Telecom Churn Prediction Project Documentation

Aim :-

To predict the customers who are likely to churn in the next N months & facilitate in taking business actions for reducing the churn.

Objective: -

Churn Prediction in Telecom using machine learning.

Estimating churners before they dis continue using a product or service is extremely important. In this Machine Learning project, you will develop a churn prediction model in telecom to predict customers who are most likely subject to churn.

In any service providing industry, when a customer decides to stop using the service either by cancelling the subscription or not paying for the service, we call this customer churn. Churn is defined as how many customers are not using the service for a certain period. Hence, customer churn is one of the essential metrics that every business must evaluate to grow. The churn rate is calculated by dividing the number of lost customers by the last number of customers. Thus, a company churn rate must be as low as possible, ideally 0%. But why is it so important to calculate the churn rate? Does it affect the business if you lose around 5% of customers? Yes, the answer is that it costs more to acquire a new customer than retain the existing customers. Retaining the current customers, any company can spend less on operating costs needed to reach new customers. So, we will use advanced machine learning techniques to predict the potential churners who are about to leave a company’s service and take the necessary steps to prevent it. This project aims to build a deep learning model that will help predict customers who are likely to churn in the next N months and facilitate in taking business actions for reducing the churn.

Project Template Outcomes

1. Understanding the Business problem:
2. Understanding what churn is, why to solve it and how to solve it.
3. Importing the dataset and required libraries.
4. Why have to do EDA (Exploratory Data Analysis).
5. Data cleaning and missing data handling if required, using appropriate methods.
6. Checking for class distributions.
7. Outlier Treatment.
8. Splitting dataset as per requirement.
9. Sampling the dataset.
10. Perform feature Engineering.
11. Building model using all ML algorithms.
12. Understanding different performance metrics like Confusion matrix, AUC, Recall, F1-score.
13. Building feature importance.
14. Analyzing the correlation plot and gaining insights.
15. Model refinement and implementation.

Understanding the Business problem: -

By using the business objective we are prediction the Telecom Churn prediction project process. when a customer decides to stop using the service either by cancelling the subscription or not paying for the service, we call this customer churn. Churn is defined as how many customers are not using the service for a certain period. Hence, customer churn is one of the essential metrics that every business must evaluate to grow. The churn rate is calculated by dividing the number of lost customers by the last number of customers. Thus, a company churn rate must be as low as possible, ideally 0%. But why is it so important to calculate the churn rate? Does it affect the business if you lose around 5% of customers? Yes, the answer is that it costs more to acquire a new customer than retain the existing customers. Retaining the current customers, any company can spend less on operating costs needed to reach new customers. So, we will use advanced machine learning techniques to predict the potential churners who are about to leave a company’s service and take the necessary steps to prevent it.

Understanding what churn is, why to solve it and how to solve it:-

By understanding this concept there is plenty of algorithm to solve like a Linear Regression, Logistic regression, Decision Tree, Random Forest method, K-NN concept, K-Mean Clusters, NLP, so many methods are available.

Each one explore it own way so that is like own protocols.

There is a python code with data science, the python exploration is efficient. Churn is nothing but user to drop a member from the

particular service provider. And also the vendors retain the customers at the same times. The vendor giving the service and there after continue with the real time customers. Suppose other vendors having previlages the drop using the customer is unavoidable.

particular service provider. And also the vendors retain the customers at the same times. The vendor giving the service and there after continue with the real time customers. Suppose other vendors having previlages the drop using the customer is unavoidable.

Importing the dataset and required libraries: -

1. According to the program aspect we are always using the Technology start with importing all the built-in libraries functions like as numpy, pandas, seaborn, matplot lib as pyplot, and also from sklearn that is called science kit tool function and so on.

