# Final Exam. AMS 580

# Name:\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_SBU ID:\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

# Dear all, the Final Exam (Wednesday, May 17, 8:00-10:45am) is an in-person exam to be held in the same classroom, and with the same zoom link. This is an open-book, open-internet exam – however you must do so entirely on your own. Please email the completed Final exam to Professor Zhu (and not your TA) with one email entitled “Final Exam, AMS 580, Firstname\_Lastname, SBU ID” with two files attached (1. the RMD file, and 2. the Output file of your RMD in either word or pdf – the file names should be: Firstname\_Lastname\_ID.\*\*\*) no later than 10:45am, at: [wei.zhu@stonybrook.edu](mailto:song.jiecheng@stonybrook.edu)

#### Part I. Supervised Learning with the Diabetes Prediction Data

The **diabetes\_prediction\_dataset.csv** data contain 8 predictors and one binary response variable diabetes (= 0 or 1), which is the true class label**.**

For this dataset, **sensitivity** is defined as a case labeled 1 being classified to label 1, while **specificity** is defined as a case labeled 0 being classified to label 0.

**1.** For the entire dataset, please perform the data cleaning as instructed before; namely, delete observations with missing value(s). Please report how many cases (namely, data points) are left after this step. Then please use the random seed 123 to divide the cleaned data into 75% training and 25% testing.

**2.** Next, Please use the *neuralnet* package in R to build the various neural network classifiers. For this question, we shall NOT perform data standardization (normalization).

1. Now we shall build the best classifier to predict the class label using the training data and the Perceptron model with (i) one hidden layer with 4 neurons, (ii) the default loss function of “sse”, and (iii) the default activation function of “logistic”. Please compute the Confusion matrix and report the sensitivity, specificity, and the overall accuracy using the testing data.
2. Next we shall build the best classifier to predict the class label using the training data and the Perceptron model with (i) one hidden layer with 4 neurons, (ii) the loss function of “ce” (namely, cross-entropy, or the negative log likelihood), and (iii) the default activation function of “logistic”. Please plot the perceptron model obtained using the training data. Please compute the Confusion matrix and report the sensitivity, specificity, and the overall accuracy using the testing data.
3. Which neural network model provides the best overall accuracy among (a) and (b)? For this best model only, please add the predicted label for every test data point. (Please do not print this data set out! It will be used in Question 5 below.)
4. Now we shall use the *randomforest* function in R to build the random forest classifiers. For this question, we shall NOT perform data standardization (normalization).

1. Please first build the best random forest to predict the class label using the training data. Please compute the Confusion matrix and report the sensitivity, specificity, and the overall accuracy using the out of bag (OOB) samples.
2. Next please use this random forest to predict the class label in the testing data. Please add the predicted label for every test data point. (Please do not print this data set out! It will be used in Question 5 below.) Please compute the Confusion matrix and report the sensitivity, specificity and the overall accuracy for the testing data.
3. Please plot the variables importance measures using
   1. *MeanDecreaseAccuracy*, which is the average decrease of model accuracy in predicting the outcome of the out-of-bag samples when a specific variable is excluded from the model.
   2. *MeanDecreaseGini*, which is the average decrease in node impurity that results from splits over that variable. The Gini impurity index is only used for classification problem.
4. Please show the importance of each variable in percentage based on *MeanDecreaseGini*.
5. Now we shall build the best classifier to predict class label using the training data and the SVM with polynomial kernel. We shall find the optimal tuning parameters C, degree and scale by using the command line:

tuneLength = 4

Please (i) report the optimal parameter values, and (ii) compute the Confusion matrix and report the sensitivity, specificity, and the overall accuracy using the testing data. Please add the predicted label for every test data point. (Please do not print this data set out! It will be used in Question 5 below.

1. Now, we shall build an ensemble classifier using the majority vote of: (1) the random forest, (2) the best neural network model, and (3) the SVM model obtained above. The way the majority vote works is that each case in the testing data is classified into the label with the majority vote from the three classifiers. For example, if a case is predicted to be 1, 1, 0 – then the case is classified as 1. Now, please compare the majority vote to the true class label (y) --- compute the Confusion matrix and report the sensitivity, specificity, and the overall accuracy of our new ensemble classifier using the testing data.

#### Part II. RNN & LSTM with the AAPL Data

This dataset contains the historical stock prices of **Apple (AAPL)**.

We will use a Long Short-Term Memory (LSTM) neural network to predict the daily close price of **AAPL** in this assignment. The data spans from 2021-03-29 to 2022-03-28.

When applying LSTM, it is very important to **normalize** the data. One widely-used method is **min-max scaler**, which means we transform x to be **(x - min(x)) / (max(x) - min(x)).** Instead of using the min and max of that day, we shall use the min and max of the **previous** day (lagged min & max).

