Programming assignment #3

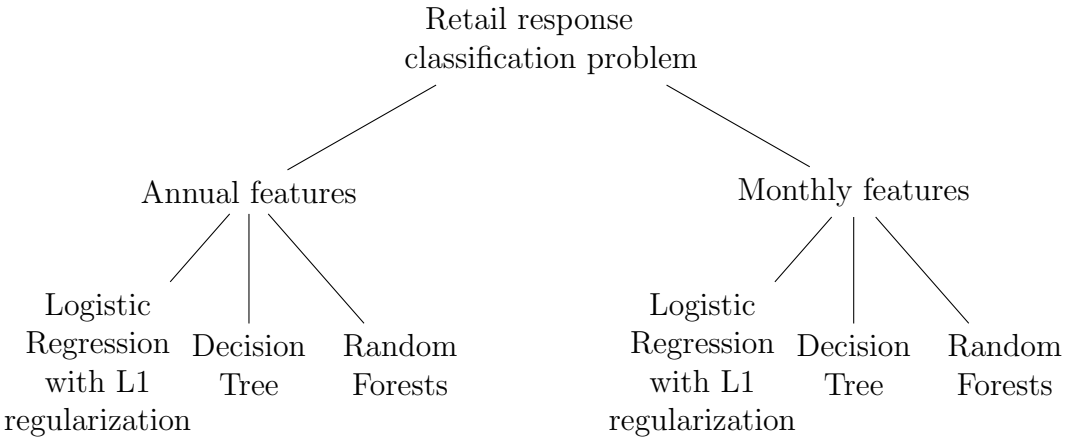
Course: CHE1148H - Data Mining in Engineering

1 Supervised learning

With the features built in Assignment #2, you are now asked to build a model that predicts clients response to a promotion campaign using 3 MLlib algorithms. This is a typical classification problem in the retail industry, but the formulation of the problem is similar to industries such as fraud detection, marketing and manufacturing.

The clients responses are stored in the Retail_Data_Response.csv file from Kaggle. The responses are **binary**: **0** for clients who responded negatively to the promotional campaign and **1** for clients who responded positively to the campaign.

You will explore solving the classification problem with two different sets of features (i.e. annual and monthly) and three different algorithms as shown in the image below.



1.1 Import the monthly and annual data and join

In Assignment #2, you created five different feature families that capture annual and monthly aggregations. Here, you will model the retail problem with two approaches: using annual and monthly features. Therefore, you need to create the joined tables based on the following logic:

Table	annual_features_outputs	monthly_features_outputs
#1	annual_features.xlsx	mth_rolling_features.xlsx
#2	annual_day_of_week_counts_pivot.xlsx	mth_day_counts.xlsx
#3		days_since_last_txn.xlsx
#4	Retail_Data_Response.csv	Retail_Data_Response.csv

In both the annual and monthly features approach, you need to join at the end with table #4, the clients responses. This is simply a table that contains the binary response of the client to our marketing effort as described above and that is the **output** or **label** or **target** that makes this a supervised learning problem.

1.2 Steps for each method (15 points)

Important note 1: When you set up a new cluster in Databricks make sure that you select a runtime version that supports ML (any **10.x ML** should work).
Important note 2: The Learning Spark book github has many useful notebooks in Chapter 10 relevant to ML pipelines.

1. Separate the inputs **X** and the output **y** in two data frames.
2. Split the data in train and test set. Use a test_size value of 2/3 and set the random_state equal to 1147 for consistency (i.e. the course code value). Use the following names for consistency.

Annual	X_train_annual	y_train_annual	X_test_annual	y_test_annual
Monthly	X_train_monthly	y_train_monthly	X_test_monthly	y_test_monthly

3. Pre-process (if necessary for the method).
4. Fit the training dataset and optimize the hyperparameters of the method.
5. Plot coefficient values or feature importance.
6. Plot probability distribution for test set.
7. Plot confusion matrix and ROC curves of train/test set. Calculate precision/recall.

1.3 Comparison of methods (5 points)

Compare the two feature engineering (annual and monthly) and the three modeling approaches (L1 log-reg, tree, forests) in terms of the outcomes of steps 5-7. Which combination of feature engineering and modeling approach do you select as the best to deploy in a production environment and why? **Tabularize** your findings in steps 5-7 to summarize the results and support your decision (how to organize information with tables in Markdown).