PROPOSAL FOR DATA 428 PROJECT

Background

I have gained knowledge of many parts of Data Science through eight courses as part of Masters of Applied Data Science over last one year. I wish to exploit the facility of a project under DATA 428 to further strengthen my knowledge and confidence while working on a dataset related to my previous work domain i.e., telecommunications.

Problem Statement

Analyse 'Telecom Customer' Dataset for predicting likelihood of customer churn.

Data Source

The 'Telecom Customer' dataset contains data for 100,000 customers with 99 variables collected over five years. It has been downloaded from 'https://www.kaggle.com/abhinav89/telecom-customer'.

Scope

I intend covering following aspects while analysing the data:

(a) Data Exploration and Manual Feature Selection:

- Distribution of variables – apply suitable transformations if required.
- Correlation of quantitative variables.
- Handling the outliers.
- Handling missing values.

(b) Feature Selection. Explore different methods for selection of features and dimension reduction like Elastic Net, RFE and PCA.

(c) Building a Predictive Model.

- Explore different possibilities to build a suitable predictive model
- Judge each candidate model based on various criteria
    - **AuC / Accuracy / minimised "cost function"** (looking for best performer on unseen data)
    - **Prediction speed** (understand the average speed of the method at prediction)
    - **Transparency** (understand the degree of transparency that a method exhibits)
- Recommend appropriate model for the telecom company.

(d) Performance Analysis. Analyse the model performance through confusion matrix and ROC – AUC.

The above work will be undertaken in Python, exploiting pandas, numpy, scipy, seaborne, matplotlib and scikitlearn – packages which I wish to strengthen my hold on.