1. Plot the close price vs. date to visualize the data we will analyze. Then, use the ‘min-max scaler’ to normalize our stock price data. Scaled\_x = (x-lagged\_min(x))/(lagged\_max(x) – lagged\_min(x)). Please report the values of the last 10 scaled close prices.
2. Divide the cleaned dataset into two parts, the last 10 prices for testing y and the rest for training (the last 10 prices in the training set will be used as testing x). Please report how many days of stock price are divided into the training set.
3. LSTM algorithm creates predictions based on the lagged values, which means we need to look back as many previous values as many points we wish to predict. Here we want to do a 10-days forecast, so we need to base each prediction on 10 data points. (We lag the data 10 times, so that each prediction is based on 10 values, and arrange lagged values into columns) Additionally, keras LSTM expects specific tensor format of shape of a 3D array of the form [samples, timesteps, features] for predictors (X) and for response (y) values. Please create matrices for training and testing predictors and response in the 3D form, and report their dimensions by using dim().

1. Please first build the predictive model to predict 10-days stock close price using the training data and the LSTM method with only one LSTM layer with 200 hidden units, and the loss function of ‘mse’. Please make predictions on the 10 observations in the testing set by using the last 10 in the training dataset and compute the Test MSE using the testing data. Scale the predicted stock price back and plot the 10-days predictions and the true stock close price in the same figure. Also, try to predict the close price of 2022-03-29.
2. Please first build the predictive model to predict 10-days stock close price using the training data and the RNN method with only one RNN layer with 200 hidden units, and the loss function of ‘mse’. Please make predictions on the 10 observations in the testing set by using the last 10 in the training dataset and compute the Test MSE using the testing data. Scale the predicted stock price back and plot the 10-days predictions and the true stock close price in the same figure. Also, try to predict the close price of 2022-03-29.
3. Please compare the performance of the algorithms in Question 4 and Question 5 in 10-days forecast using the testing data.

#### Part III. Regression Analyses & Variable Selections with the Data Science Salary Data

The Data Science Job Salaries Dataset ‘ds\_salaries.csv’ contains 11 columns, each are:

work\_year: The year the salary was paid

experience\_level: The experience level in the job during the year

employment\_type: The type of employment for the role

job\_title: The role worked in during the year

salary: The total gross salary amount paid

salary\_currency: The currency of the salary paid as an ISO 4217 currency code

salaryinusd: The salary in USD

employee\_residence: Employee's primary country of residence in during the work year as an ISO 3166 country code

remote\_ratio: The overall amount of work done remotely

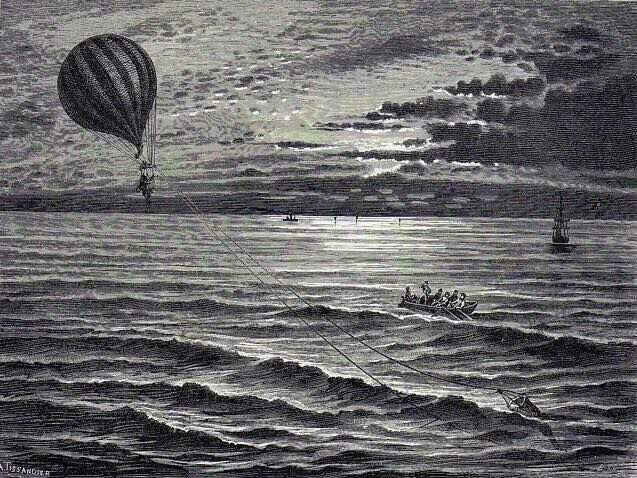
company\_location: The country of the employer's main office or contracting branch

company\_size: The median number of people that worked for the company during the year.

Our goal is to establish a sensible regression model predicting the output variable ‘salary\_in\_usd’.

1. Please perform data cleaning by checking whether there are any missing values and if so, please delete observations with missing values. Please report how many observations with missing values we have in our dataset. Please use the random seed 123 to divide the data into 75% training and 25% testing.
2. Now we shall perform penalized regression analysis with three different methods.

1. Please first find the best Ridge Regression model using the training data. Please (a) find the best **λ** valuethrough cross-validation and display this value; (b) display the coefficients of the fitted model; and (c) make prediction on the testing data, plot the observed response variable on the x-axis, and the estimated response variable on the Y-axis, and report the RMSE and the Coefficient of Determination .
2. Please first find the best LASSO model using the training data. Please (a) find the best **λ** valuethrough cross-validation and display this value; (b) display the coefficients of the fitted model; and (c) make prediction on the testing data, plot the observed response variable on the x-axis, and the estimated response variable on the Y-axis, and report the RMSE and the Coefficient of Determination .
3. Please first find the best Elastic Net model using the training data. Please (a) find the best **tuning parameter** valuesthrough cross-validation and display these values; (b) display the coefficients of the fitted model; and (c) make prediction on the testing data, plot the observed response variable on the x-axis, and the estimated response variable on the Y-axis, and report the RMSE and the Coefficient of Determination .
4. Please discuss which penalized regression method is the best for the data science salary data, and why.



**Unknown Artist?**