**Customer Churn Analysis for Telecommunication company using Machine Learning Techniques**

* **Jayashri Dandare**

**Abstract :**

In telecommunication companies, ‘churn’ means customers’ decision to move from one service provider to another. The competition environment in telecom companies' makes their aim is to maintains their customers who are likelihood to leave and earns their satisfaction, so to avoid the problem of churn they need churn predictive models. Data mining techniques can be used to build churn prediction model for telecommunication companies to identify churner and non-churner customers because it can extract the predictive information from large databases.

**Introduction:**

Churn is a term used in many companies which is mean loss of customers of the company for many reasons one of them is the dissatisfaction of customer. In telecommunication companies “churn” term refers to customer's decision to leave the current service provider and move to other service provider, it can be easily happened especially for prepaid customers because they have not any contract same as to post-paid customers.

Churn occurs easily because of the strong and breeding competition environment in services which are providing especially in telecommunication sector, also churn can be happen for another rezones for examples customer's dissatisfaction with services and high cost of these services which can be in another service provider with best quality and lower cost. So, churn become a concern issue in telecom sector because retaining of existing customer is costly than acquiring new one. To identify churner and non-churner customers and understand the rezones of this churn to reduce it, these companies can build churn prediction model which can help them in churn issue to build this model they can use Data mining (DT) this can be useful because DM can extract a predictive information from large databases, it's had many techniques which can for example: Grid Search CV, Random Forest Regressor etc.

For all companies that bill customers on a regular basis, one of the main variables is churning. Having worked on some telecom projects in a professional capacity, this area wasn’t new for us.

However, applying ML techniques to predicting customer behavior isn’t something that either one of us has done before. We were keen to see if we could identify the key reasons behind customer churn & if possible, create a model that would help retain the same customers. We came across a telecom company’s customer churn [dataset](https://www.ibm.com/communities/analytics/watson-analytics-blog/guide-to-sample-datasets/) on the IBM Watson Analytics website.

**Problem Statement:**

Customer churn is when a company’s customers stop doing business with that company. Businesses are very keen on measuring churn because keeping an existing customer is far less expensive than acquiring a new customer. New business involves working leads through a sales funnel, using marketing and sales budgets to gain additional customers. Existing customers will often have a higher volume of service consumption and can generate additional customer referrals.

Customer retention can be achieved with good customer service and products. But the most effective way for a company to prevent attrition of customers is to truly know them. The vast volumes of data collected about customers can be used to build churn prediction models. Knowing who is most likely to defect means that a company can prioritise focused marketing efforts on that subset of their customer base..

**Problem Objective:**

The objective of this problem is to build a churn prediction model which can identify churner customers and non-churner customers and implementing this model by using Random Forest Regressor Method.

**Methodology:**

The dataset consists of 21 variables in all. A few are continuous, rest are categorical. The control variable was customer churn with 2 levels Y/N (i.e. customer has left or not). Initially, we started out with some basic EDA & some visual plots that would help us understand the data better. Since many machine learning algorithms cannot operate on categorical variables, we had to convert this data into numerical variables. We decided to use the Ordinal Encoder method to convert the variable type. After the initial analysis, we proceeded to use techniques like Feature Importance & Feature Selection to see if we had any variables that were redundant & could be discarded in the process of building the models. We decided to use models like Linear Regression, Min-Max Scaler, Grid-Search CV, Random Forest Regressor for this analysis. We split the data into train & test & built a model using each of these classifiers. We used Cross-validation to evaluate the quality of the predictive models by partitioning the original data into a training set to train the model, and a test set to evaluate it.

### **Dataframe Description:**

The dataset contains the data of the employee. On the basis of the data we have to predict the attrition of company. The dataset contains tha data like 'customerID', 'gender', 'SeniorCitizen', 'Partner', 'Dependents', 'tenure', 'Phone Service', 'MultipleLines', 'InternetService', 'OnlineSecurity', 'OnlineBackup', 'DeviceProtection', 'TechSupport', 'StreamingTV', 'StreamingMovies', 'Contract', 'PaperlessBilling', 'PaymentMethod', 'MonthlyCharges', 'TotalCharges', 'Churn'

In the above dataset the target is to predict the 'Total Charges' of the company.

