**Texas Tech University**

**Department of Computer Science**

**Course:** Introduction to Artificial Intelligence **Group:** 1

**Instructor:** Dr. Juan Carlos Rojas **Email:** [Juan-Carlos.Rojas@ttu.edu](mailto:Juan-Carlos.Rojas@ttu.edu)

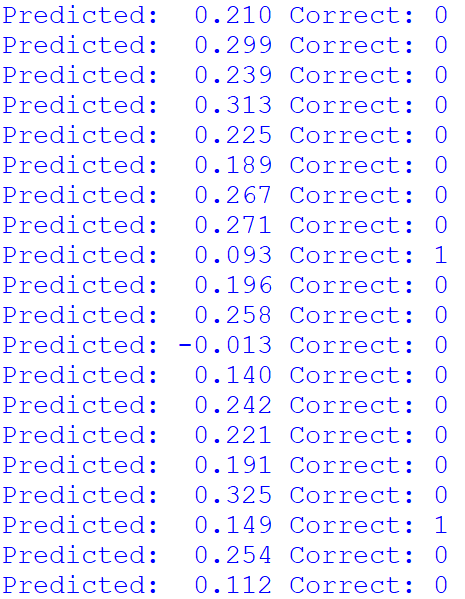
**Hours:** 8:00 – 12:00 (Saturdays) **Room:** 320

# Homework 3

Due Saturday, June 8 at 8:00am.

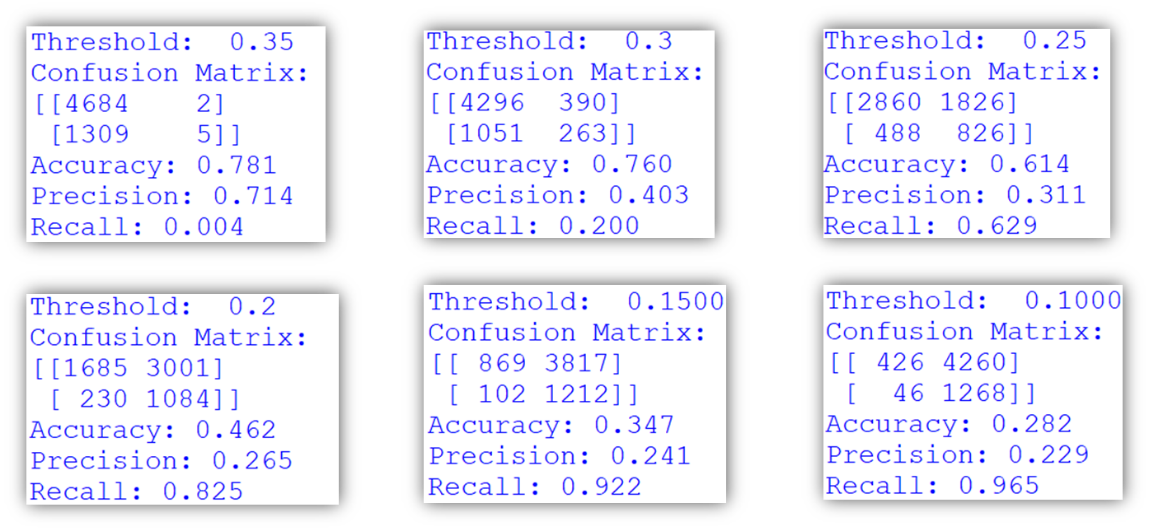
## Practice 1

* Load the credit default dataset pickle file (credit\_card\_default\_dataset.pickle)
* Train a linear prediction model against **all** the available features
* Predict new labels for the test data
  + Print the first 20 predictions, and also the first 20 correct labels



## Practice 2

* Load the credit default dataset pickle file (credit\_card\_default\_dataset.pickle)
* Train a linear prediction model against all the available features
* Compute the prediction probabilities
* For each of the following thresholds: 0.35, 0.30, 0.25, 0.20, 0.15, 0.10
  + Create a vector of binary predictions
  + Compute the confusion matrix
  + Compute the accuracy
  + Compute the precision
  + Compute the recall

