FIT 5202 Assignment 2A Feedback Sheet

Student Name: AKSHAYA KUMAR CHANDRASEKARAN

Marked By: Peter Liu

Part A : Working with RDDs and DataFrames

Tasks		Criteria	Yes	Partial	No	Comments
1.1 Creating Spark Session	**1**	No of processors and title of application	☑	□	□	
		Config for spark.sql.files.maxPartitionBytes	☑	□	□	
	2	SparkSession created using the SparkConf	☑	□	□	
1.2 Loading the data	**1**	Schema specified for Process activity data correctly	□	□	☑	
		Schema specified for Memory activity data correctly	□	□	☑	
		-Data loaded into Ds correctly using the schema for both Process & Memory -Row count displayed for both	□	☑	□	Should define the schema instead of inferring them here
		DF cache for both Process & Memory	☑	□	□	
	2	Display the missing data count in each DF for both Process & Memory	☑	□	□	
		Data transformation to proper format	☑	□	□	
1.3 Exploring the data	**1**	Show the count of attack and non-attack for both Process & Memory (for column 'attack')	☑	□	□	
		Show the count of attack TYPE for Process(for column 'type')	☑	□	□	
		Describe the class imbalance (for both 'attack' & 'type' columns)	☑	□	□	
	2	Show the basic statistics for numeric features	☑	□	□	
		Show the top-10 values for non-numeric features (excluding attack label and attack type)	☑	□	□	
		Process plot 1 and description	□	☑	□	No need to create spark df from list and then convert it back to pandas

	3	Process plot 2 and description	☑	□	□	
		Memory plot 1 and description	☑	□	□	
		Memory plot 2 and description	☑	□	□	
2.1 Preparing training data and testing data	1	Randomly split each DF into 80% training and 20% testing for both Process & Memory	☑	□	□	
	2	Use 20% (or lower if due to VM constraint) attack events from 2.1.1 training and maintain 1:2 ratio of stratefied sampling for both Process & Memory	☑	□	□	
		Cache the training data for both Process & Memory	☑	□	□	
		Display the count of each events' data for both Process & Memory	☑	□	□	
2.2 Preparing features, labels and models	1	Discussion on feature selection for Process	☑	□	□	
		Discussion on how to transform features for Process	☑	□	□	
		Discussion on feature selection for Memory	☑	□	□	
		Discussion on how to transform features for Memory	☑	□	□	
	2	Feature transformer / estimator creation for Process	□	☑	□	Normalizer is not mentioned in the discussion
		Feature transformer / estimator creation for Memory	□	☑	□	Normalizer is not mentioned in the discussion
		Bonus task for the custom transformer	□	□	☑	
	3	ML model estimators DT + GBT for Process & Memory	☑	□	□	
		Four Pipelines (DT + GBT) including the above transformers / estimators for both Process & Memory	☑	□	□	
	1	Train ML pipelines (DT + GBT) for Process	☑	□	□	
		Train ML pipelines (DT + GBT) for Memory	☑	□	□	
	2	Test ML pipelines (DT + GBT) for Process, and display the confusion-matrix count (no formatting required for confusion matrix)	☑	□	□	
		Test ML pipelines (DT + GBT) for Memory, and display the confusion-matrix count (no formatting required for confusion matrix)	☑	□	□	
		Compute AUC, accuracy, recall, precision for attack label for Process	☑	□	□	

2.3 Training and evaluating models	**3**	Compute AUC, accuracy, recall, precision for attack label for Memory	☑	□	□	
		Discussion on which metric is more proper	☑	□	□	
	4	Top features: A) extract the feature importance vecctor from model for both Process & Memory	☑	□	□	
		Top features: B) Map feature vector position correctly onto the original feature names for both Process & Memory	☑	□	□	
		Top features: C) Display the top-5 feature names and the importances correctly for both Process & Memory	☑	□	□	
		Discussion on which pipeline model is better for both Process & Memory	□	☑	□	Also consider overfitting issue, interpretability, benefits of using Boosting
		Discussion on whether "ts" column should be added	☑	□	□	
		ROC curve: A) correctly getting ROC data (TPR & FPR) under different thresholds for both Process & Memory	☑	□	□	
		ROC curve: B) properly plotting the curve for both Process & Memory	☑	□	□	
	5	Prepare rebalanced data from full data for both Process & Memory and re-train the corresponding pipeline models	☑	□	□	
		Persist the models	□	□	☑	
3. Knowledge sharing	**1**	Answer number of kmeans jobs and attaching screenshot	☑	□	□	
	2	Explain what the job steps are	☑	□	□	
		Explain in the context of k-means and in the distributed context	□	☑	□	The process in Spark is an optimised version, not entirely the same as the lecture. Note the first seven jobs are related to kmeans|| cluster centre initialisation, while the next two are related to Lloyds's algorithm iteration for clustering

Qualitative Aspect		Organization of tasks in jupyter notebook Adherance to python standards Use of appropriate comments, output readability	☑	□	□	Overall acceptable notebook presentation, consider adding inline reference as well
		Final Grade	Late Submission		0	**HD**