Project Report on Customer churn prediction based on Machine Learning Techniques

By Kamaljeet Singh

Chapter 1

Problem statement

The problem statement of the project deals with the "customer churn". Customer attrition, also known as customer churn, customer turnover, or customer defection, is the loss of clients or customers. Customer churn analysis is usually adopted the banks, insurance, telecom companies etc as one of the key business metrics. The cost of retaining an existing customer is far less than acquiring a new customer. So the companies need to ensure that the existing customer or client won't churn out easily as the attrition of customer would add new expenses to the company. Objective of this case is to predict whether the customer will churn out or not

Data

Based on the problem statement, a classification model required to whether the customer will churn out or not.

We have a total of 20 independent variables and 1 dependent variables. The training dataset contains 3333 observations.

Our dataset contains customer details whether customer have took voice mail plan, international plan or not, call duration, call charges during day time, night time, evening time and international calls. For the churn variable we have two values True or False. False means the customer did not churn out or moved out while the true means that the customer did churn out or moved out. So the prime objective would be to get the correct prediction whether the customer will churn out or not for test dataset containing the same features.

Chapter 2

Data Preprocessing

While building any of the Data Science project it needs to be ensured that the data set contains less or no redundant data. To remove all the unwanted data, Data Preprocessing needs to be done. Data preprocessing is a data mining technique that involves transforming raw data into understandable format. Real world data is incomplete, inconsistent and usually contains too many errors. Although complete removal of the errors is not possible but it can be reduced to quite an extent using Data Preprocessing techniques.

Understanding the data

In our datasetwe have 5 categorical variables (Object type) 16 Numerical Variables (Int and Float Type)

Analysis of the Numerical Data

In the data set 5 categorical variables are present 16 numerical variables (Float and int type). On analyzing the numerical statistics of the given data it is found that minimum value among all numerical values is 0 and maximum is 395. This means that the data is in same range in all numerical columns. A simple machine learning algorithm is based on the distance between two points. This means that higher values will dominate the lower values while calculating the distance. If such is the case, FEATURE SCALING is applied. Since in the given dataset there is no such thing as large differences in the distance between two points. Hence feature scaling is not required.

Analysis of the Categorical Data

The variable named as “State” contains 51 unique values. So feeding this variable will not provide effective results. Moreover while making the dummy variables maximization of the dimension of dataset would be required leading to curse of dimensionality. So it would be better to drop the state column and ‘area code’ would be a better alternative as it contains only 3 categories. For area code and other categories feature selection test will be required to verify the importance

In the given dataset for churn true only 14.5% are related. It is generally found that that if minor class is less than 5% then target class imbalance may occur. In such case target class imbalance will be removed at the end for improvement of model. Target class imbalance is the scenario where the number of observations belonging to one class belonging to one class is significantly lower than the other one. Phone numbers have all different values, in case if this variable is dropped it would not yield significant improvement in the model. Rest categorical variables have only two categories, so they would be used as per the importance derived from the Feature Selection test.

Missing Value Analysis

As per missing value analysis given dataset does not have any missing value analysis. In case of presence of missing value it should be imputed using mean, median, K-Nearest Neighbour, Linear Regression and other similar methods.

Outlier Analysis

An outlier is an observation that lies at an abnormal distance from other values in a random sample from a population. Outlier analysis is also called as anomaly detection. Outlier analysis can be a tricky part when the dataset is small. The boxplot method is used for the outlier detection, if any value is greater than (Q3+(1.5*IQR)) or less than (Q1 –(1.5*IQR))

Where

Q1 > 25% of data or less than or equal to this value or 25^{TH} %ile

Q2> 50% of data or less than or equal to this value or Median or 50^{th} %ile

Q3 > 75% of data or less than or equal to this value or 75^{th} %ile

IQR(Inter Quartile Range) = Q3 –Q1

Outlier should be removed in such a manner that important information is not lost in the process for correct prediction. Before the removal of outlier, the nature of outlier must be known. Some outliers are just an outcomes of human measurement errors.

For the given dataset there is a possibility that the customer is consuming high amount of data or is making lots of calls. In such case Box plot will consider as an outlier but actually it is just information.

Since the given dataset is small in size and removing outlier may also result in reducing the dataset even further and can also cause the loss of important data so the best alternative should be to build the machine learning models and feed the model with two different dataset. One with the outliers and one without the outlier. On getting the results the with two different dataset, further performance tuning will be done as required. Tree based algorithms are insensitive to the outliers compared to the other algorithms. Hence tree based algorithms could provide the best possible outcome without losing information.

Analysis

For total day minutes, eve minutes, international minutes, outliers are on both sides and distribution is also normal

For the charges and minutes the distribution is almost same for such case multicollinearity will be tested in feature selection

‘vmail messages‘ displays binomial distribution due to the fact that most customers have 0 values hence a strong peek displayed in 0 and normal distribution at values other than 0.

Further steps would involve creating a copy of data frame and performing the outlier removal process and feeding the model with two different dataframe (one with the outliers and other without outlier)

Removal of outliers using box plot

Outlier removal will be done in following features

- 'account length',
- '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 charge',
- 'total intl calls'

Outlier removal will not be done for following features

- customer service calls
- number of vmail messages

As these both features more no of 0 and median is more centered towards zero

Analysis

There are very less outliers present even after the outlier removal process because outliers are based on distance from median, after removal of outliers the median shifts towards the center and increasing the distance between other points hence few outlier sre still present.

Feature Selection

For the feature selection the correlation between independent variable must be known. For any machine learning model, its performance decreases with multicollinearity, so it must be ensured that two variables carrying same information must be dropped.

