**Assignment Questions**

* What is R^2 value?? How is it different from accuracy?
* What is Cost function?? How can one reduce it?
* Difference between a gradient descent & stochastic gradient descent (SGD) ? How does it help reducing the error rate?
* Interpretation of Logistic regression output.. what is pseudo R^2? What is log-likelihood? What is Z-value with respect to logistic model? What is std. error in logistic regression? What does it say about each variable in the model?
* What is multicollinearity? (Discussed in class) what is dummy variable trap? What is perfect multicollinearity and imperfect multicollinearity? What is the best way to prevent dummy variable trap?
* What is the use of entropy and Information Gain in Decision trees? How does it help in deciding the split? (Explanation with example would be great)
* What is Gini coefficient?? What role does it play in Decision Trees.
* What is sensitivity & Specificity with respect to the confusion matrix?? What does it say about the model?
* How to know if your data has overfit or underfit? How does one prevent overfitting of data??
* Other examples of Ensemble techniques other than random forest? How does random forest combine weak classifiers (Decision trees) to produce a smaller error rate?
* What are boosting and bagging? How are they different from each other? What advantage does Random forest have over these 2 ?
* What is TF-IDF value? Formula & explanation how it is better from just a DTM? I said that it tells us about the word importance in a document / corpus.. how does it do that?
* Why is Exploratory Data Analysis (EDA) important before building your model?