Churn rate (sometime called attrition rate), in its broadest sense, is a measure of the number of individuals or items moving out of a collective group over a specific period. It is one of two primay factors that determine the steady state level of customers a business will support. The term is used in many contexts, but is most widely applied in business with respect to a contractual customer base, for example in business with a subscriber-based service model such as mobile telephone networks and pay TV operators. The term is also used to refer to participant turnover in peer-to-peer networks. Churn rate is an input into customer lifetime value modeling, and can be part of stimulator used to measure return on marketing investment using marketing mix modeling.

**Data Analysis:**

For the project scope, Jupyter Notebook is used with Python 3 environment. Sample dataset is used from IBM imaginary customer churn data. Features and format of the data can be seen in Table 1. Before machine learning algorithms are implemented, data cleaning and transformation steps are provided. Dataset included, firstly, 7043 rows and 21 features. These features are grouped into demographic parts, location, services provided by company, payment methods etc. The customer Data Preprocessing Churn Prediction Model Evaluation Customer Segmentation Figure 1 - Analysis Framework will be churn or not. Hence, we can eliminate these columns before machine learning algorithms are implemented and tested.

Table 1 : Features and format of the data

|  |  |
| --- | --- |
| **Features** | **Format** |
| Gender | Male/Female |
| SeniorCitizen | Yes/No |
| Partner | Yes/No |
| Dependents | Yes/No |
| Tenure | Numerical |
| PhoneService | Yes/No |
| MultipleLines | Yes/No/No phone service |
| InternetService | DSL/Fiber optic /No |
| OnlineSecurity | Yes/No/No internet service |
| OnlineBackup | Yes/No/No internet service |
| DeviceProtection | Yes/No/No internet service |
| TechSupport | Yes/No/No internet service |
| StreamingTV | Yes/No/No internet service |
| StreamingMovies | Yes/No/No internet service |
| Contract | Month-to-Month /One Year /Two Year |
| PaperlessBilling | Yes/No |
| PaymentMethod | Bank Transfer /Credit Card /Electronic Check / Mailed Check |
| MonthlyCharges | Numerical |
| TotalCharges | Numerical |
| Churn | Yes/No |

**EDA Concluding Remark:**

We started out with some basic EDA. We have around 7000 observations. We got rid of 11 incomplete observations from our dataset. We have 20 features & 1 target variable. Only 4 features out of 19 were numeric. All the rest were categorical variables. Our customer churn v/s stay data split is in the ratio of 1:3. Ratio of males to females is around 50:50. When we plotted the Churn variable against the customer’s tenure, one interesting observation was that most of the customers who leave the telco provider, usually do it within the first year. Beyond the first year, they tend to stick around. Another interesting bit is that their biggest customer base are their oldest customers followed by the newest. Obviously, we would not like to have multicollinearity in our dataset. So, we plotted a heatmap/correlation plot for all the variables. The variables with the highest positive correlation are TotalCharges, MonthlyCharges & Tenure. Looking at the data, we figured that TotalCharges is nothing but Tenure times MonthlyCharges. Hence, we discarded the TotalCharges variable. Now, customerID is another such redundant feature, so we got rid of that variable, too.

**Pre-Processing Pipeline:**

As the first step in our data preprocessing, we split the variables into Feature variables & Target variables. Then, after this we split our data into training, & testing. To bring the variables on the same scale, we standardized the data. MonthlyCharges & Tenure are the features with the highest Feature Importance (~22%). No other feature exceeds more than 10% in Feature Importance. In order to check if we can get rid of features that are not contributing to the model much, we tried to use the Recursive Feature Elimination method. The output of the RFE is surprising since it mentions that the optimal number of features are **21**. We decided to include all the features in our analysis & move ahead with our model building & comparison.

