**Laxmi Prasanna Chennadi(jg55734)**

**Spring 2020**

**DATA 606 Capstone Project**

**Analysis of Telecom Churn Data**

**Project Overview:**

In the current market many numbers of telecommunication networks are available, and we have luxury to choose one of them according to our requirement. The increased number of telecom networks in the market has led to increased number of customers moving to another network. This has become very common problem in Telecom industry and we often come across such scenarios in our daily life and this is another factor effecting the revenue of the organizations. The main goal of my project is to predict churn of telecom data which is nothing but number of people leaving current network and moving to another company. I want to analyze the factors that affect churn which in turn would be helpful for companies in retaining their customers and in acquiring new ones. To accomplish my goal, I am planning to build classification and regression models, furthermore I would perform few Machine Learning algorithms, evaluate their accuracy, and then choose the best fit algorithm for the business.

**Motivation:**

Being a super enthusiastic socialite, I have started using gadgets from a very young age and mobile phone is one of them without which no daily activities can be completed with ease. Approximately 5-6 years back, during my high school, back in my country, many telecom networks entered the market, and customers started jumping from one network to another. That was the first time when question arouse in my mind why people want to move to another network? What could be the reasons? Later, after 2 years, another network came into picture which was adopted by 60% and more population of the country. This network even surpassed most of the oldest and experienced companies which were in the same field since long time. This circumstance further had more impact on me. These were the two instances which motivated me on working on this project.

**DataSet:**

The dataset which will be used for this project is Telecom data which comprises of 6999 entries and 21 columns which describe the various factors that affect churn and churn column would be my target variable. It is available in Kaggle. 21 columns of dataset are, State, account length, area code, phone number, international plan, voice mail plan, number of voicemail messages, total day minutes, total day calls, day charge, evening minutes, eve calls, evening charge, night minutes, night charge, international minutes, international calls, international charge, customer service calls, churn etc.

**SOURCE:** <https://www.kaggle.com/becksddf/churn-in-telecoms-dataset>

**Methodology:**

Initially I would start with data preprocessing, i.e data cleaning, removing NaN values. Then I will perform exploratory data analysis, observe the factors influencing churn column and create visuals of the data accordingly. My next step would be applying logistic regression and Random forest classification algorithms. Later perform few other machine algorithms like KNN, PCA etc, evaluate their accuracy and choose the best algorithm which serves the business purpose. With this we will be able to predict churn, control and also improve the revenue of the company.

**Literature Survey/Preliminary Work:**

Churn analysis is the evaluation usually done by the organization on the customer’s loss rate. It is very helpful for many organizations as it will help them in analyzing the reason for the churn and come up with precautionary methods as well as solutions for solving the issues related to churn in an organization. Frequent analysis of churn data is very important. The organizations are very interested in analyzing these trends and use it to reduce churn rate as the price for acquiring a new customer is usually higher than retaining the old one. It is important to deal with churn for every organization in the consumer market and enterprise sectors as it has an impact on the revenue and policy decisions of the organization. According to the authors of “Leading on the Edge of Chaos”, a 2% increase in customer retention (or decreasing churn) is equivalent to 10% reduction in costs. So, it is no wonder that companies that care about customers pay a lot of attention to Churn Analysis.

**Exploratory Data Analysis:**

Initially, we perform the basic steps of data loading and analyzing to understand the characteristics of the dataset. We then start the data cleaning process where we check for missing values as well as other inconsistencies. In this dataset, we can observe that there are no missing values. We perform conversion of data types to make the data more reliable and relevant for analysis.

A screenshot of a cell phone

Description automatically generated

To do the exploratory analysis, heat map and correlation methodology is used to understand the correlation between the columns. From this, we can observe that there is correlation between churn and few other columns like total day charge, customer service calls and international plan. They are dependent on each other and by analyzing this using visualization techniques, we can find out the dependency (direct or inverse). We can find out the impact of these columns on the churn rate.

A picture containing object, clock

Description automatically generated

From the above map, we can see that churn has highest correlation with columns total day charge, total day minutes, total eve charge, total eve minutes, total night charge, total night minutes,total intl charge and total intl minutes.

The columns that have highest correlation with the churn variable are the total\_day\_charge, the total\_day\_minutes and the number of customer service calls.

