REPORT

CHURN REDUCTION

DIBYANSHU KUMAR

18/02/2019

**Contents:**

* Introduction
* EDA
* Target variable analysis
* Feature Selection
* Bivariate Analysis
* Univariate Analysis
* Sampling
* Modeling

**Introduction**

**1.1 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.

**1.2 Data**

Our task is to build classification models which will classify that whether a customer will churn out or not based on the given training data.

**Training data**: The dimension of training data is (3333, 21) which means 3333 rows and 21 columns. Name of the variables are as follows:

['state', 'account length', 'area code', 'phone number',

'international plan', 'voice mail plan', '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', 'Churn']

In these 21 variables 20 variables are independent variables and ‘Churn’ is our dependent variable

or target variable.

**Test data**: The dimension of test data is (1667,21) . It contains the same set of variable as train data.

We would build our model which will train by the training data and will check the accuracy of the model by predicting the results for test data.

**Train data information:**

<class 'pandas.core.frame.DataFrame'>

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

**Test data information:**

<class 'pandas.core.frame.DataFrame'>

RangeIndex: 1667 entries, 0 to 1666

Data columns (total 21 columns):

state 1667 non-null object

account length 1667 non-null int64

area code 1667 non-null int64

phone number 1667 non-null object

international plan 1667 non-null object

voice mail plan 1667 non-null object

number vmail messages 1667 non-null int64

total day minutes 1667 non-null float64

total day calls 1667 non-null int64

total day charge 1667 non-null float64

total eve minutes 1667 non-null float64

total eve calls 1667 non-null int64

total eve charge 1667 non-null float64

total night minutes 1667 non-null float64

total night calls 1667 non-null int64

total night charge 1667 non-null float64

total intl minutes 1667 non-null float64

total intl calls 1667 non-null int64

total intl charge 1667 non-null float64

number customer service calls 1667 non-null int64

Churn 1667 non-null object

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

memory usage: 273.6+ KB

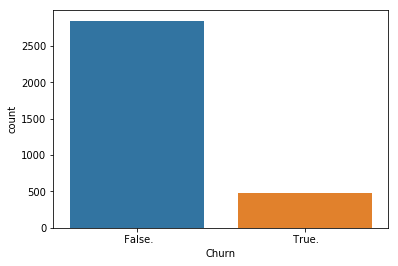
**EDA**

We will focus on:

1. **Understand the problem**. We'll look at each variable and do a philosophical analysis about their meaning and importance for this problem.
2. **Univariate Analysis**. We'll just focus on the dependent variable ('Churn') and try to know a little bit more about it.
3. **Bivariate Analysis**. We'll try to understand how the dependent variable and independent variables relate.
4. **Basic cleaning**. We'll clean the dataset and handle the missing data, outliers and categorical variables if available.
5. **Test assumptions**. We'll check if our data meets the assumptions required by most multivariate techniques.

**Target Variable Analysis:**

Our target variable is churn and it contains two categorical values i.e. ‘True’ and ‘False’. The value of churn is ‘True’ if the customer has left and is ‘False’ elsewise.



Count of ‘True’ values = 483

Count of ‘False’ values = 2850

Percentage of ‘True’ values = (483 / (2850+483))\*100 = 14.4914

This plot shows the count of churn. We can see that the people churning out are quite less in number than those who are not. So there can also be a **Target imbalance problem** for our model, which means more no of outcomes favouring a particular class (False). This could lead to more number of wrong predictions.

