**Project: Telco Customer Churn**

DSC630\_Predective\_Analytics

Milestone2

Name: Bhargava Gaggainpali

**Introduction**

**Background:**

Customer churn is a major problem and one of the most important concerns for every company. Retaining an existing customer is many times more effective thanks gaining new customers. Generally, people only switch to a different company only when the service is not to the level of expectation in one or the other factor when compared to competitors, more than getting attracted to the specials and offers which are offered by the other companies. As the number of peoples who use a phone or a telecom product does not increase in huge volume, any company looking to improve the profits is trying to attract other company customers. At the same time retaining existing customers is very important. As retaining a current customer is ten times more productive than gaining a new one.

**Problem Statement:**

As customer churn directly effects the revenues of the companies, companies are seeking to develop means to predict potential customer to churn. Therefore, finding factors that increase customer churn is important to take necessary actions to reduce this churn and retain the customers.

**Scope:**

Scope of the project is to use the sample dataset from Kaggle, which shows several categories of customers and also some influencing factors causing the customer churn in the telecom industry to perform customer segmentation and identify a pattern which influence the customer churn. For this, I am planning to implement supervised machine learning process to achieve the goal.

**Document Overview:**

In the document, I am planning to detail all the steps which are followed to find the factors which are influencing the customer churn in the telecom industry. Which includes the source of the data, tools used to access data, processing mechanism to achieve the goal. Also include the validation process to make sure the process is not biased.

**Preliminary Requirement**

**Technical Approach:**

I am planning to follow the technical approach of performing Exploratory data analysis as the first step and disregard the non-relevant data and apply supervised machine learning algorithms to fit the model and validate the model by using confusion matrix.

**Data sources or plan for data:**

I am planning to use the sample data source from Kaggle. 7043 observations and 21 variables

[Telco Customer Churn | Kaggle](https://www.kaggle.com/blastchar/telco-customer-churn)

Given are list of the variables:

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

**Analysis:**

Will perform initial stats to check on the duplicate rows and null values to make sure the valid content of the data. Generate multiple charts on the variables to see the spread of the values in the variables. Perform Univariant analysis by using histograms and box plots. Perform Bivariant analysis by using correlation matrix and heat maps.

**Requirement Development:**

To Achieve the requirement, I will use python as my primary tool to build charts and stats to get better understanding of the data. As a next step will use multiple supervised machine learning algorithms to build couple of models to pick the most efficient model.

**Model Deployment:**

Planning to build multiple models using Support Vector Machines, Logistic Regression and Random Forest, as part of the build and based on the accuracy of the confusion matrix, planning to deploy the best model for implementation.

**Testing and Evaluation:**

While in the process of building the model, planning to split the data as 70% for building the model and 30% to evaluate the model. This gives us enough sample data to test in all the scenarios to make sure the model is accurate.

**Expected Results**

Expected result of this project is to use the sample data and build a model by identify the factors which are causing the customer churn with a model accuracy of more than 85%.

**Execution and Management of Project**

**Project Plan:**

I am planning to implement the project in a step-by-step process.

1. Understand the data by running stats.
2. Remove duplicates and handle null values.
3. Using charts analyze the spread of the data, univariant and bivariant relationship.
4. Perform dimensionality reduction and Convert categorical values to numerical values.
5. Run supervised machine learning algorithms to build model.
6. Evaluate the model by using test data. Based on the performance, try to implement multiple algorithms.

**Project Risk:**

Based on the limited data in the sample dataset, I am guessing it might impact the performance of the model. If needed planning to source additional supporting data from other sources.

**Reference:**

[Telco Customer Churn | Kaggle](https://www.kaggle.com/blastchar/telco-customer-churn)

[How Costly Is Customer Churn in the Telecom Industry? - The European Business Review](https://www.europeanbusinessreview.com/how-costly-is-customer-churn-in-the-telecom-industry/?__cf_chl_jschl_tk__=ec58e0210ef4250a8414742b7aaa8d8e28384753-1611086044-0-AXu6Mldw8Cpmi6L6s5SgV1PXJkIGPIbNtM2k3mk8dG4GwrfiMZ4fT3rVf0CPVbws4ipH6NuFN2fITmeBiQb8uSJYkXQKBcXB-e3XBH6IJWYli7zOM4eqnO9BCh0JE11v9IHHQZeKKMNRTUecYHq5Mahj6uNBBJ23n_LWFAdEF677ZQEQpPYruho0OLOLbdjBj-oKjPGHAZ52v2QMgPAHjyXRIDb-FEUu_sFhz8CW6q0KxFhOgoPWW_R8KZhMo4AOUlv2fbzwIi3oYhYXMSMOdAjpE9u7gRwD-0RQeFCEtTEzQqRmjC2QEbIGAHZKgbkn99TTuu5kBJLOuOoRLFhaD9d-oHKuQ27yx7WrU_mDPxYdp7qkByeE3o25LzRg460JyA)