

dtypes: float64(8), int64(8), object(5)

memory usage: 546.9+ KB

From EDA we came to know that apart from Target variable, there are 15 continuous variables and 5 categorical variables.

**Continuous variables:**

account length, number vmail messages, total day minutes, total day calls, total day charge, total eve minutes, total eve calls, total eve charge, total night minutes, total night calls, total night charge, total intl minutes, total intl calls, total intl charge, number customer service calls, phone number

**Categorical variables:**

state, area code, international plan, voice mail plan

**Target variable:**

Churn

**Chapter 2**

**Methodology**

Before feeding 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. Few pre-processing techniques are applied on the data set to bring it to proper shape.

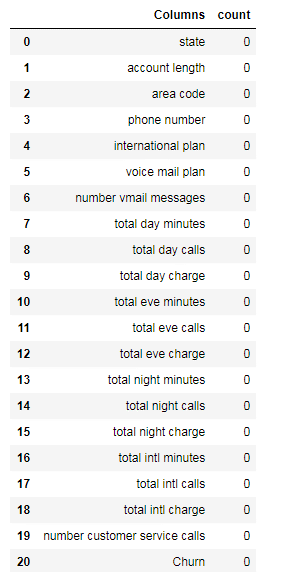
**2.1 Pre-Processing**

Any predictive modeling requires that we look at the data before we start modeling. 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.

As part of data pre-processing, all the variables with object data type are converted to numerical data type.

**2.1.1 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 column has more than 30% of missing values, either we ignore the entire column or we ignore those observations.

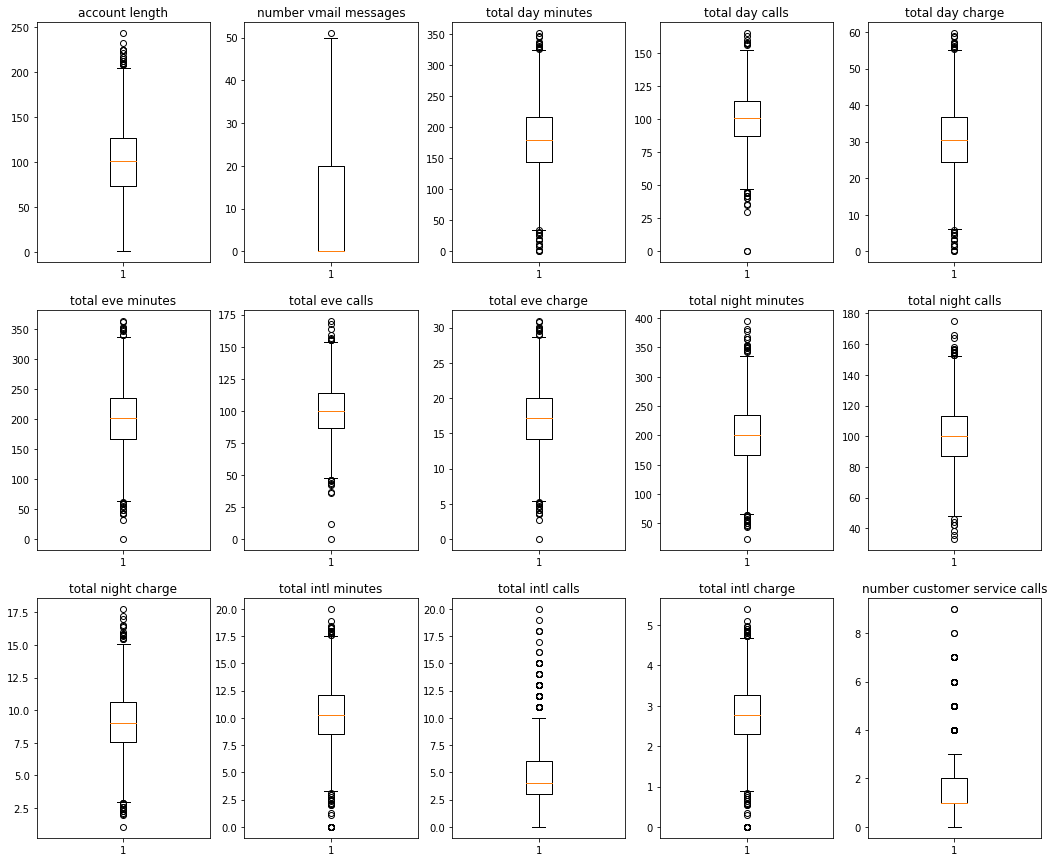


Since there are no missing values in the data set, we can proceed without any missing value analysis.

