Computer Science Department
CS675 – Introduction to Data Science (CRN: 72458)
Fall 2022

Project #2 / Due 30-Nov-2022(*)

This is the continuation of Project #1.
You have performed the very important step of Explanatory Data Analysis (EDA) on Customer Churn data within the Telecommunication industry.

It is very common for companies to focus energy on combating customer churn, but the reality is that once they've identified a customer who is about to leave, it's often too late to entice them to stay.

The key is to identify customers at risk of churn early, before they get too far down the path, and to take preemptive action to retain them and improve the customer relationship.
This is where Advanced Analytics (Machine Learning and Deep Learning modeling) comes into play.
Using Data Science and AI approaches to predict which of the customers will be loyal and which will lapse.
Your ML model(s) will help the company's executives to understand what makes a great or bad customer so they can take action.

Teaching machines to predict customer behavior, and communicate which customer attributes predict specific behaviors, allows management to help build an organizational playbook for acquiring and keeping happy customers. For example, a client success organization can reach out before there's a problem, their marketing department can reach new customers that are less likely to churn, and their sales organization can bring these better customers on board.

This problem is a typical classification task. You must **build Machine Learning models to predict whether a customer will churn or not**.

You are asked to **perform two (2) stages of analysis**, based on different distribution of data:
1- Fit your models on the original (given) dataset
2- Fit your models on the modified dataset, after applying the **SMOTE** technique.

The Machine Learning models to be used are:

- Naïve Bayes
- Logistic Regression
- Random Forests
- XGBoost

The metrics of performance of the chosen models should be, with huge emphasis to **recall**:

- Accuracy
- Precision
- **Recall**
- F1-Score

You should follow the standard ML workflow process while building your models:

- Explanatory Data Analysis (already done!)
- Data Visualization (already done!)
- Data Preprocessing (Data Imputation, Feature Selection & Scaling, Encode Categorical Features) (partially done)

- Address Data Imbalance (apply the SMOTE technique)
- Split the training/test datasets in the 80/20 % ratio
- Algorithm Selection
- Modeling Building
- Modeling Evaluation
- Model Tuning (Hyperparameter Tuning)

Churn rate is a critical metric of customer satisfaction. Low churn rates mean happy customers; high churn rates mean customers are leaving you. A small rate of monthly/quarterly churn compounds over time. 1% monthly churn quickly translates to almost 12% yearly churn.

The data (**telco-customer-churn.csv**) is available for you to download.

The dataset has 7043 rows and 21 columns.

There are 17 categorical features:
CustomerID: Customer ID unique for each customer
gender: Whether the customer is a male or a female
SeniorCitizen: Whether the customer is a senior citizen or not (1, 0)
Partner: Whether the customer has a partner or not (Yes, No)
Dependent: Whether the customer has dependents or not (Yes, No)
PhoneService: Whether the customer has a phone service or not (Yes, No)
MultipeLines: Whether the customer has multiple lines or not (Yes, No, No phone service)
InternetService: Customer's internet service provider (DSL, Fiber optic, No)
OnlineSecurity: Whether the customer has online security or not (Yes, No, No internet service)
OnlineBackup: Whether the customer has an online backup or not (Yes, No, No internet service)
DeviceProtection: Whether the customer has device protection or not (Yes, No, No internet service)
TechSupport: Whether the customer has tech support or not (Yes, No, No internet service)
StreamingTV: Whether the customer has streaming TV or not (Yes, No, No internet service)
StreamingMovies: Whether the customer has streaming movies or not (Yes, No, No internet service)
Contract: The contract term of the customer (Month-to-month, One year, Two years)
PaperlessBilling: The contract term of the customer (Month-to-month, One year, Two years)
PaymentMethod: The customer's payment method (Electronic check, Mailed check, Bank transfer (automatic), Credit card (automatic))

Next, there are 3 numerical features:
Tenure: Number of months the customer has stayed with the company
MonthlyCharges: The amount charged to the customer monthly
TotalCharges: The total amount charged to the customer

Finally, there's a prediction feature:
Churn: Whether the customer churned or not (Yes or No)

Write **Python** scripts in order to complete the following tasks along with their output. All work should be done and submitted in a single **Jupyter Notebook.**
For each of the two (2) stages of analysis, perform the following:
1- Build each of the four (4) models independently, by taking the default parameters.
2- Present the accuracy of each model; list the best model (in terms of 'recall' metric)
3- Tune (Hyperparameter tuning) only the Random Forest and XGBoost Models
4- Present the accuracy of each model; list the best model (in terms of 'recall' metric)
5- Which of the two stages of analysis produces better results? Which is the best model, overall?