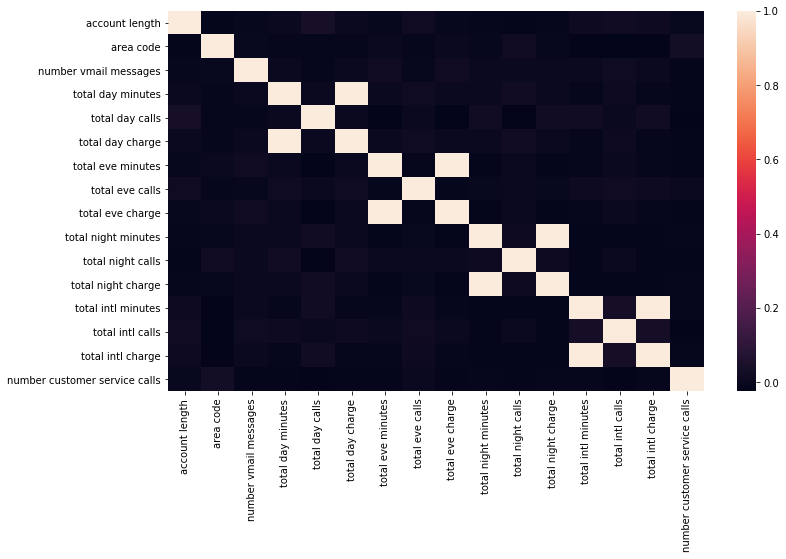
For this problem we are more concerned about the False Negative Rate so that we don’t predict wrong outcome for those who would churn out. We need to keep the False Negative Rate as low as possible.

To decrease FNR and to avoid target imbalance problem we would randomly over sample the minority class.

**Feature Selection:**

To select features which will be representing our data is one of the important task to perform since we don’t want to take unnecessary variables into our model. There could be variables present in our dataset which would depend on other variables. If the correlation is very high amongst two independent variable then we have to drop one of them since it would decrease the model performance. High correlation amongst two variables means that both of them are carrying almost same information. For categorical variables we will use chi-square test to see which categorical value should be taken in the model.

Heatmap for the numerical variables:



In the heatmap we can see that ‘total day minutes’ is quite correlated with ‘total day charge’ which is same for the evening, night, international minutes and charge respectively. We would take only of the highly correlated variables.

For the categorical variables we will perform chi-square test. In chi-square test we find the dependency between the dependent variable and categorical independent variables by calculating the p-value. If p-value is less than 0.05 then we reject null hypothesis and we say the variables are dependent or else the variable are independent.

The p-values for this problem is as follows:

state

0.002296221552011188

phone number

0.49185608455943547

international plan

2.4931077033159556e-50

voice mail plan

5.15063965903898e-09

We can see that the phone number has very high p-value thus there is no relation between ‘phone number’ and ‘Churn’. So we would drop the variable ‘phone number’. Also the variable ‘state’ has too many categories and it is also not affecting our target variable so much. So we will remove this variable too.

Thus after doing the Feature Selection we are left with the following variables.

'account length', 'area code', 'international plan', 'voice mail plan', 'number vmail

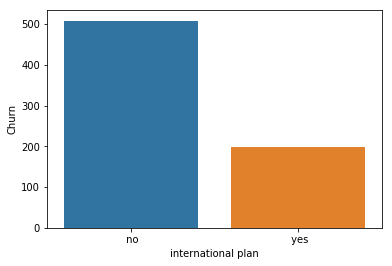
messages', 'total day calls', 'total day charge', 'total eve calls', 'total eve charge', 'total

night calls', 'total night charge', 'total intl calls', 'total intl charge', 'number customer service calls', 'Churn’

**Bivariate Analysis:**

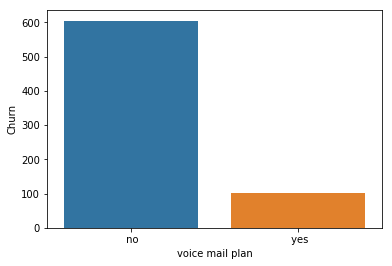
We would now do the analysis of the independent variables with respect to the dependent variable.

* First we consider ‘international plan’ and 'Churn’:



In the x-axis we have international plan and on y-axis we have count of people who churn out. From the graph we see that more number of people churn if they are not taking any international plans.

* ‘voice mail plan’ and ‘Churn’



Also more no of people churn if they did not took voice mail plan.

* ‘Churn’ and ‘number vmail message’



On an average the number of vmail messages sent by a person who churn out is less than a person who won’t churn.