**2.1.2 Outlier Analysis**

One of the pre-processing step after imputing the missing values is checking for the presence of outliers. In this case we use a classic approach of removing outliers. We visualize the outliers using boxplots.

Boxplots are plotted for each of the continuous variables and based on the insights and proper understanding about the data, no outliers are removed from the variables.

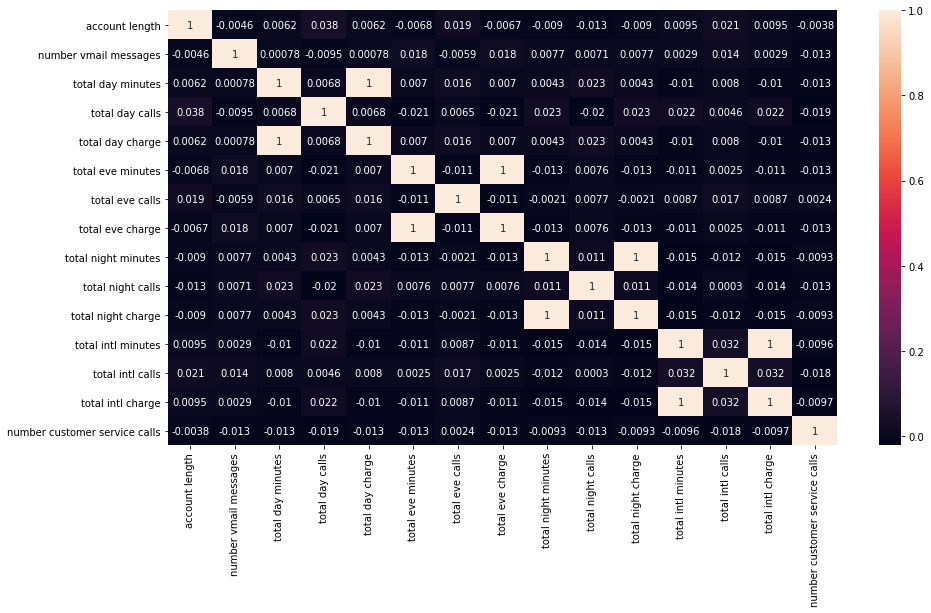


From the boxplot, we see that all the variables except ‘number vmail messages’ have outliers. But all the values which are depicted as outliers cannot be considered as outliers since all the observations fall in the possible range.

**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. 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. This in turn helps us to avoid the problem of multi collinearity.

In this project we have selected **Correlation** Analysis for numerical variable to remove multi collinearity from the data set.



From correlation analysis, it is evident that variables ‘total day minutes’, ‘total eve minutes’, ‘total nght minutes’ and ‘total intl minutes’ have high correlation of 1. So these variables are removed from the data set.

**Chi square test:**

In order to find the relationship between categorical variables, we use chi square test. From chi square test, we found that the variable ‘area code’ is independent of our target variable.

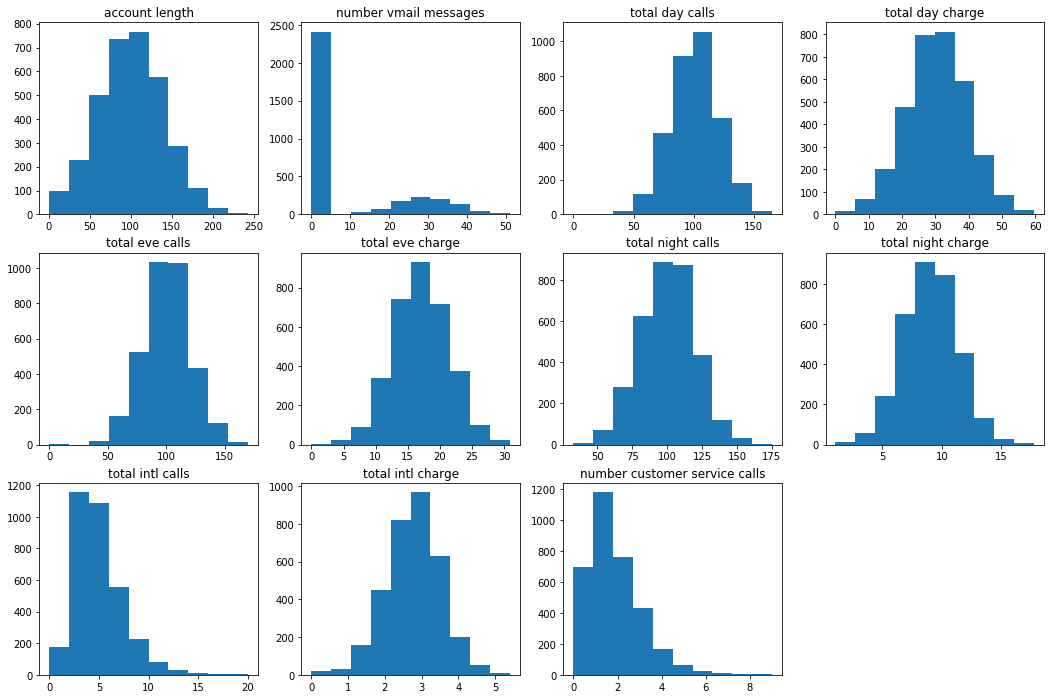
Hence this variable is removed from the data set.

