# **DATATRAINED ACADEMY**

# **Blog Article –1**

Submitted by: Sindhu Shree N

Batch Number: 1832

## **Customer Churn Analysis:**

1. **Problem Statement:**

Customer churn is when a company’s customers stop doing business with that company. Businesses are very keen 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 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 prioritize 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.

**Understanding a problem:**

It’s important to understand what insights one needs to get from the analysis. In short, we must decide what question to ask and consequently what type of machine learning problem to solve: classification or regression.

1. **Data Analysis:**

The first stage of this analysis is to describe the dataset (where we can find mean, median, standard deviation, minimum and maximum values set in a table), understanding the meaning of each variable, detecting possible patterns and performing the necessary adjustments to ensure that the data will be proceeded correctly during the machine learning process. Each prompt them to leave companies could do more to prevent the loss of customers.

This project is based on customer data from IBM sample datasets with the aim of building and comparing several customer churn prediction models. It has 7043 data points and 20 features plus one target feature describing customers gender, dependents and whether they are senior citizens; and labelled(supervised learning) with whether they did churn or not. Machine learning models can help to understand and determine how these factors relate to workforce attrition (churn).

**Data Preparation and Cleaning:**

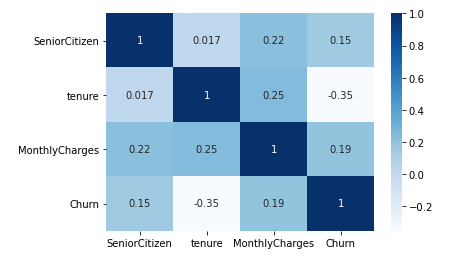
* Reading the CSV file and doing initial statistical analysis (like shape, info etc)
* Data pre-processing: Reading the unique values foe each column and removing those which won’t be significant in the analysis further.
* Create a new data frame to proceed with the analysis further.

Historical data that was selected for solving the problem must be transformed into a format suitable for machine learning. Since model performance and therefore the quality of received insights depend on the quality of data, the primarily aim is to make sure all the data points are presented using the same logic, and the overall dataset is free of inconsistencies.

Dataset Contains:

1. customerID
2. gender
3. SeniorCitizen
4. Partner
5. Dependents
6. tenure
7. PhoneService
8. MultipleLines
9. InternetService
10. OnlineSecurity
11. OnlineBackup
12. DeviceProtection
13. TechSupport
14. StreamingTV
15. StreamingMovies
16. Contract
17. PaperlessBilling
18. PaymentMethod
19. MonthlyCharges
20. TotalCharges
21. Churn

Next we will proceed to find the correlation of the features with the target features. In which we can find which are positively correlated and negatively correlated with the dependent feature (Churn)

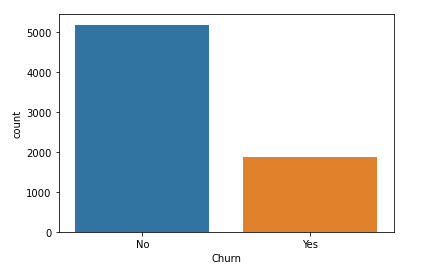


However the correlation matrix does not indicate any high degree of correlation with the dependent variable but it does provide us with the holistic view off some of the factors.

1. **EDA concluding Remark:**

* Finding patterns of data through visualization and reveal the hidden trends from data.
* Using both matplotlib and seaborn library to visualize the data.
* Finding relationships between features using bar graphs, histograms, box plots, heatmap.
* Analyzing both the numerical and the categorical columns separately.

Here churn is the target variable. The dataset is clean with no missing values. The target class is imbalanced, with churn rate of 35%.



So the churn rate is high where the employees are paid monthly charges of $10 to $2000, and the churn attrition seems to happen at every level regardless of employee monthly charges. This can be confirmed later at feature importance.

1. **Pre-Processing Pipeline:**

For the model to proceed with the data efficiently, the categorical variables should be encoded.

Data has to be pre-processed as a machine learning models are better at reading numbers than words. Using get\_dummies, the categorical data can be replaced with numbers with increasing the features (no. of input features now is increased to 25)

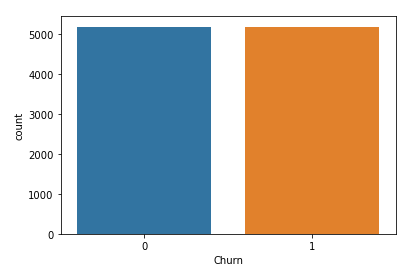
**Feature engineering, extraction and selection:**

Feature engineering is a very important part of dataset preparation. During the process, we will scale the data to standardize the independent features present in the data in a affixed range. If this is not done then the machine learning algorithms tends to weigh greater values, higher and consider smaller values as the lower values, regardless of the unit of the values. Since in our present data the dependent class is imbalanced so we have used the “MinMaxScaler” to scale the data.

**Resampling:**

Resampling is the technique we can increase or decrease the data points synthetically if the target class is imbalanced. It is used in most of the classification problems (binary classification). And most of the times we use upsampling in order to increase the data points which has less compare to other.

For the present data since our target class is imbalanced we will apply upsampling (for positive class), where it synthetically generates data points (corresponding to minority class) and are injected into the dataset. After this process, the counts of both labels are almost same. This equalization procedure prevents model from inclining towards the majority class. Furthermore, the interaction between the target classes remains unaltered.



So we can see that the positive class is upsampled now the target class looks balanced.

Since we need to predict the Churn which is a discrete value we will use the classification algorithms to predict those discrete values.

1. **Building machine learning models:**

The main goal of this project is to develop a churn prediction model. Where we usually train numerous models, tune, evaluate, and test them to define the one that detects potential churners with the desired level of accuracy on training data.

