CSCI - 4146 - The Process of Data Science - Summer 2022

Assignment 2

The submission must be done through Brightspace. Due date and time as shown on Brightspace under Assignments.

- To prepare your assignment solution use the assignment template notebook available on Brightspace.
- The detailed requirements for your writing and code can be found in the evaluation rubric document on Brightspace.
- Questions will be marked individually with a letter grade. Their weights are shown in parentheses after the question.
- Assignments can be done by a pair of students, or individually. If the submission is by a pair of students, only one of the students should submit the assignment on Brightspace.
- We will use plagiarism tools to detect any type of cheating and copying (your code and PDF).
- Your submission is a single Jupyter notebook and a PDF (with the compiled results generated by your Jupyter notebook). File names should be:
  - **A2-<your_name1>-<your_name2>.ipynb**
  - **A2-<your_name1>-<your_name2>.pdf**
- **Forgetting to submit both files results in 0 markings for both students.**

About the dataset : Telco Customer Churn

Context:

Customer churn is the percentage of customers that stopped using the company's product or service during a given time period, and it's one of the most important metrics for businesses to evaluate customer loyalty. Customer churn is a critical metric because it is much less expensive to retain existing customers than it's to acquire new ones.

A model that can predict reliably whether a customer is going to leave or not, is a valuable asset to companies. Companies can improve their customer retention by knowing which customers they need to put their efforts to retain. Telecom companies tend to experience significant fluctuation in their customers, so such models can be very useful to this type of companies.

Some relevant information about the data set:

- Each row represents a customer, and each column contains the customer's attributes.
- The dataset contains customers who left within the last month—the column is called "Churn".
- The dataset contains services that each customer has signed up for—phone, multiple lines, internet, online security, online backup, device protection, tech support, and streaming TV and movies
- Customer account information—how long they've been a customer, contract, payment method, paperless billing, monthly charges, and total charges
- Demographic info about customers—gender, age range, and if they have partners and dependents

1. **Data understanding and feature engineering (0.1)**
    - **a.** Extract the numerical and categorical features from the dataset and build the data quality report.
    - **b.** Identify data quality issues and build the data quality plan.
    - **c.** Preprocess your data according to the data quality plan
    - **d.** Explore the data set to find patterns in the data (e.g. correlation, trends, etc) and potentially form some hypotheses.
        - **i.** Customer Account Information: Consider the tenure, and contract variables, and then plot graphs and describe your observations.
        - **ii.** Services that each customer has signed up for—phone, multiple lines, internet, online security, online backup, device protection, tech support, then visualize plots and describe your observations.

2. **Build a baseline model to predict customer churn (0.35).** Using the CSV file, split the dataset into training(40%), validation(30%), and test(30%) splits.
    - **a.** Explain what the task you're solving is (e.g., supervised x unsupervised, classification x regression x clustering or similarity matching x, etc).
    - **b.** Use a feature selection method to select the features to build a model.
    - **c.** Select the evaluation metric. Justify your choice.
    - **d.** Perform hyperparameter tuning if applicable.
    - **e.** Train and evaluate your model on test data.

f. How do you make sure that your model is not overfitting the data?

g. Plot the learning curve. What can you conclude from this plot?

h. Analyze and discuss model performance.

3. **Build a NN model to predict customer churn (0.35).** Repeat question #2 above but now use a neural network model to predict churn. You can use a simple feedforward neural network. Compare the model to your baseline model with a statistical significance test. Use a box plot to visualize your comparison.

4. **Concept drift detection (0.2).** Use concept drift methods and find out if there is any drift in the data that can be detected. If so, what type of drift is that? Suggest specific actions to adapt your model to the new concept.