So the sample library functions are:

import numpy as np

import pandas as pd

import seaborn as sns

import matplotlib.pyplot as plt

from IPython.display import display

from sklearn.preprocessing import PowerTransformer

from sklearn.model\_selection import train\_test\_split

from sklearn.preprocessing import StandardScaler

from sklearn.metrics import roc\_auc\_score

from sklearn.metrics import roc\_curve

from sklearn.feature\_selection import RFE

After importing libraries position the progammers who get the excel exact file name and pd.read\_csv to import the files.

Whether varname.head( ) to display the content which approximately 5 \* 21 rows and columns will displayed is it having missing values or not.

By similar way varname.tail( ) way it is decending order.

Also whether to find a null values through data information using.

If we have no null value here then the data information will easiest.

Suppose we are having null values we have to clear the data reducing using data cleaning.

Exploratory Data Analysis:

**Exploratory Data Analysis (EDA)**is an approach to analyze the data using visual techniques. It is used to discover trends, patterns, or to check assumptions with the help of statistical summary and graphical representations.

**Dataset Used:**

For the simplicity of the article, we will use a single dataset. We will use the employee data for this. It contains 8 columns namely – First Name, Gender, Start Date, Last Login, Salary, Bonus%, Senior Management, and Team.

**Dataset Used:**[Employees.csv](https://media.geeksforgeeks.org/wp-content/uploads/employees.csv)

Let’s read the dataset using the Pandas module and print the 1st five rows. To print the first five rows we will use the [**head()**](https://www.geeksforgeeks.org/python-pandas-dataframe-series-head-method/) function.

**Steps Involved in Exploratory Data Analysis:**

1. Data Collection. Data collection is an essential part of exploratory data analysis. ...
2. Data Cleaning. Data cleaning refers to the process of removing unwanted variables and values from your dataset and getting rid of any irregularities in it. ...
3. Univariate Analysis. ...
4. Bivariate Analysis.

Data cleaning and Missing data Methods:

Data cleaning is the process of fixing or removing incorrect, corrupted, incorrectly formatted, duplicate, or incomplete data within a dataset. When combining multiple data sources, there are many opportunities for data to be duplicated or mislabeled. If data is incorrect, outcomes and algorithms are unreliable, even though they may look correct. There is no one absolute way to prescribe the exact steps in the data cleaning process because the processes will vary from dataset to dataset. But it is crucial to establish a template for your data cleaning process so you know you are doing it the right way every time.

Having clean data will ultimately increase overall productivity and allow for the highest quality information in your decision-making. Benefits include:

* Removal of errors when multiple sources of data are at play.
* Fewer errors make for happier clients and less-frustrated employees.
* Ability to map the different functions and what your data is intended to do.
* Monitoring errors and better reporting to see where errors are coming from, making it easier to fix incorrect or corrupt data for future applications.
* Using tools for data cleaning will make for more efficient business practices and quicker decision-making.

Checking for class distributions:

In a normal distribution, data is symmetrically distributed with no [skew](https://www.scribbr.com/statistics/skewness/). When plotted on a graph, the data follows a bell shape, with most values clustering around a [central region](https://www.scribbr.com/statistics/central-tendency/) and tapering off as they go further away from the center.

Normal distributions are also called Gaussian distributions or bell curves because of their shape.

All kinds of variables in natural and social sciences are normally or approximately normally distributed. Height, birth weight, reading ability, job satisfaction, or SAT scores are just a few examples of such variables.

Because normally distributed variables are so common, many[statistical tests](https://www.scribbr.com/statistics/statistical-tests/) are designed for normally distributed populations.

Understanding the properties of normal distributions means you can use [inferential statistics](https://www.scribbr.com/statistics/inferential-statistics/) to compare different groups and make estimates about populations using samples.

Outlier Treatment:

**Outlier** **Treatment** is One of the important part of **data** pre -processing is the handling outlier. If your data contains outliers that affect our result which will depend on the data. So to remove these **outliers** from **data** **Outlier** **Treatment** is used.

