**Introduction:**

Machine Learning (ML) by looking at the definition, we can say that it is a field of computer science. The telecommunications sector has become one of the main industries in developed countries.

**Problem Definition:**

Customer churn is when a company’s customers stop doing business with that company or service. Customer churn is the percentage of customers who stopped using your company’s product or service during a certain period of time. The retention and the acquisition of the users are the major concerns in the telecom industry. Machine learning algorithm techniques make us sure that there are a certain amount of approaches needed to lessen the churn rate. We will use logistic regression to build the model. Here, we need to predict the churn which has two values i.e. Yes and No. There are 7043 rows and 21 attributes in this dataset. The data contains categorical, integer and float type values. There are no missing values present in this dataset.

**Exploratory Data Analysis (EDA):**

EDA is a phenomenon under which data analysis is used for gaining a better understanding of data aspects like a.) Main features of data b.) Variable and relationships that hold between them c.) Identifying which variables are important for our problem.

There are 7043 samples in this dataset. There are float, integer and categorical values present in this dataset. The target variable churn is imbalanced. We will be balancing our target variable by using random oversampling which will add copies to the minority class and therefore, our class will be balanced. We will plot the box plots, distribution plots, count plots etc. Here, we will be plotting the count plots for all the categorical variables in this dataset. While plotting the distribution plots for the columns in this dataset, if there is any skewness present in any of the columns, we will be removing it by applying some transforms like sqrt, cbrt and log transforms. We also need to make sure that the skewness range is between -0.5 to +0.5 and we plot the distribution plots for each columns without any skewness. We plot the box plots to detect if there are any outliers present in them or not. We will also plot the histogram plots for all the column to check whether there is any skewness present in each of the features or not. Multiple libraries are available to perform basic EDA such as pandas, matplot library and seaborn.

**EDA Concluding Remarks:**

After doing all the EDA process, I was successful in removing all the skewness from the data and I removed all the outliers which were present in the data by using Z-score.

**Pre-Processing Pipeline:**

The pre-processing of data involves three steps namely data cleaning, feature selection and data transformation. The data cleaning involves missing data. We need to replace the missing values by using either mean, median or mode. Encoding the categorical variables which were done by using a label encoder as the categories are assigned starting from zero. Before, training the model feature selection is one of the important factors that can influence the model’s performance.

**Building Machine Learning Models:**

I have used Logistic Regression since it is one of the best model of regression analysis to utilize when our target variable is binary. It describes the relationship between an independent variable and a dependent variable. I have also used some of the classification algorithms to check the accuracy of our data and check which model is performing better when compared to the other models. Let’s see the introduction of all the models which I have used in my dataset.

1. Logistic Regression: It works when there are only two values i.e. 0 and 1. It is estimating the parameters of a logistic model which is a form of binary regression.

2. Gaussian Naïve Bayes: It is a supervised machine learning algorithm. It is a classification technique that works only when there are binary values.

3. K Nearest Neighbor: This algorithm is working by calculating the distance of the nodes and we call this distance Euclidian distance. Here, k= number of neighbors.

4. Support Vector Machine (SVM): In SVM, we have a parameter called a kernel. We can give different values to the kernel to improve the accuracy. Radius base function (RBF), linear and poly are kernels in svc.

5. Decision Tree: It is one of the classification techniques. It works better with classification problem. We have two parameters in the decision tree i.e. Gini and entropy.

6. Random Forest: It is a supervised machine learning technique. It is an ensemble learning method for classification. It also works very well with classification.

Firstly, I have imported all the algorithms which were required to build the model. I have made sure that my training data is one side and the testing data is on another side. I have split the data into training and testing datasets. I have made 68% of the data available for training and 32% of the data available for testing. Then, I have taken the logistic regression model to check on which random state I am getting the best accuracy score such that the random state can be finalized. Then, by using the fit method, I have trained all the models. By using predict method, I have predicted the values for all the models. Predicted data is nothing but the answer given by the x\_test model and y\_test is the actual data. We will use an accuracy score from sklearn to find the accuracy of the model. We will also use a confusion matrix and classification report from sklearn and because of the confusion matrix, we can tell what will be the accuracy of the model. By using a classification report, we will be getting scores like recall, precision, specificity and model performance.

The formulae are as follows:

TP-True Positive, TN-True Negative, FP-False Positive, FN-False Negative

Recall or Sensitivity=TP/TP+FN

Precision=TP/TP+FP

Specificity=TN/FP+TN

Model Performance=F1=2(Precision\*Recall/Precision + Recall)

The accuracy score for logistic regression is 75.84%, accuracy score for gaussian naive bayes is 75.44%, accuracy score for support vector machine is 74.07%, decision tree classifier is 85.05%, accuracy for K Neighbors classifier is 79.14% and accuracy score for random forest classifier is 87.47%. After checking the accuracy scores for all the models, the Random Forest model is giving me the highest accuracy score of 87.47%. But this accuracy score can be because of overfitting. Let’s check with cross-validation.

Cross-Validation: It is a resampling technique with a basic idea of dividing the dataset into training and testing. On one part we train the model and on the second part i.e. testing the data which is unseen for the model. If the model works with good accuracy on test data, it means that the model has not over fitted the training data and it can be trusted with the prediction whereas if it performs with bad accuracy then our model is not to be trusted and we need to tweak our algorithm. We will use cross val score for cross-validation. I have taken the mean of all the scores obtained in each iteration as the final score of our model. The cross val score for logistic regression model is 76.43%, cross val score for gaussian naïve bayes is 75.51%, cross val score for support vector machine is 74.54%, cross val score for decision tree classifier is 87.25%, cross val score for K Neighbors classifier is 81.02% and cross val score for random forest classifier is 88.12%. After checking the cross val score for all the models, the least difference between the highest accuracy score and cross val score says that the random forest model is working better when compared to the other models. Now, let’s fine-tune the model by using hyper-parameter tuning.

Hyper Parameter Tuning: We use this technique to improve the accuracy score of the best model that we have chosen. Sklearn comes with Grid Search CV to do the search over specified parameter values for an estimator. It helps to optimize the model’s performance. With the best parameters, the model will identify the patterns within the dataset in a better way. I used the parameters like criterion, max depth, min samples split, max features and n estimators for the random forest classifier model to find the improved accuracy of the model. After tuning the model, the accuracy of the model with the best parameters is around 89.69% which suggests that there is an increase in the performance of the model by 2.5%.

Saving the model: We are done with testing the model. Now, we will save our model by importing joblib. Now, we need to load the model from the joblib.

**Conclusion:**

Customer churn is often a critical problem for the telecom sector as the customers do not hesitate to leave if they are not able to get the product or the service that they are looking for. It can be concluded that with machine learning, we were able to predict effective ways for recognizing churn in customers. In this way, we have learned to build a machine learning project. We also learned to fine-tune our model and save it for further use.