MULTICOLLINEAR VARIABLES Night Mins, Night Charge 0.999999

Eve Charge, Eve Mins 1.000000

Day Charge, Day Mins 1.000000

Intl Charge, Intl Mins 0.999993

**A picture containing blue, toy, red, woman

Description automatically generated**

**As day charge increases, rate of churn is also increasing**

**A picture containing drawing

Description automatically generated**

**From the above graph, we can observe that churn is directly proportional to customer service calls. That is, churn is increasing with increase in customer service calls**.

**A screenshot of a cell phone

Description automatically generated**

**churn rate is high for customers with international plan when compared to customers with no international plan.**

**Data Preparation:**

Moving forward I tried detecting the outliers in the columns and tried clipping them for better analysis. Later, to prepare data for Model building I have created two data frames with predictor variables and target variable. Here international plan and voice mail plan are categorical types(yes/no), pd.get\_dummies are used to convert them into numerical (1/0). Now data is normalized i.e, all features are set to one range of scale for feature scaling of training data using Min-Max scaling is done, by importing MinMaxScaler from sklearn.preprocessing. Later Features and Target data frames are split into training and testing sets using train\_test\_split from sklearn.model\_selection. Finally, now that data is prepared, we can proceed with Model Construction.

**Model Building:**

Our goal is to predict customers who are most likely to churn or change telecom networks. As it turns out to be a classification problem, I am using Logistic Regression from regression algorithms, K-Nearest Neighbors to learn how Nearest-Neighbors would work on this dataset and Random forest algorithms from decision trees family to perform classification analysis on data to predict churn.

Applying Binomial Logistic regression would give better results and tried fitting the model on the Features\_train and Target\_train dataframes and performed prediction on train data where 77% of accuracy is recorded and applying the same classifier on test data i.e on target data frame, logistic regression gives 78% accuracy.

Secondly, I applied k-Nearest s

Neighbors classifier by iterating k value from 1-10 and plotting the accuracy of each iteration shows 1 as the optimum value of k, therefore by taking n\_neighbors as 1, I tried fitting the model on train data and obtained 100% accuracy but when applied on test data the accuracy rate has dropped by 14% which happens in case of overfitting of data. Overfitting occurs when a model learns the detail and noise in the training data to the extent that it negatively impacts the performance of the model on new data. This means that the noise or random fluctuations in the training data is picked up and learned as concepts by the model. The problem is that these concepts do not apply to new data and negatively impact the model’s ability to generalize.

Random Forest Classifier is an ensemble method which combines more than one algorithm and it belongs to the family of decision trees. It is based on the principle of creating a set of decision trees from random subsets of a training set and then aggregates their votes to decide the final class of test object. When applied on train set 95% accuracy is observed and on test set 94% which clearly depicts Random Forest Classifier has given best results so far.

Later I observed unique identity columns like state, area code and phone number which are just providing information about customers, though these columns do not have any significant impact on the churn column, but similarities between data points of these columns might impact churn. For example, if customers from single state or area code are changing networks this analysis helps to detect such patterns from data. So, to predict such patterns we are creating a new column called cluster by detecting similarities among all the rows and giving them a group id, which is a cluster.

K-Means clustering is used to cluster data points with similar properties and will help train models to learn from those similarities. How many clusters to create? To know the number of clusters to create, find WCSS(within cluster sum of squares) using k-means classifier by iterating k value in range of 1-10 and plotting an elbow curve to know where the value of WCSS tends to flatten. By applying this on our dataset gives 5 as the optimum value of k. Why do we choose k as the optimum value where the elbow curve flattens? This method calculates SSE (sum of squared distance) between each data point and mean of the cluster. As we go on increasing the value of k WCSS becomes zero which is not ideal to find the similarities. So, we choose the value of k where the curve starts to flatten.

After creating clusters, I have observed that there are columns have multicollinearity, applying dimensionality reduction on these features reduces the computational complexity by reducing the number of samples created by model and interpreting parameters of the model. I chose PCA (Principal Component Analysis) algorithm and created two new components. After creating these 2 new components, a new data frame is created by reloading the dataset to append two PCA components to churn, international plan, voicemail plan and clusters.

