DTI 5126: Fundamentals of Data Science

Summer 2022

Assignment 2

Submission Deadline: 14th July 2022 on Brightspace.

This assignment should be **completed individually using R**. Upon completion, present your result in one submission, including the answers generated or plots **(Note: not more than 15 pages**). Where applicable, submit the source codes used to generate your results as a separate attachment.

Part A: Classification (50 points)

Customer churn rate is an important performance metric in the Telecoms industry due to the highly competitive markets. The churn rate enables companies to understand why their customers are leaving. You are hereby provided with the *churn dataset* containing randomly collected data from a telecom company's database. Develop ML models that can help the retention team predict high risk churn customers before they leave. Complete the following:

1. Generate a scatterplot matrix to show the relationships between the variables and a heatmap to determine correlated attributes (10 points)
2. Ensure data is in the correct format for downstream processes (e.g., remove redundant information, convert categorical to numerical values, address missing values, etc.) (10 points)
3. Split the dataset into 80 training/20 test set and fit a decision tree to the training data. Plot the tree, and interpret the results. (10 points)
4. Try different ways to improve the decision tree algorithm (e.g., use different splitting strategies, prune tree after splitting). Does pruning the tree improves the accuracy? (10 points)
5. Classify the data using the XGBoost model with *nrounds = 70* and *max depth = 3*. Evaluate the performance. Is there any sign of overfitting? (10 points)
6. Train a deep neural network using Keras with 3 dense layers. Try changing the activation function or dropout rate. What effects does any of these have on the result? (10 points)
7. Compare the performance of the models in terms of the following criteria: precision, recall, accuracy, F-measure. Identify the model that performed best and worst according to each criterion. (10 points)
8. Use a ROC graph to compare the performance of the DT, XGboost & DNN techniques. (10 points)

Part B

(20 points) A store is interested in determining the associations between items purchased from its Departments. The store chose to conduct a market basket analysis of specific items purchased to analyze customer's buying behavior.

You are hereby provided with a file '*transactions.csv*' containing information for transactions made over the past 3 months.

a) Generate a plot of the top 10 transactions (5 points)
b) Generate association rules using minimum support of 0.002, minimum confidence of 0.20, and maximum length of 3. Display the rules, sorted by descending lift value (5 points).
c) Select the rule from QII-b with the greatest lift. Compare this rule with the highest lift rule for maximum length of 2 (10 points).
    i) Which rule has the better lift? Which rule has the greater support?
    ii) If you were a marketing manager, and could fund only one of these rules, which would it be, and why?