Splitting dataset as per requirement:

1. Training set (Has to be the largest set)
2. Cross-Validation set or Development set or Dev set
3. Testing Set

The test set can be sometimes omitted too. It is meant to get an unbiased estimate of algorithms performance in the real world. People who divide their dataset into just two parts usually call their Dev set the Test set.

We try to build a model upon training set then try to optimize hyperparameters on the device set as much as possible then after our model is ready, we try and evaluate the testing set.

## **Training Set**:

The sample of data used to fit the model, that is the actual subset of the dataset that we use to train the model. The model observes and learns from this data and optimize its parameters.

## **Cross-Validation Set:**

We select the appropriate model or the degree of the polynomial (if using regression model only) by minimizing the error on the cross-validation set.

## **Test set:**

The sample of data used to provide an unbiased evaluation of a final model fit on the training dataset. It is only used once the model is completely trained using the training and validation sets. Therefore test set is the one used to replicate the type of situation that will be encountered once the model is deployed for real-time use.

Sampling the Dataset:

After this, we create a Python function called random\_sampling() that takes population data and desired sample size and produces as output a random sample. Systematic Sampling Systematic sampling is defined as a probability sampling approach where the elements from a target population are selected from a random starting point and after a fixed sampling interval.

Building model using all ML algorithms:

**Machine** **learning** happens to be a small part of this process. The **model** **building** process involves setting up ways of collecting data, understanding and paying attention to what is important in the data to answer the questions you are asking, finding a statistical, mathematical or a simulation model to gain understanding and make predictions.

Building model using all ML algorithms:

**Supervised Learning:** In [Supervised Learning](https://www.educba.com/what-is-supervised-learning/), the data set is labeled, i.e., for every feature or independent variable, there is a corresponding target data which we would use to train the model.

**Un-Supervised Learning:** Unlike in Supervised Learning, the data set is not labeled in this case. Thus clustering technique is used to group the data based on its similarity among the data points in the same group.

**Reinforcement Learning:** A special type of Machine Learning where the model learns from each action taken. The model is rewarded for any correct decision made and penalized for any wrong decision, which allows it to learn the patterns and make better accurate decisions on unknown data.

**Regression:** There is a continuous relationship between the dependent and the independent variables. The target variable is numeric in nature, while the independent variables could be numeric or categorical.

**Classification:** The most common problem statement you would find in the real world is classifying a data point into some binary, multinomial, or ordinal class. The target variable has only two outcomes (Yes/No, 0/1, True/False). In the Multinomial Classification problem, there are multiple classes in the target variable (Apple/ Orange/Mango, and so on). In the Ordinal classification problem, the target variable is ordered (e.g., the grade of students).

To solve this kind of problem, programmers and scientists have developed some programs or algorithms that could be used on the data to make predictions. These algorithms could be divided into linear and non-linear or tree-based algorithms. Linear algorithms like Linear Regression, Logistic Regression are generally used when there is a linear relationship between the feature and the target variable, whereas the data exhibits non-linear patterns, the tree-based methods such as [Decision Tree](https://www.educba.com/decision-tree-in-machine-learning/), Random Forest, [Gradient Boosting](https://www.educba.com/gradient-boosting-algorithm/), etc., are preferred.

**Linear Regression:**

As the name suggests, this algorithm could be used in cases where the target variable, which is continuous in nature, is linearly dependent on the dependent variables.

It is represented by:

**y = a\*x + b + e**, where y is the target variable we are trying to predict, a is the intercept, and b is the slope, x is our dependent variable used to make the prediction. This is a Simple Linear Regression as there is only one independent variable.

In the case of [Multiple Linear Regression](https://www.educba.com/multiple-linear-regression/), the equation would have been:

y = a1\*x1 + a2\*x2 + …… + a(n)\*x(n) + b + e

Here, e is the error term, and a1, a2.. a (n) are the coefficient of the independent variables.