1. **Identify domain problem**: Churn prediction is a phenomenon which is used to identify the possible churners in advance before they leave the network. This helps the CRM department to prevent subscribers who are likely to churn in future by taking the required retention policies to attract the likely churners and to retain them. Thereby, the potential loss of the company could be avoided. This research utilizes data mining techniques to identify the churners.
2. **Data Selection**: Churn prediction models require the past history or the usage behavior of customers during a specific period of time to predict their behavior in the near future. Data Acquisition Identify Problem Data Selection Data Preparation Data mining technique (Naïve Bayes Algorithm) Churners NonChurners 8 | P a g e is a difficult problem for the researchers to acquire the actual dataset from the telecom industries. This is because the customer’s private details may be misused.
3. **Data Preparation**: After collection of required datasets, we require to prepare the dataset so that they can be processed by appropriate tools. These datasets are generally large volumes of unstructured data which cannot be handled by traditional data processing techniques. So for processing them we require special computing frameworks which process the data and makes the data ready for analysis to be applied on.

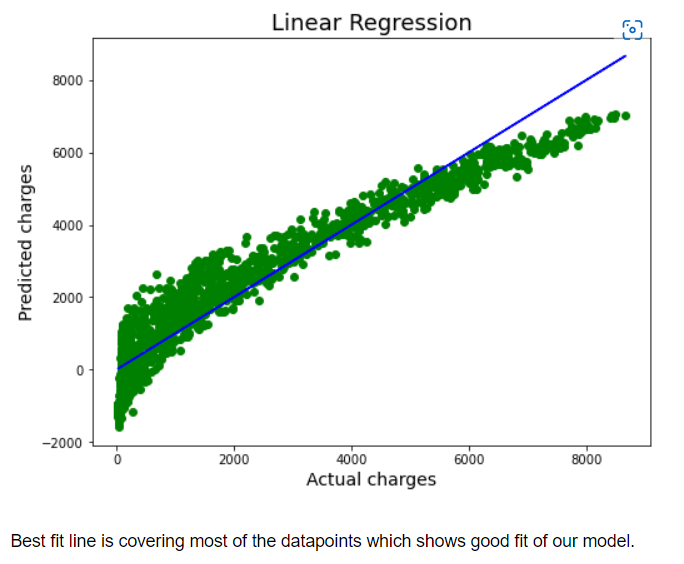
**Building Machine Learning Models:**

We've used quite a few models to check which fits best on our data. Models used are –

* Linear Regression
* Random Forest Regressor
* Lasso
* cross\_val\_score
* GridSearchCV

Our target variable is a binary variable. The customer either stays or leaves. Hence, we decided to use the Logistic Regression as our first model to check how the model fits the data.

To check the model’s predictive performance, we’ve used the Cross Validation across almost each and every classifier barring Random Forest Regressor (since that is essentially what a Random Forest does). We plotted the Linear Regression curve for all the classifiers as a diagnostic test to evaluate the performance of the models.



**Future Enhancement:**

The efficiency that we have got after the calculation is almost 99.8% which can be enhanced using some developed formula and further calculations, so that it will be easier for us to calculate the churn probability of a customer, which will provide better stability and better flexibility.

**Conclusion:**

\* It would be interesting to have more features, some continuous ones preferably, in our model & do this exercise all over again. This would probably increase the efficiency of our models.

\* One of the more surprising outputs was that Recursive Feature Elimination (RFE) didn’t help eliminate any redundant features.

\* We came across a couple of articles that mentioned that Random Forest Regressor categorical variables better than continuous variables. That did not seem to be the case in our models.

\* Based on the model scores, to predict customer churn Linear Regression seems to be the best model for this dataset.

\* We are getting model accuracy and cross validation both as 99.8% which shows our model is performing extremely well.