Before applying the machine learning algorithms we need to split the data in the way that 80% of data should go the training phase and rest of the data should go to the testing phase. This is done by using “train\_test\_split” function in sklearn model selection.

Later we use cross validation in training the models, and each baseline model performance can be tabulated.

The model will be cross-validated using a 5-fold cross validation returning the average accuracy. This method will be applied at every modelling step, to ensure that the model is not biased by the train\_test\_split.

Classic machine learning models are commonly used for predicting customer attrition, for example, ridge, decision tree, random forest and others. Using Random forest as a baseline model, then the performance of such models as XGBoost can be assessed. We generally use a baseline model’s performance as a metric to compare the prediction accuracy of complex algorithms.

There are 5 classification methods used in this particular analysis as discussed above:

**1. Ridge Classifier:** It is a type of linear classifier which is used to predict the churn. The baseline model performance results are quite good, with F1-Scores ranging from 70% to 80% for most of the models. After tuning hyper parameters and the threshold, the Ridge Classifier has achieved F1-Score of 76% and Recall also 76%.

**2. Decision Tree Classifier:** It is a type of supervised learning algorithm. This is almost used in classification problems. This algorithm splits a data sample into two or more homogeneous sets based on the input variables. A part of a tree is generated with each split. As a result, a tree with decision nodes and leaf nodes is developed. A tree starts from a root node which is a best predictor. Here in this case this model provides a good results, with F1-Scores ranging from 70% to 80% for most of the models. After tuning hyper parameters and the threshold, the Decision Tree Classifier has achieved F1-Score of 89% and Recall 88%.

**3. KNeighbors Classifier:** It is one of the simplest machine learning algorithms based on supervised learning technique. K is the number of nearest neighbors. The number of neighbors is the core deciding factor. This works by finding the distances between a target point and all input data points. Euclidean distance is taking as a measure of distance. Here this model achieved a good accuracy with F1-Scores ranging from 70% to 80% for most of the models. After tuning hyper parameters and the threshold, the KNeighbors Classifier has achieved F1-Score of 78% and Recall 76%.

**4. Random Forest Classifier:** It is a type of ensemble learning method that uses numerous decision trees to achieve higher prediction accuracy and model stability. Every tree classifies a data instance based on attributes, and the forest chooses the classification that received most instances. Here it achieved F1-Scores ranging from 70% to 80% for most of the models. After tuning hyper parameters and the threshold, the Random Forest Classifier has achieved F1-Score of 91% and also Recall 91%.

**5. XGBoost Classifier:** It is the implementation of the gradient boosted tree algorithms that is commonly used for classification and regression problems (In this case it is classification). This algorithm consisting of a group of weaker models or trees, which sums up their estimates to predict a target variable with more accuracy, with F1-Scores ranging from 70% to 80% for most of the models. After tuning hyper parameters and the threshold, the XGBoost Classifier has achieved F1-Score of 85% and Recall 86 %.

Also, if features are closely related to one another. One of them has to be removed to prevent misleading results to linear models like RidgeClassifier. Although tree-based models are not directly affected, they could also lead to over-fitting.

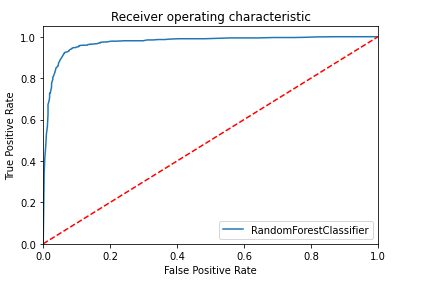
The highest final model of the dataset: In the end, we can see that utilizing data science on customers churn provided significant benefit to the business as we can tag each customer with the churn attrition score and come up with customized churn strategy for each group.

**AUC ROC Curve:**

Finally we will draw a Roc curve since it is a useful tool for the reasons like:

The curves of different models can be compared directly in general or for different thresholds.

The area under the curve (AUC) can be used as a summary of the model skill.



According to the classification report the accuracy of the model is 91% however its recall is lower 30% of positive cases. The RandomForestClassifier is proving excellent results. However, the purpose of the problem is to identify customers who are left. This is the reason that recall then becomes a very important measure. Recall measures the fraction of values that are identified correctly.

RandomForestClassifier has emerged as the final winning model with F1-Score of 91% and the highest Recall is also of 91%. Hope this could be the highest possible score achieved with the inherent limitations in the dataset.

The top factor for customer churn in this hypothetical organization seems to be tenure, emerged at the top. The customers with short-term contract are more likely to churn. This clearly explains the motivation for companies to have long term relationship with their customers.

Machine learning models are as good as the data to feed it, and more data would strengthen the model. In the real-life situation, getting the right data is often more challenging than the analytics itself.

1. **Concluding Remarks:**
2. With the help of notebook I learnt how **EDA** can be carried out using **Pandas and other plotting libraries**.
3. Also, I have seen making use of packages like **matplotlib and**

**seaborn** to develop better insights about the data.

1. I have also seen how **preprocessing** helps in dealing with **missing values and irregularities** present in the data. I also learnt **how to create new features** which will in turn help us to better predict the survival.
2. I also make use of **pandas profiling** feature to generate an html report containing all the information of the various features present in the dataset.
3. **Churn rate** is a health indicator for subscription based companies. The companies should know the customer’s requirements and try to provide them.
4. The decisive improvements can be made to company processes. While some level of churn rate is inevitable, it should be kept at the minimal possible level.
5. This model will allow the company to calculate the probability of an customer to unsubscribe the company and to act on key-factors to avoid departures. The **tenure** of customers seem to be important causes of withdrawals.