* Analysis between ‘Churn’ and ‘total day calls’ , ‘total eve calls’ , ‘total night calls’ and ‘total intl calls’



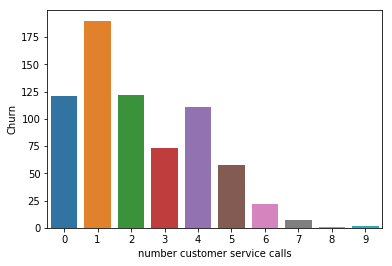
From the graphs we could see that churning does not depend much on the average no of calls made. Although it slightly varies in case of ‘total int calls’.

* Analysis between ‘Churn’ and ‘total day charge’ , ‘total eve charge’ , ‘total night charge’ and ‘total intl charge’



People who churn have higher average charges than those who does not churn. Although there is not much difference.

* ‘Churn’ and ‘number customer service calls’

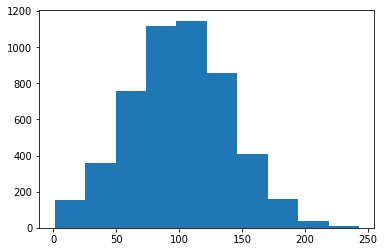


On x-axis we have count of ‘Churn’ and on y-axis we have 'number customer service calls'. The bar graph tends to decrease but there is not a linear tendency.

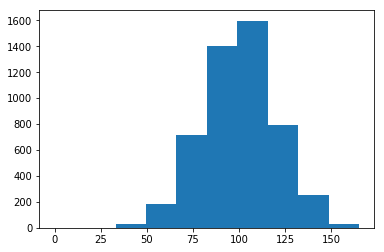
**Univariate Analysis:**

We would now check the distribution of continuous independent variable.

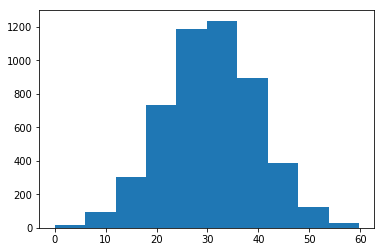
* account length



* Total day calls



* Total day charge



We could see that distributions of most of the continuous variable is normally distributed.

**Scaling and Outlier Analysis:**

For this problem we won’t be scaling the data because our modeling will be done on the algorithm of random forest which does not require the variables to be scaled.

Also our dataset may contain outliers but removing outliers from this data would not be good if we want a better accuracy because our dataset is not very large enough and also the no of people who churn is less in the training data. If we remove the outliers then it would decrease more. This could increase the problem of target class imbalance.

**Sampling:**

In our training dataset the value count for churn is:

False : 2850

True : 483

So we could see that there is a target class imbalance problem.

For this problem one we have to reconstruct our data by resampling. There are two methods for that, Let me explain with an example

a) Under sampling the majority class

b) Oversampling the minority class

Since our dataset is not very large so we will go for second method. We would randomly take 800 more samples of minority class and will merge it to the training data set.

**Modeling:**

**Random Forest Classification:**

As we are completed with most of our work for analysis, now we are only left with the modeling part. For modeling we have used Random Forest Classifier algorithm.

Random Forest adds additional randomness to the model, while growing the trees. Instead of searching for the most important feature while splitting a node, it searches for the best feature among a random subset of features. This results in a wide diversity that generally results in a better model.

Therefore, in Random Forest, only a random subset of the features is taken into consideration by the algorithm for splitting a node. You can even make trees more random, by additionally using random thresholds for each feature rather than searching for the best possible thresholds (like a normal decision tree does).

With the random forest classification algorithm we get an accuracy of about 96%.

Confusion matrix:

[1433, 10]

[ 58, 166]

Here we have 1599 right predictions and 68 wrong predictions.

FNR = (58 / (166+58))

= 0. 2589

**XGBoost Classification:**

XGBoost is an implementation of gradient boosted decision trees designed for speed

and performance.

For this problem Xgboost gave better results. Although the efficiency did not increase

much but the FNR decreases remarkably.

The score for Xgboost is 0.962

Confusion matrix:

[1432, 11]

[ 52, 172]

We have 63 wrong predictions.

FNR = (52 / (172+52))

= 0. 2321