COM6012 Assignment 2 - Deadline: 5:00 PM, Wed 03 April 2019

Assignment Brief

How and what to submit

A. Create a .zip file containing the following:

1) **AS2_report.pdf**: A report in PDF containing answers to all questions. The report should be concise. You may include appendices/references to make it self-contained.
2) **Code, script, and output files:** All files used to generate the answers for individual questions above, **except the data**. These files should be named properly starting with the question number: e.g., your python code as **Q2**_xxx**.py** (**one each question**), your script for HPC as **Q2**_HPC.sh, and your output files on HPC such as **Q2**_output.txt or **Q2**_figB.jpg. Alternatively, now with the support of **Jupyter Hub** (see Lab 1 1.1b for details), you can also develop your answers in **Jupyter Notebook** and submit the as **AS2_xxx.ipynb** and anything else needed to reproduce the results in your report, with Question numbers clearly labelled (e.g., in **bold** or bigger font) in the notebook (note the results should be from HPC, not your local machine for verification purposes).

B. Upload your .zip file to MOLE before the deadline above. Name your .zip file as **USERNAME_STUDENTID_AS2.zip**, where USERNAME is your username such as **abc18de**, and STUDENTID is your student ID such as 18xxxxxxx.

C. **NO DATA UPLOAD**: Please do not upload the data files used. We have a copy already. Instead, please use a **relative file path in your code (data files under folder 'Data')**, as in the lab notebook so that we can run your code smoothly.

D. **Code and output**. 1) Use **PySpark** as covered in the lecture and lab sessions to complete the tasks; 2) **Submit your PySpark job to HPC** with **qsub** to obtain the output.

Assessment Criteria (Scope: Session 6-9; Total marks: 20)

1. Being able to use pipelines, cross-validators and a different range of supervised learning methods for large datasets
2. Being able to analyse and put in place a suitable course of action to address a large scale data analytics challenge

Late submissions: We follow Department's guidelines about late submissions, i.e., a deduction of 5% of the mark each working day the work is late after the deadline, but **NO late submission will be marked one week after the deadline** because we will release a solution by then. Please see this link.

Use of unfair means: "*Any form of unfair means is treated as a serious academic offence and action may be taken under the Discipline Regulations.*" (from the MSc Handbook). Please carefully read this link on what constitutes Unfair Means if not sure.

Please, only use interactive HPC or Jupyter Hub when you work with small data to test that your algorithms are working fine. If you use rse-com6012 in interactive HPC or with Jupyter hub, the performance for the whole group of students will be better if you only use up to four cores and up to 15G per core. When you want to produce your results on the large datasets and want to request access to more cores and more memory, YOU NEED TO USE BATCH HPC. This will be mandatory. We will monitor the time your jobs are taking to run and will automatically "qdel" the job if it is taken much more than expected. We want to promote good code practices (e.g. memory usage) so, please, once more, make sure that what you run on HPC has already been tested enough for a smaller dataset. Based on our own solution to both questions, with the proper setting, you need a maximum of 10 mins for running Question 1 and 20 mins for running Question 2 in batch mode in HPC (these times are for the large datasets using the rse-com6012 queue). It is OK to attempt to produce results several times in HPC, but please, be mindful that extensive running jobs will affect the access of other users to the pool of resources.

Question 1. Searching for exotic particles in high-energy physics using classic supervised learning algorithms [10 marks]

In this question, you will explore the use of supervised classification algorithms to identify Higgs bosons from particle collisions, like the ones produced in the Large Hadron Collider. In particular, you will use the HIGGS dataset.

About the data: "The data has been produced using Monte Carlo simulations. The first 21 features (columns 2-22) are kinematic properties measured by the particle detectors in the accelerator. The last seven features are functions of the first 21 features; these are high-level features derived by physicists to help discriminate between the two classes. There is an interest in using deep learning methods to obviate the need for physicists to manually develop such features. Benchmark results using Bayesian Decision Trees from a standard physics package and 5-layer neural networks are presented in the original paper. The last 500,000 examples are used as a test set."

You will apply Decision Trees for Classification, Decision Trees for Regression and Logistic Regression over a subset of the dataset and over the full dataset. As performance measures use classification accuracy and area under the curve. For Regression, use a sensible threshold for binarizing the decision.

1. Use pipelines and cross-validation to find the best configuration of parameters and their accuracy. Use a sensible grid for the parameters (for example, three options for each parameter). Use the same splits of training and test data when comparing performances between the algorithms. *For finding the best configuration of parameters, use 25% of the data chosen randomly from the whole set* (5 marks).

2. Once you have found the best parameter configurations for each algorithm in the smaller subset of the data, use the full dataset to compare the performance of the three algorithms in the cluster. Once again, use the same splits of training and test data when comparing performances between the algorithms. Provide training times when using 10 CORES and 20 CORES. **Remember to use the batch mode to work on this** (3 marks)
3. Report the three most relevant features for classification or regression for each method obtained in step 2 (2 marks).

Do not try to upload the dataset to MOLE when returning your work. **It is 2.6Gb.**

HINTS: 1) An old, but very powerful engineering principle says: *divide and conquer*. If you are unable to analyse your datasets out of the box, you can always start with a smaller one, and build your way from it. **2)** This dataset was used in the paper "Searching for Exotic Particles in High-energy Physics with Deep Learning" by P. Baldi, P. Sadowski, and D. Whiteson, published in Nature Communications 5 (July 2, 2014). You can compare the results that you get against Table 1 of the paper.

Question 2. Senior Data Analyst at *Intelligent Insurances Co.* [10 marks]

You are hired as a Senior Data Analyst at *Intelligent Insurances Co.* The company wants to develop a predictive model that uses vehicle characteristics to accurately predict insurance claim payments. Such a model will allow the company to assess the potential risk that a vehicle represents.
The company puts you in charge of coming up with a solution for this problem and provides you with a historic dataset of previous insurance claims. The Claimed amount can be zero or greater than zero and it is given in US dollars. A more detailed description of the problem and the available historic dataset is here The website contains several files. You only need to work with the .csv file in **train_set.zip.** The uncompressed file is **2.66 Gb**.

1. Preprocessing (4 marks)
    a. the dataset has several fields with missing data. It is up to you whether you want to remove the rows with missing fields or you want to use an imputation method for filling the missing data.
    b. convert categorical values to a suitable representation.
    c. the data is highly unbalanced: most of the records contain zero claims. When designing your predictive model, you need to account for this.
2. The predictive model. Two strategies for prediction are as follows. Implement both:
    a. You can see the problem as a regression problem where the variable to predict is continuous. In this case, use linear regression in PySpark as the predictive model. Be careful about the preprocessing step above. The performance of the linear regression will depend on the quality of your training data (3 marks).
    b. You can build two separate models. The first will be a binary classifier (of your choice) that will tell whether the claim was zero or different from zero. If

the claim was different from zero, you use Gamma regression to predict the value of the claim. Once again, be careful about the preprocessing step above. The performance of the classifier will depend on the quality of your training data (3 marks).

Use Pyspark to implement your whole pipeline and provide training times when using 10 CORES and 20 CORES. **Remember to use the batch mode to work on the large dataset.**

HINT: An old, but very powerful engineering principle says: *divide and conquer*. If you are unable to analyse your datasets out of the box, you can always start with a smaller one, and build your way from it.