**2.1.4 Feature Scaling**

Feature scaling is a method used to bring the data to a suitable range of values. Normalization and Standardization are the two major techniques used to scale the data. 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.

We have plotted the histogram to check the distribution of data across all the continuous variables.

Since most of the variables are not uniformly distributed, data is fed to the model as is.



After pre-processing, it is found that variable ‘phone number’ is not important for modeling. Hence it is removed from the data set.

After all the data wrangling, data set contains 14 independent variables and 1 dependent variables without any missing values and unwanted outliers. Will freeze this dataset and apply different classification algorithms.

**2.2 Modeling**

After a thorough preprocessing we will be using some classification models on our processed data to predict the target variable. Models are trained using train dataset and evaluated using test dataset.

**2.2.1 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. Each branch connects nodes with “and” and multiple branches are connected by “or”. 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.

Decision tree model is applied on the train dataset and the values are predicted for the test data using predict function. Predicted values and actual values of test data are compared and **Accuracy, False negative rate and Sensitivity** are calculated.

|  |  |  |
| --- | --- | --- |
| Decision Tree | R | PYTHON |
| Accuracy | 95.62 | 92.50 |
| FNR | 29.4 | 30.35 |
| Sensitivity | 70.53 | 69.64 |

**2.2.2 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 randomly selects n no of variables and n no of observations to build each decision tree.

Random Forest model is applied on the train dataset and the values are predicted for the test data using predict function. Predicted values and actual values of test data are compared and **Accuracy, False negative rate and Sensitivity** are calculated.

|  |  |  |
| --- | --- | --- |
| Random Forest | R | PYTHON |
| Accuracy | 94.99 | 95.74 |
| FNR | 29.01 | 29.91 |
| Sensitivity | 70.98 | 70.08 |

**2.2.3 Logistic Regression**

Linear Regression is one of the statistical methods of prediction. It is used to predict categorical data. It predicts the probability of the target variable. Then probability values are converted to actual values.

Logistic regression model is applied on the train dataset and the values are predicted for the test data using predict function. Predicted values and actual values of test data are compared and **Accuracy, False negative rate and Sensitivity** are calculated.

|  |  |  |
| --- | --- | --- |
| Linear Regression | R | PYTHON |
| Accuracy | 87.4 | 87.22 |
| FNR | 84.37 | 77.67 |
| Sensitivity | 15.62 | 22.32 |

**2.2.4 KNN**

KNN is simple algorithm that stores all available cases and classifies new cases based on a similarity measure. KNN measures the distance between given data point and all other data points. Then based on the specified k value, it selects the closest neighbors to that data point. In case of classification, it will take the majority of neighbor values and assigns that value.

KNN model is applied on the train dataset and the values are predicted for the test data using predict function. Predicted values and actual values of test data are compared and **Accuracy, False negative rate and Sensitivity** are calculated.

|  |  |  |
| --- | --- | --- |
| KNN | R | PYTHON |
| Accuracy | 80.26 | 85.84 |
| FNR | 79.46 | 86.60 |
| Sensitivity | 20.53 | 13.39 |

**2.2.5 Naïve Bayes**

Naïve Bayes algorithm is used only for classification problems. It predicts the probability of the target variable. It is the only classification model where multi collinearity will not affect the outcomes because it works on the assumption that all the variables are independent.