Scatterplot Analysis

As per the scatterplot analysis there is strong collinearity between few columns. There is straight line in scatterplot for four pair of columns

Columns which are showing multicollinearity :

- Total day Minutes and total day charges
- Total eve minutes and total eve charges
- Total night minutes and total night charges
- Total intl minutes and total intl charges

Scatterplot and heatmap Analysiis

From scatter plot and heat map analysis it is found that there is an exact linear relationship between charges and minute for each column of day, evening, night and international call. So both ‘charges’ and ‘minutes’ should be dropped as they contain similar information

Analysis

P-value obtained from the Chi-Square Test among the variables ‘voice mail plan’ and ‘international plan’ with ‘Churn’ is 0.05 rather very less. Evidently we can reject the ‘null hypothesis’ and accept the ‘alternate hypothesis’. Large P- values between ‘churn’ and ‘area code’ indicates the independence of the variables. So the variable area code should be dropped

‘Voicemail plan’ and ‘International plan’ are independent to each other and are not multicollinear, moreover they both have P-value sufficient enough to reject the “Null Hypothesis”.

So these two variables should be taken in the final dataset.

Analysis of numerical variables

VIF Analysis

VIF value for any variable having less than value 10 depicts the multicollinearity. Multi collinear variables should be dropped from the final dataset. Hence ‘total day minutes’, ‘total eve charge’, ‘total night charge’, ‘total intl charge’ will be dropped.

Feature Scaling

For the given set of data Feature Scaling is not required. Feature scaling is important for those variables for which their values are either too high or too low. Algorithms using distance method, are affected by out of range values. Higher value dominates the lesser values in calculating distance. Since the given dataset has a uniform distribution so feature scaling is not required.

Building the Machine Learning model using Algorithm Stated in the Coursework

key terms to avoid any confusion in next steps.

- train_set: training dataset containing all observations.
- No_outlier: training dataset , containing observation which left after outlier removal
- X_train : containing independent variables of train_set
- y_train: containing dependent variable (Churn) of train_set
- X_train_wo : containing independent variables of no_outlier
- y_train_wo : containing dependent variable (Churn) of no_outlier

Cross Validation

Before Building models cross validation must taken into consideration

K-fold Cross Validation

K-fold cross validation is used to check performance of model which is checked on K different test dataset. K-fold cross validation helps to divide the training data in k sets and will build a model using k-1 training set and one left set would be used to test the performance. In this way it would build k times model and each time there would be different test dataset to check performance and at the end all k model's accuracy mean value would be considered as model accuracy.

Grid Search CV

GridsearchCV

Hyperparameter are the parameters which are passed as argument to the building functions, like kernel, criterion, n_estimators etc. So to get best values of these, gridserchcv is used. In this technique, a list of these different parameters are made and then gridsearchcv build model for every combination of these parameters and then check crossvalidation score and based on score it, further it gives the best combination of hyperparameters.
And then the model is build with the values of hyperparameter given by GridSearchCV.
This is called performance tuning and this would be used to tune the model..

Model building

Model Performance:- For the given problem statement. Classification model is required and then model performance will be used to decide the final model and algorithm.

Logistic Regression

Analysis

Almost same performance for both dataset is obtained i.e. 86% K- fold accuracy. Logistic regression gives better result for linearly separable data. 183 observations are predicted as churning False and in actual these observation were Churning as True.

(K – Nearest Neighbors)
Analysis

Almost same performance for both dataset is obtained i.e. 85% K-fold accuracy which is slightly less accurate than logistic regression. 207 observations are predicted as churning False and in actual these observations having Churning as True.

Naïve Bayes
Analysis

Almost same performance for both dataset is obtained i.e. 85% K-fold accuracy and slightly less accurate than logistic regression. 135 observations are predicted as churning False and in actual these observations having Churning as True.

Decision tree:

Analysis

92% K-fold accuracy for dataset with outliers and 91% K-fold accuracy for dataset without outliers is obtained. A slightly higher performance for dataset with outliers than dataset without outliers is recorded. 65 observations are predicted as churning False and in actual these observations have Churning as True.

So, not drop outliers should be the correct decision moreover Decision tree algorithm outperformed the other models.

Hyperparameter tuning

Hyperparameter tuning is used to find optimum values of arguments used in building models like n_estimators, max_depth, kernel etc. so that better result with these tuned parameter can be gained. So hyperparameter tuning will be done for two models which gave us accuracy more than 90% i.e. Decision Tree Classifier and Random Forest

Decision Tree Model Hyperparameter tuning:

Tuning decision tree for following parameters on dataset with outliers i.e. train_set

Now, building DecisionTreeclassifier with parameter suggested by GridSearchCV for dataset train_set.

Analysis

Decision Tree model is improved after parameter tuning. For dataset with outliers model is improved from 92.26% to 94.42% K-fold accuracy and for dataset without outliers improvement is from 91.95% to 94.03% K-fold accuracy.

Chapter 3

Conclusion

3.1 Final Model and Training Dataset

From the above models we selected below dataset and model for predicting our test dataset. As below model giving us less error in predicting true churning which was our main motive to reduce churning.

Dataset:

• First take whole training dataset.

• Drop columns ‘area code’, ‘state’, ‘phone number’, ‘total day minutes’, ‘total eve charge’, ‘total night charge’, ‘total intl charge’.

• Change ‘international plan’, ‘voice mail plan’ and ‘Churn’ columns to category and then to levels of category (0 and 1)

• Do same thing with test dataset

Model:

• Use random Forest model and train using dataset which we prepared with above steps.

• Perform hyperparameter tuning.

• Build model using tuned hyperparameter.

• Prediction can be obtained from model