# ISSS610
# Applied Machine Learning

## Assignment 1

## Linear Regression and Logistic Regression

### [Question 1] Data Exploration

In this assignment, you are given an air ticket sales dataset from airport PEK to airport SHA. Unfortunately, there is not much information about this dataset except column names, you need to carry out your own analysis to understand this dataset first. The information below is known to you at this moment:

- cabinClass: Y = Econ Class, C = Business Class, F = First Class
- rate: ranged from 0 to 1, stands for discount rate, lower is cheaper
- createDate: the actual date this data is generated
- dateDifference: the date difference between the create date and the actual departure date. For example, if the value is 7, it simply means I am booking the ticket one week ahead.

Answer the following questions. You need to write code to answer questions labelled by "code required".

(1) What is the minimum time interval between createDate and departureDate for valid records (code required, 2 marks)? Does it make sense (0 mark)?

(2) Is the column price the original price or the discounted price (code required, 2 marks)? Column rate stands are the discount rates, the lower the cheaper. You may look at column rate to answer this question.

(3) What are the original prices for each cabin class and each flight (a flight is identified by its flight number)? Tabulate flight number, departure time in hh:mm, cabin class and original price for each flight (code required, 2 marks). You may use your own reasonable assumptions to find out departure time and the original prices.

## [Question 2] Linear Regression

(4) Are the original air ticket prices related to the departure time of the flights for each airline (code required, 4 marks)? Airlines are identified by the first two letters of the flight numbers.

(5) Are the discount rates related to how long in advance tickets are purchased (code required, 3 marks)? You may use your own assumptions or divide the dataset into smaller subsets according to your own assumptions. Hint: in case you need to translate a categorical column into a set of columns where each represents one specific value on the original column you may use pandas.get_dummies. You may refer to section “Dummy coding” under Wikipedia article on categorical variable: en.wikipedia.org/wiki/Categorical_variable#Dummy_coding

(6) Choose your own attributes to build two Lasso or ridge regression models, one for predicting rate and the other for predicting price (code required, 4 marks). Are the performances of the two regression models significantly different? Why is it so (0 marks)? Hint: When choosing attributes, you always need to think about use cases of the machine learning models.

## [Question 3] Logistic Regression

(7) Choose your own attributes to build a binary classifier on if there is any discount on the purchase price: class 0 for no discount, class 1 for price being discounted. Identify 2 to 5 important factors that determine if there is discount (code required, 4 marks).

(8) Let us now create three classes for all records with $rate < 1$: class 1 for $0.75 < rate < 1$, class 2 for $0.5 < rate \le 0.75$ and class 3 for $rate \le 0.5$ Choose your own attributes to build an ovr multi-class logistic regression classifier for this purpose. Plot the 5 ROC curves in one plot, 3 for each class, 1 for micro-average and another for macro-average, and report the respective AUC under each of the ROC curves (code required, 4 marks). You may refer to sklearn.metrics.roc_curve: scikit-learn.org/stable/auto_examples/model_selection/plot_roc.html

(9) Imagine we now stack the binary classifier and the multi-class classifier together, we first predict if there is any discount on the price, and then predict which category the discount belongs to. The final prediction is thus labelled with 0 – no discount, 1 – discount rate between 0.75 and 1 exclusive both, 2 – discount rate between 0.5 exclusive and 0.75, and 3 – discount rate is less than or equal to 0.5. Build the confusion matrix for these four classes (code required, 2 marks). Tabulate precision, recall and F1-score: en.wikipedia.org/wiki/F1_score with macro averaging and weighted averaging: scikit-learn.org/stable/modules/generated/sklearn.metrics.f1_score.html (code required, 2 mark). Does the performance improve by replace ovr multi-class logistic regression to multinomial multi-class logistic regression (code required, 1 mark)?