Naïve Bayes model is applied on the train dataset and the values are predicted for the test data using predict function. Predicted values and actual values of test data are compared and **Accuracy, False negative rate and Sensitivity** are calculated.

|  |  |  |
| --- | --- | --- |
| Naïve Bayes | R | PYTHON |
| Accuracy | 88.12 | 85.84 |
| FNR | 71.4 | 60.26 |
| Sensitivity | 28.57 | 39.73 |

**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 calculated accuracy, false negative rate and sensitivity for different models. Since the main objective of this model is churn reduction, accuracy alone cannot decide the performance of the model. Hence sensitivity and false negative rate are calculated.

Model with high accuracy and sensitivity and low false negative rate, can be selected as the best model for this data.

**3.2 Model Selection**

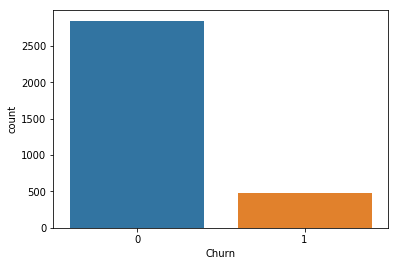
After observing the accuracy, sensitivity and false negative rate values of all the models, we can conclude that Random forest has higher values accuracy and sensitivity and low value of false negative rate. Therefore, we can freeze Random forest model for this data set.

**Appendix**

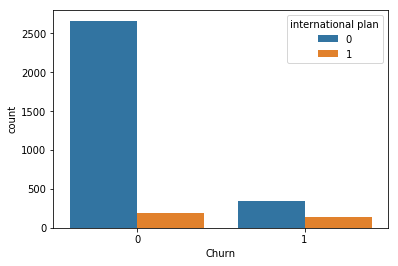
**Extra Figures**

Relationship of our target variable (Churn) with other variables.

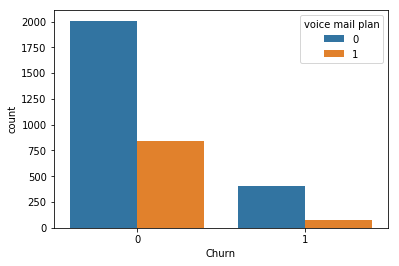
1. Graph of number of customers who are churned out



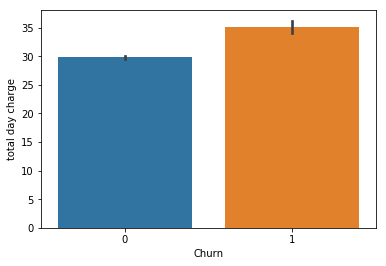
1. Graph of customers’ behavior having international plan



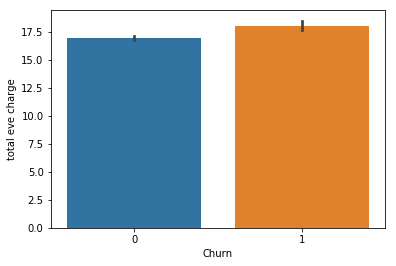
1. Graph of customers’ behavior having voice mail plan



1. Graph of 'total day charge' vs 'churn'



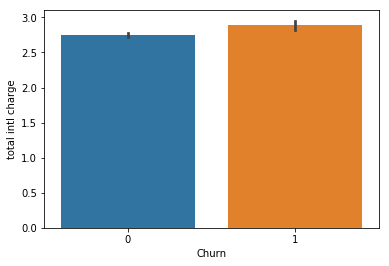
1. Graph of 'total eve charge' vs 'churn'

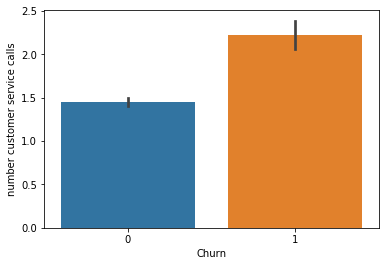


1. Graph of 'total night charge' vs 'churn'



1. Graph of 'total intl charge' vs 'churn'



1. Graph of 'number customer service calls' vs 'churn'

**References**

1. For Data Cleaning and Model Development -

<https://edwisor.com/career-data-scientist>