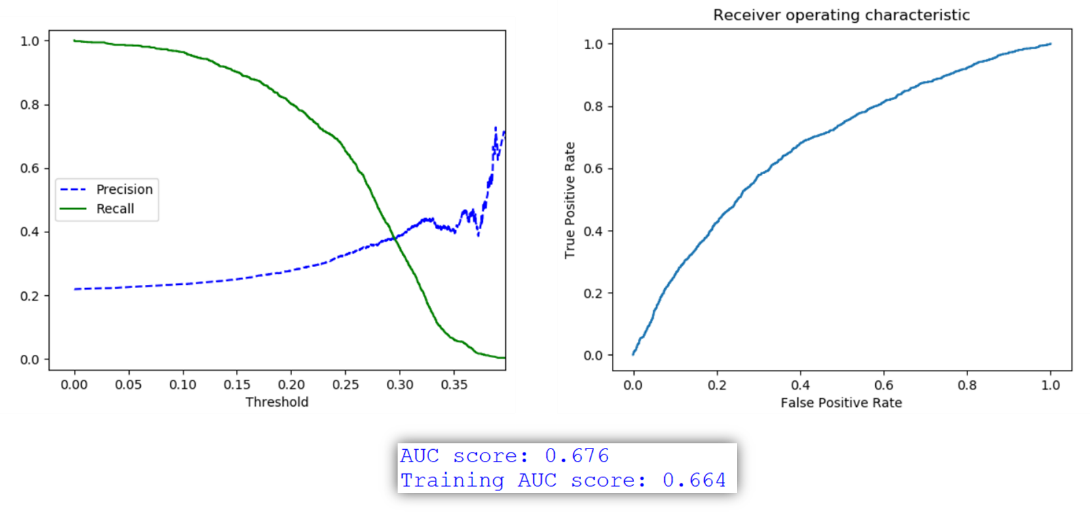


## Practice 2b (Optional)

* Implement linear regression using weight-balanced class values
  + Scale the outputs in a ratio inversely proportional to the ratio of classes
  + Compute a Precision & Recall graph
    - How does it look, compared to the previous case?
  + Compute the ROC and AUC
    - Did it improve?
    - Is weight-balancing the classes helpful?

## Practice 3

* Train a logistic regression model for the credit default dataset (credit\_card\_default\_dataset.pickle)
  + Compute and plot a Precision & Recall plot
  + Compute and plot a Receiver Operating Characteristic (ROC) plot
  + Compute the AUC score for the test data
  + Compute the AUC score for the training data

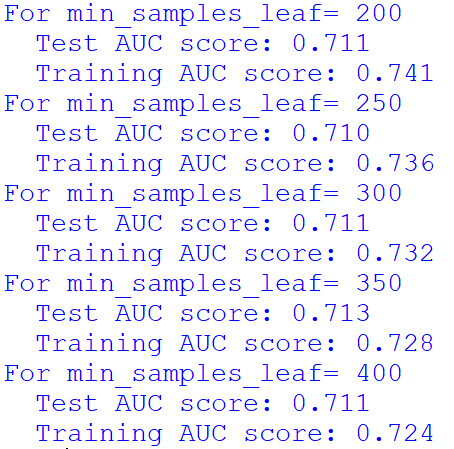


## Practice 3b (Optional)

* Extract a sorted list of feature importances, based on the logistic regression coefficients
* Are they the same as for linear regression?

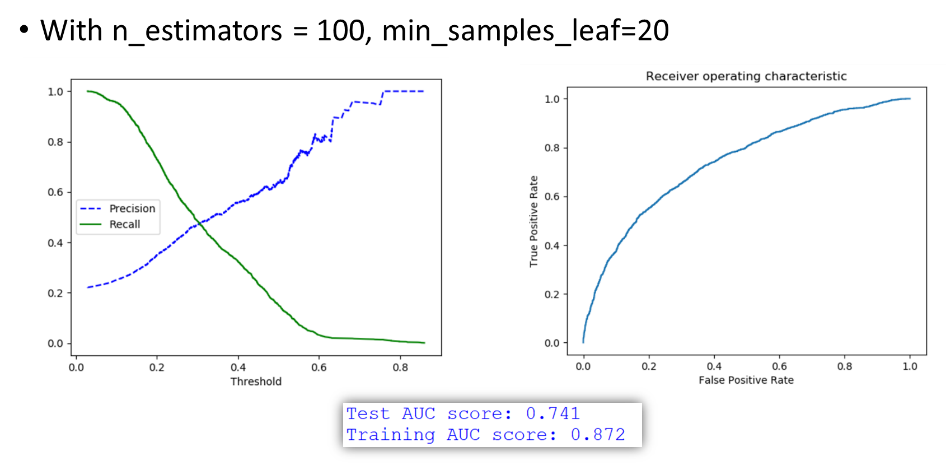
## Practice 4

* Train a decision tree classifier for the credit default dataset (credit\_card\_default\_dataset.pickle)
* Try different values of the parameter *min\_samples\_leaf.* For each:
  + Compute the AUC score for the test data
  + Compute the AUC score for the training data
* What is the best value of *min\_samples\_leaf* that you could find?



## Practice 5

* Train a random forest classifier for the credit default dataset (credit\_card\_default\_dataset.pickle)
* Try different values for the parameter *n\_estimators.* For each:
  + Compute the AUC score for the test data
  + Run multiple times. Are your results repeatable?
  + Pick a value that provides consistent results
* Once you are satisfied with the number of estimators, try different values of *min\_samples\_leaf*
  + What is the best AUC score you could get?



## Practice 5b (Optional)

* Use GridSearchCV in scikit-learn, as a way to automatically search for the best tuning parameters in your random forest.

## Practice 6 (Optional)

* Consider pruning some features that may not be adding independent information
  + Try removing Prev\_Bill5 and Paid\_Amt5
  + Re-train some of your classifiers (logistic, random forest)
    - Did results degrade, stay the same, or improve?
  + Try removing more features and see what happens