As a final step, after altering the dataset I have applied the Random-Forest classifier with an expectation of obtaining better results. On fitting on train data accuracy of 100% has been recorded whereas, on test data prediction accuracy has dropped to 88-89%. This was supposed to be the best model, but an overfitting problem is observed.

**Model Results:**

**Predictions on Test sets:**

A screenshot of a cell phone

Description automatically generated

**Clustering+ PCA+Random Forest Classifier:**

**Train and Test set:**

A screenshot of a social media post

Description automatically generated

Here, Clustering+ PCA+ Random Forest was supposed to give better results, but over fitting has occurred. Usually the problem of overfitting is observed when there is high variance in the data. PCA performs dimensionality reduction by reducing large set of variables to smaller ones and still contains important information. These variables are called principle factors or principle components and these components capture maximum variance within data.

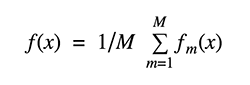
In PCA dimensionality reduction is performed either by feature elimination i.e removing features from the data which have less importance which is done previously by creating 2 principle components or by feature extraction where we create more independent variables which are the combination of already existing variables. These components are ordered accordingly on how well they predict our target variable and remove least important ones. PCA is also used to overcome overfitting by reducing the variance. Hence, I have performed feature extraction by creating 15 new components using 15 old components. After all the hard work of creating clusters, PCA components and feeding it to Random Forest Classifier 1% increase in accuracy on test data is recorded compared to simple Random Forest classifier model. Though it is just 1%, model got better in predicting True churn data by 2% on testing data.

A screenshot of a social media post

Description automatically generated

**Bagging:**

Since only 1% increase is observed, I was not satisfied with the results and further decided to proceed with Ensemble Methods. Ensemble methods are meta-algorithms that combine several algorithms into one predictive model. They are Bagging, Boosting, Stacking etc. Here, Bagging is applied to reduce the variance whereas boosting is used on highly biased and to improve predictions stacking is used. Since, we have issue of overfitting I have decided to perform Bagging. Bagging itself stands for bootstrap aggregation. One way to reduce the variance of an estimate is to average together multiple estimates. For example, we can train M different trees on different subsets of the data (chosen randomly with replacement) and compute the ensembl

Bagging uses bootstrap sampling to obtain the data subsets for training the base learners. For aggregating the outputs of base learners, bagging uses voting for classification and averaging for regression. On applying Bagging, accuracy on train set is 96% whereas, accuracy on validation set is 95%.

A screenshot of a cell phone

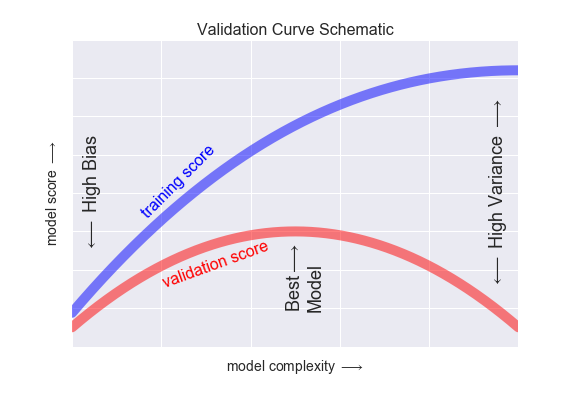
Description automatically generated

From the above output, I observed that accuracy on applying bagging is recorded similar to that of Random Forest Classifier.

**Evaluating Performance of the Model:**

* For high-bias models(underfitting), the performance of the model on the validation set is similar to the performance on the training set.
* For high-variance models(overfitting), the performance of the model on the validation set is far worse than the performance on the training set.

If we imagine that we have some ability to tune the model complexity, we would expect the training score and validation score to behave as illustrated in the following figure:



The diagram shown here is often called a *validation curve*, and we see the following essential features:

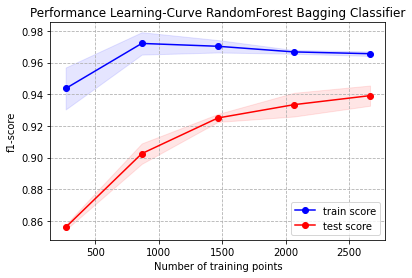
* For very low model complexity (a high-bias model), the training data is under-fit, which means that the model is a poor predictor both for the training data and for any previously unseen data.
* For very high model complexity (a high-variance model), the training data is over-fit, which means that the model predicts the training data very well but fails for any previously unseen data.
* For some intermediate value, the validation curve has a maximum. This level of complexity indicates a suitable trade-off between bias and variance.

