# COM6509 Assignment 1 - Deadline: 12:00 PM, Friday 27th Nov 2020

## Assignment Brief

**Please READ the whole assignment first, before starting to work on it.**

How and what to submit

A. A **Jupyter Notebook** with the code in all the cells executed.

B. Upload your notebook to Blackboard before the deadline above. Name your file as **COM6509_Assignment_1_USERNAME_XXXXXXXXX.ipynb**, where XXXXXXXXX is your username such as **abc18de**

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. So '**./data/**', instead of '**/User/dicaprio/myfiles/mlai/assignment1/'**

**Assessment Criteria** (Scope: Sessions 1-5; Total marks: 30)

1) Being able to follow the steps involved in an end-to-end project in machine learning.
2) Being able to use scikit-learn to design a machine learning model pipeline.

**Late submissions:** We follow the 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.

**Senior Data Analyst at *Intelligent Insurances Co.* [30 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.** The company provides the file **train.csv** that contains records of historic claims. They also provide a **data dictionary** file that describes the variables in the dataset. You will follow the checklist in a machine learning project and provide a predictive model that better generalises to a test set.

In this exercise, you will use train.csv to design your ML model. We will use your ML model over an independent test set that we will not share with you while you work on the assignment. We want to simulate the real life scenario where an ML model that has been designed using a dataset is then deployed into production. Imagine that assessing your

model over this independent test set is equivalent to having your model working on in production. When we provide the feedback for your Assignment, we will let you know the performance of your model over this test set.

Before you begin, partition the data into train and test sets. Do not have a look at your test data until you have designed your final model. Use the root-mean-squared error (RMSE) as the performance measure. You are allowed to use scikit-learn for the Assignment.

1. **Data preprocessing [7 marks]**
    a. The dataset has several fields with missing data. Choose a method to deal with missing data and justify your choice [2 marks].
    b. Convert categorical values to a suitable representation. *Notice that there are many categorical variables in the dataset. If you use all the categorical variables you will end up with a large feature space. Feel free to ignore categorical variables that will increase your feature space considerably but use at least five categorical variables* [2 marks].
    c. The data is highly imbalanced: more records contain zero claims than not. When designing your predictive model, you need to account for this [3 marks].

    Steps a, b and c above are the minimum requirements. You may choose to perform additional steps of data preprocessing (e.g. transform the variables, standardisation, normalisation, etc).

2. **Performance using a single model [8 marks]**

    You can see the problem as a regression problem where the variable to predict is continuous (the claimed amount in USD). Be careful about the preprocessing step above. The performance of the regression model will depend on the quality of your training data. Compare the performance of the following models:

    a. Linear regression [2 marks].
    b. Ridge regression [2 marks].
    c. Random forests for regression [2 marks].
    d. Gradient tree boosting for regression [2 marks].

    For each model, use grid search with at least three options for each parameter and clearly report the performance measure over a validation set.

3. **Performance using a combination of two models [6 marks]**

    You can build a prediction model based on two separate models in tandem (one after the other, see Figure 1). Once again, be careful about the preprocessing step 1. For this step, use the same training and validation sets from step 2.

    a. The first model will be a binary classifier that will tell whether the claim was zero or different from zero. Compare the following classifiers: random forests for classification and gradient boosting for classification [2 marks].

b. For the second model, if the claim was different from zero, train a regression model to predict the actual value of the claim. Compare the same models that you used in step 2 [2 marks].
c. Use the tandem model built from steps a and b, for predicting in the same validation data used in Step 2, and report the performance [2 marks]

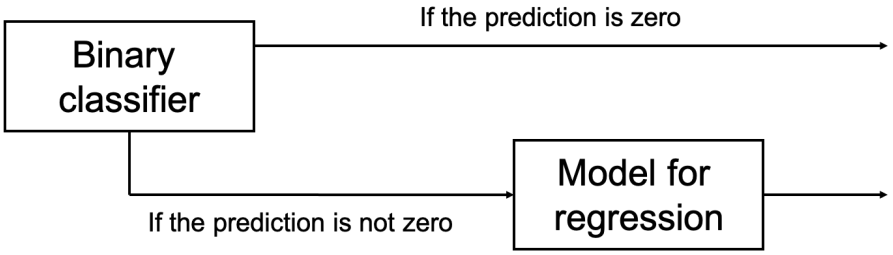


Figure 1. A predictive model that uses a binary classifier and a model for regression.

**4. Report the performance of the best models over the test set [2 marks]**

Compute the performance metric over the test set for the best model in Step 2 and the best model in Step 3. **WARNING: Be careful to not confuse the test set given in data.zip, test.csv, with the test set you create as part of the assignment and your machine learning check-list.**

**5. Present your solution [4 marks]**

Provide four interesting and meaningful observations/comments about your machine learning pipeline, with minimum three sentences for each observation/comment.

**6. Create a function that contains the best model you built from Steps 1 to 4 that we will use to assess the performance of your design over an independent test set [3 marks].**

Your function should have the following form:

```
In [ ]: prediction = my_insurance_claim_predictor(Xtest)
```

We have provided you with a test.csv file that contains a few instances of the larger test data we will use. **Notice that this file DOES NOT contain the Claim Amount.** Import the test.csv and name the input test data as Xtest that the above function will use. When testing your code, we will only change the file test.csv that we provided here with a larger test.csv file. We will then use the "prediction" that your function provides and the true claims in the test set that we have to compute the RMSE.

As users of your ML project, we are only expected to be able to use the end product of your design which is summarised in the instruction above. If your function does not

provide the predictions as expected, you will lose these marks. Assume that the test input data we use here will have the same form than the one given in the train.csv file. Verify with the test.csv file provided. We will use your predictions to compute the RMSE over the independent test set and report the performance back to you when we provide you with the feedback.

## Additional comments

### Regarding GridSearch

As you will notice, the gridsearch steps in the Assignment take some while to run (they can take until two hours for gradient boosting, if you use the whole training data). When marking your assignment, we would prefer not to wait for a long time for the grid search to finish, therefore may I please ask you to comment the code for the grid search, once you're happy with the best model that you get from each grid search?