A BLOG ON CUSTOMER CHURN ANALYSIS

BY

SATU VINAYAK POLE

(DataTrained Batch DS0522)

**Introduction about topic**:

The meaning of customer churn is that when a customer stops doing business with that company. In this case the data is all about telecommunications company that has customers data and kind of services their customers are subscribed to and whether that customer will be renewing that subscription or not. From this dataset I need to analyse that data and predict why there is customer churn and which features are responsible for it, so that company can improve that area to minimize churning rate.

**Problem Defination:**

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.

Preventing customer churn is critically important to the telecommunications sector, as the barriers to entry for switching services are so low.

In this case we will examine customer data from IBM Sample Data Sets with the aim of building and comparing several customer churn prediction models. As said earlier this models will be used by tele companies to understand the reason for customer churn and stopping it as far as possible.

**Data Analysis:**

Jupyter notebook was used to carry out all the python operations pertaining to this project. For start, pandas , numPy, seaborn and matplotlib libraries were imported. After importing libraries, dataset was loaded as csv file and it was printed.

Dataset contained 7043 rows and 21 columns. The columns are mentioned and described as:

1. customerID- represents the customer ID for every customer
2. Gender- whether customer is male or female
3. SeniorCitizen- whether the customer is a senior citizen or not
4. Partner- whether the customer has partner or not.
5. Dependents- whether the customer has dependents or not.
6. Tenure- Number of months the customer has stayed with the company
7. PhoneService- Whether the customer has a phone service or not
8. MultipleLines- Whether the customer has multiple lines or not (Yes, No, No phone service)
9. InternetService-Customer’s internet service provider (DSL, Fiber optic, No)
10. OnlineSecurity-Whether the customer has online security or not (Yes, No, No internet service)
11. OnlineBackup-Whether the customer has online backup or not (Yes, No, No internet service)
12. DeviceProtection- Whether the customer has device protection or not (Yes, No, No internet service)
13. TechSupport-Whether the customer has tech support or not (Yes, No, No internet service)
14. StreamingTV- Whether the customer has streaming TV or not (Yes, No, No internet service)
15. StreamingMovies- Whether the customer has streaming movies or not (Yes, No, No internet service)
16. Contract-The contract term of the customer (Month-to-month, One year, Two year)
17. PaperlessBilling-Whether the customer has paperless billing or not (Yes, No)
18. PaymentMethod- The customer’s payment method (Electronic check, Mailed check, Bank transfer (automatic), Credit card (automatic))
19. MonthlyCharges-The amount charged to the customer monthly
20. TotalCharges- The amount charged to the customer annually.
21. Churn- Whether the customer has churned out or not.

Further, the dataset was checked for null values and it returned that it had zero null values in it. The dataset was also examined if it had any duplicate values.There were no duplicate data seen in dataset.

While examining value counts for all columns it was noticed that the customerID column has all the unique values in it. This means for each customer unique customerID was seen. This kind of column data is not useful for model building as such data lowers the training and testing accuracy of machine learning model.Hence this column was dropped and dataset was saved.

Countplots were plotted for every column to check value count using visualization methods. Also barplots were plotted to show relation between label(Churn) and features (columns). Further it was seen that TotalCharges column had empty spaces in some rows. This I was able to identify as the TotalCharges column was full of float values but its datatype was object. Once locating the empty spaces they were replaced with null values I.e. nan’s which inturn were filled with mean values of TotalCharges column.

The dataset with ‘object’ datatype were encoded using label Encoder. Further, a heatmap of correlation within every features and label was plotted which showed that TotalCharges column was showing multicollinearity with tenure column. This analysis was confirmed with scatter plot which showed positive correlation.

**EDA Concluding Remark:**

The conclusion of findings from EDA are summarized below

* No null and duplicate values in dataset. One column had presence of empty spaces which were later removed and cleaned.
* The columns with object datatype were encoded using Label encoding method.
* The strongest positive correlations with the target features are: Monthly Charges, paperless billing, senior citizen and payment method.
* The strongest negative correlations with the target features are: Contract, tenure, online security,tech support and total charges.

Other observations include:

* It is seen that most number of customers that pay high monthly charges are likely to churn out.
* Also higher number of month to month contract customers are likely to stay with present company.
* Most people that have subscribed to services like Streaming movies,streaming TV, device protection, Online backup and security and phone services are very likely to continue with present company.
* Churn rate is less for customers who have opted for DSL and fiber as internet service.
* Churn rate is more for recent customers I.e. users within 6 months. Older customers that are happy with present services are tend to stay with present company.
* Most Senior citizens are not likely to churn out.

**Pre-processing Pipeline:**

This includes steps undertaken to prepare the datasets for machine learning algorithm and its implementation.

Encoding: The dataset with ‘object’ datatype columns were encoded using Label Encoder I.e. the data was not converted into numerical values.

Outlier and Skewness Removal: The dataset with continuous columns were examined for outliers and no outliers were seen. It was further identified that TotalCharges column had skewness present which was later removed using log transformation.

Feature Selection technique,data scaling and splitting dataset : The dataset was split into X and Y variables as X representing all the feature columns and Y representing only label column and applied with SelectKBest, f classif feature selection method to identify 16 best features that are related to label. Further new dataset was formed with new features and was scaled using Standard Scaler method. This scaled feature data was further used in machine learning algorithm.

Handling Imbalanced Classes of Label: Since the Label classes were imbalanced, it had to be balanced using SMOTE method which generated synthetic samples for minority class to make both classes equal or to a number specified.

**Building Machine Learning Models:**

After performing all the exploratory data analysis which is useful in understanding the dataset, data cleaning, and performing all pre-processing pipeline steps, the dataset was ready to build machine learning model. The steps to build machine learning models are as follows:

* The data was split into train-test split. This splits the label and feature data into 4 parts namely x\_train,x\_test,y\_train and y\_test. Train data represents data that will be used model training and test represents data that will be used for model testing.
* The first model that was trained was Random Forest Classifier and to find this model’s accuracy best random state which gives max.accuracy was found out.
* After finding best random state the model was again trained for training data and tested on test data. Further the accuracy score, confusion matrix, classification report and f1\_score was printed. The cross validation score which shows whether the model is overfitted or not was also printed.
* Same process was implemented for more 3 model algorithms namely AdaBoost Classifier,Decision Tree classifier and Gradient Boosting classifier and earlier mentioned scores alongwith model accuracies were printed.
* All the scores and accuracies of each model were compared with every model and best model was selected. The model that has higher training and testing accuracy, has almost similar or nearly similar F1 and cross validation scores (implying that model is not overfitting) to testing is considered to be the best model.
* In this case, Random Forest Classifier model was considered best model as it had best training and testing accuracy with respect to other models and its cross validation score was same as testing score.
* The best model is always further tuned with a hope that we can increase the training and testing accuracy but the best parameters that will help to increase the score are not known always. Hence its always a trial and error method. This is done using GridSearchCV method which is a type of hyperparameter tuning. Here all the parameters of your model are specified and GridSearchCV runs combination and permutations on best model considering all those parameters. After completing this process best parameters are fetched that were responsible to provide best test and training scores based on parameters specified. This is always a time consuming process.
* Further this parameters are fetched into new model of same algorithm that was used above and the model is again trained and tested with accuracies and scores.
* This accuracy provided this model was less compared to the original Random Forest classifier model hence original best model was saved.
* The model was saved using Pickle.

**Concluding Remarks:**

As the telecommunication company generates more data on its customers (on new and churned out customers) the algorithm can be re-trained using the additional data and theoretically generate more accurate predictions to identify churning rate and reason behind it.