Performance of the model is evaluated by using validation\_curve. is used which establishes relation between training score and validation score. This gives useful insight on performance of model. Performance of model is at its best at the intermediate point where test accuracy is maximum which is a suitable tradeoff between bias and variance. One important aspect of model complexity is that the optimal model will generally depend on the size of your training data. Hence, validation curve has 2 important inputs. They are:

1. Model Complexity
2. Number of Training Data points.

Plot of training/validation score with respect to size of training data set is known as **“Learning Curve”.**

The notable feature of the learning curve is the convergence to a score as the number of training samples grows. Once you have enough points that a particular model has converged, adding more training samples will not help in improving the model.

****

Hence, in order to find relevance of model with training data inputs, I have plotted a learning curve, from the above plot we can understand that model is at its best and further increase in training data points does not improve the model. The only way to increase model performance in this case is to use another complex model.

**Grid Search:**

A model parameter is a configuration variable that is internal to the model and whose value can be estimated from the given data. Whereashyperparameter is a characteristic of a model that is external to the model and whose value cannot be estimated from data. The value of the hyperparameter has to be set before the learning process begins. Grid-Search is used to find the optimal hyperparameters of a model which results in the most ‘accurate’ predictions. Grid search is an approach to hyperparameter tuning that will methodically build and evaluate a model for each combination of algorithm parameters specified in a grid.

To tune the hyperparameters and to improve the model performance, I have applied Grid-Search. On using this parameter tuning model, the accuracy rate is observed as 99% on training set and 96% in test set i.e score has increased further by 1% more when compared to Bagging classifier.

A screenshot of a cell phone

Description automatically generated

**Conclusions:**

* The rate of churn i.e at which people move from one network to other is 14.491%.
* Target variable Churn is affected by day charge, Customer service calls and International plan.

1. As day charge increases, rate of churn is also increasing
2. Churn is directly proportional to customer service calls. That is, churn is increasing with increase in customer service calls.
3. churn rate is high for customers with international plan when compared to customers with no international plan.

* Successfully applied various Machine Learning algorithms (Logistic Regression, K-Nearest Neighbors, Random Forest Classifier, K-clustering, and PCA) as stated in the proposal and calculated the accuracy of various models.
* To improve the model and to perform hyper parameter tuning, I have successfully applied Ensemble method (Bagging) and Grid Search.
* From my analysis, I have observed that Random Forest Classifier is best model fit to the data set being an ensemble method it gave better results. Though I have applied Bagging, Grid Search, the accuracy has increased only by 2% even after lot of hard work.

**Learnings:**

* Learnt to apply various model and explored various classification and regression algorithms.
* What is overfitting? Approaches to over-come the problem and its relevance to clustering and Principle Component Analysis.
* Best way to evaluate the performance of the model and factors effecting the validation curve.
* Feature Engineering is another important concept which I understood deeply.
* Learning Curve is another concept which I have learned and how size of training data points effect performance of the model.

**References:**

* <https://www.researchgate.net/publication/321723500_Review_of_Customer_Churn_Analysis_Studies_in_Telecommunications_Industry>
* <https://towardsdatascience.com/machine-learning-powered-churn-analysis-for-modern-day-business-leaders-ad2177e1cb0d>
* <http://www.computerscijournal.org/vol10no1/churn-analysis-in-telecommunication-using-logistic-regression/>
* [https://journalofbigdata.springeropen.com/articles/10.1186/s40537-019-0191-6](https://journalofbigdata.springeropen.com/articles/10.1186/s40537-019-0191-6  )
* <https://jakevdp.github.io/PythonDataScienceHandbook/05.03-hyperparameters-and-model-validation.html>
* <https://blog.statsbot.co/ensemble-learning-d1dcd548e936>
* <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4792409/>
* <https://towardsdatascience.com/a-one-stop-shop-for-principal-component-analysis-5582fb7e0a9c>
* <https://towardsdatascience.com/ensemble-methods-in-machine-learning-what-are-they-and-why-use-them-68ec3f9fef5f>