A metric is used to evaluate the model’s performance, which could be Root Mean Square Error, which is the square root of the mean of the sum of the difference between the actual and the predicted values The goal of Linear Regression is to find the best fit line which would minimize the difference between the actual and the predicted data points.Logistic Regression

In terms of maintaining a linear relationship, it is the same as Linear Regression. However, unlike in Linear Regression, the target variable in Logistic Regression is categorical, i.e., binary, multinomial or ordinal in nature. Moreover, the choice of the activation function is important in Logistic Regression as for binary classification problems, the log of odds in favor, i.e., the sigmoid function, is used.

Machine Learning Algorithms could be used for both classification and regression problems. The idea behind the KNN method is that it predicts the value of a new data point based on its K Nearest Neighbors. K is generally preferred as an odd number to avoid any conflict. While classifying any new data point, the class with the highest mode within the Neighbors is taken into consideration. While for the regression problem, the mean is considered as the valueIn the case of a multi-class problem, the softmax function is preferred as a sigmoid function takes a lot of computation time. The metric used to evaluate a classification problem is generally Accuracy or the ROC curve. The more the area under the ROC, the better is the model. For example, a random graph would have an AUC of 0.5. The value of 1 indicates the most accuracy, whereas 0 indicates the least accuracy.

Understanding different performance metrics like Confusion matrix, AUC, Recall, F1-score:

Confusion\_Matrix:

A confusion matrix is a summarized table of the number of correct and incorrect predictions (or actual and predicted values) yielded by a classifier (or classification model) for binary classification tasks. In simple words, “ A confusion matrix is a performance measurement for machine learning algorithm ”.

AUC:

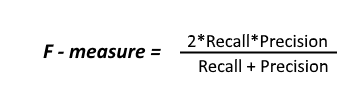
AUC stands for Area Under the Curve. ROC can be quantified using AUC. Th e way it is done is to see how much area has been covered by the ROC curve. If we obtain a perfect classifier, then the AUC score is 1.0. If the classifier is random in its guesses, then the AUC score is 0.5.

Recall:

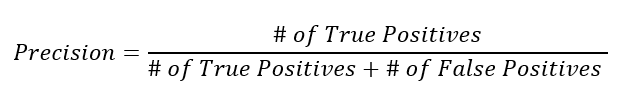
how many right hits were also **identified**, is referred to as recall. Precision (your formula is wrong) is the percentage of returned hits that were true positives, or accurate hits.

F1-Score:-

**F1-Score** or **F-measure** is an evaluation metric for a classification defined as the harmonic mean of **precision** and **recall**. It is a statistical measure of the accuracy of a test or model. Mathematically, it is expressed as follows,

  
Here, the value of F-measure(F1-score) reaches the best value at 1 end. the worst value at 0. F1-score 1 represents the perfect accuracy and recall of the model.

Precision is the first part of the F1 Score. It can also be used as an individual machine learning metric. It’s formula is shown here:



You can interpret this formula as follows. Within everything that has been predicted as a positive, precision counts the percentage that is correct:

* A **not precise model**may find a lot of the positives, but its selection method is noisy: it also wrongly detects many positives that aren’t actually positives.
* A **precise model** is very “pure”: maybe it does not find all the positives, but the ones that the model does class as positive are very likely to be correct.

Analysing the correlation plot and gaining insights:

### **Carl Pearson Correlation Coefficient:**

One of the most used ones is the Pearson Correlation Coefficient. We can call it just the correlation coefficient. This coefficient is used to calculate the correlation with the terms:

1. The data have interval or ratio scale.

2. The relationship between the two variables must be linear, it means that the distribution of data generally scatters along a straight line.

3. Data is normally distributed.

Model refinement and implementation:

The model refinement task in system-level synthesis**transforms a specification from a functional model to a chosen implementation model**. In this paper, we categorize several commonly-used implementation models and then describe a set of refinement procedures to transform a specification to each of these implementation models.