**How to Implement Customer Churn Prediction [Machine Learning]**

**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.

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

**What is the churn rate?**

Churn rate is a measure of the number of customers or employees who leave a company during a given period. It can also refer to the amount of revenue lost as a result of the departures.

Churn classify as:

1. Customer and revenue churn
2. Voluntary and involuntary churn

Customer and revenue churn : Customer churn is simply the rate at which customers cancel their subscriptions on the other hand revenue churn is the loss in you monthly recurring revenue at the beginning of the month.

Voluntary and Involuntary Churn: Voluntary churn is when the customer decides to cancel and takes the necessary steps to exit the service. Due to dissatisfaction, or not receiving the value they excepted on the other Involuntary Churn due to situations such as expired payment details, severe errors, insufficient funds.

**Data analysis**

We seem that in dataset there is no missing value found. Then we calculate value\_counts() for if any like (?) this type of value present in my dataset so I find it and then I replace to nan.

But we seem in my dataset (?) that type of value not present and we clearly see that target variable has imbalanced class distribution. Positive (Churn=Yes) is much less than negative (churn=No). Imbalanced class distributions influence the performance of a machine learnings model negatively.

The remaining categorical values have more than two values(or classes).

We seem that there is a high imbalance in SeniorCitizen and PhoneService variables. Most of the customers are not senior and similarly, most customers have a phone service.

We seem in count plot male and female are equally distributed. Average churn rate are approximately the same which indicates gender variable does not bring a valuable prediction power to a model.   
Therefore, I will not use gender variable in the machine learning model.

The other binary features have an effect on the target variable. The phone service may also be skipped if you think 2% difference can be ignored. I have decided to use this features in the model.

We can also use pandas pivot\_table to check the relationship between features and target variable.

Internet Services

6 variables that come with internet service which are streamingtv, StreamingMovies, OnineSecurity, OnlineBackup, DeviceProtection and Techsupport. We clearly see that fiber optic service is much moer expensive than DSL which may be one of the reason why customers churn.

Phone Service

If a customer does not have a phone service, he?she cannot have multiple lines. MulitpleLines column includes more specific data compares to PhoneService column. So I will not include PhoneService column as I can understand the number of people who have phone service from MultipleLines column.

More likely month-to-month service. Thjs clearly explains the motivation for copains to have long term relationships with their customers.

Payment method

We seem that countplot most of service used by people is Electronic check. Customers pay with electronic check are more likely to churn and this kind of payment is more common than other payment types.

Analysis continuous features

The continuous features is tenure, monthly charges and total charges.

It is clear that people who have been a customer for a long time tend to stay with the company. The average tenure in months for people who left the company is 20 months less than the average for people who stay.

It seems like monthly charges also have an effect on churn rate.

Contract and tenure features may be correlated because customer with long term contract are likely to stay longer with the company.

As expected, contract and tenure are highly correlated. Customers with long contracts have been a customer for longer time than customers with short-term contracts. I think contract will add little to no value to tenure feature so I will not use contract feature in the model.

**EDA Concluding Remarks**

Based on our basic exploratory analysis, we can define the important customer attributes that can give us the best insight in order to predict the type of customers that can churn. We can continue this analysis to answer some basic questions such as, ‘ Does lower estimated salary increase churn?’ or “Does lower credict score increase churn?” and so on. We can group and summarize the dataset in different ways to gain more insights from customer attributes. We will dive into more of these in the next dataset . For now, lets start thinking about predicting which customers will churn.

In this case, We can label out target(response ) variable which is churn. This means we can create a classification model and perform different algorithm methods such as Decision Tree, Random Fores, Logistic Regression or support Vector Machines. When it comes to machine learning models, we are looking for two main conditions: 1- Normal Distribution of the features set, 2- Same scale of the feature set.

**Pre-processing Pipeline**

Categorical features need to be converted to numbers so that they can be included in calculations done by a machine leaning model. The categorical variables in our data set are not ordinal (i.e. There is no order in them). For example, ‘DSL’ internet service is not superior to ‘Fiber optic’ internet service. An example for an ordinal categorical variable would be ratings from 1 to 5 or a variable with categories ‘bad’, ‘average’ and ‘good’.

When we encode the categorical variables, a number will be assigned to each category. The category with higher numbers will be considered more important or effect the model more. Therefore, we need to do encode the variables in a way that each category will be represented by a column and the value in that column will be 0 or 1.

Let’s do balancing target variable by sampling

First we separate positive class (churn = Yes) and negative class (churn= No). then upsampling the positive class. Combining positive and negative class and checking class distribution.

We seem that our target variable (churn) is balanced.

**Building Machine Learning Models**

First we separate feature and target column. Then we can further randomly split our dataset to training and testing dataset in order to fit our model with training dataset and test the predictions with the testing dataset. The idea is to train the model with the training dataset and test the prediction with the test dataset. If we didn’t use training and testing datasets and used the entire dataset instead, the algorithm will only make accurate predictions with our dataset and will fail with any new data the gets fed to it.

In this dataset, let’s use Logistic Regression, KNeighbors Classifier, Random Forest Classifier , Support Vector Machines and Gradient Boosting Classifier to create our model and prediction , further evaluate them which one is better.

Based on the metrics evaluations, while 76% of the predictions would be accurate with the Logistic Regression, 77% **o**f the predictions would be accurate with the KNeighbors Classifier, 90% **o**f the predictions would be accurate with the RandomForest Classifier, 77% **o**f the predictions would be accurate with the SVC, 79% **o**f the predictions would be accurate with the GradientBoosting Classifier. We would prefer to use Random Forest Classifier in this case.

When we look at the distribution of the customers that churn vs not churn, we see that the data is impartial. This means we can’t just rely on accuracy metric scores for the prediction models. Let’s look at the second customer data set to see if we can do better analysis and prediction models.

This time we are looking at a telecommunication company and it’s existing customer attributes such as their current plan, charges, location in terms of state, amount of customer service calls, account length and churn.

We can review additional evaluation metrics, such as cross validation matrix which will give us the amount of true positives, false positives, true and false negatives, precision, recall and f1 score. We can also see what we can do improve the model by looking at what features contributes the most to the prediction.

The model predicts 885 True Negatives, 159 False Positives, 29 False Negatives, 997 True Positive.

When we evaluate the model with Random Forest Classifier, we see that:

Precision score is 0.86

Recall score is 0.85

ROC curve is follows:

AUC score (the are under the roc curve) is 0.90 and f1 score is 0.90

Concluding Remarks

With the existing consumer insights through data, companies can predict customer’s possible needs and issues, define proper strategies and solutions against them, meet their expectations and retain their business. Based on the predictive analysis and modelling, businesses can focus their attention with targeted approach by segmenting and offering them customized solutions. Analyzing how and when the churn is happening in customer’s lifecycle with the services will allow the company to come up with more preemptive measures.

we have walked through a complete end-to-end machine learning project using the customer Churn dataset. We started by cleaning the data and analyzing it with visualization. Then, to be able to build a machine learning model, we transformed the categorical data into numeric variables (feature engineering). After transforming the data, we tried 5 different machine learning algorithms using default parameters. Finally, we tuned the hyperparameters of the Random forest Classifier (best performance model) for model optimization, obtaining an accuracy of nearly 90%.