**Project Final Report Common Mistakes**

**Problem description**

* Is the problem or focus clearly defined? Does it make sense? Does it have any business value?

**Data set description: origin, data points, variables**

* Is the data source clearly defined? Are independent variables and dependent variable clearly defined? Why did you choose these IVs?
* Does the report contain descriptive statistics for variables? For categorical variable, descriptive statistics should include counts and percentages of each value within a variable.

**Data preprocessing activities and results**

* Is there discussion about missing value problem? What is the missing data rate?
* For imbalanced dataset, is there discussion about imbalanced data problem? For extremely imbalanced dataset (e.g. credit card fraud), the report should include what process students conduct to balance the dataset.
* What data type conversion or reformat has been done?

**Intended algorithms to be used and rationale**

* Make sure the prediction makes sense – using common sense, do you believe DV is determined by the IVs?
* Naive Bayes assumes independence among all IVs, does your variables satisfy such condition? Need to provide a correlation test.
* Did the student use two predictive methods? (decision tree, naïve Bayes, SVM and neural network).
* If the student has categorical DV: all four predictive methods (decision tree, naïve Bayes, SVM and neural network) can be used.
* If the student has numeric DV (integer or continuous): all four predictive methods (decision tree, naïve Bayes, SVM and neural network) can also be used. Decision tree and neural network R package directly support continuous DV. Naïve Bayes and neural network have version supporting continuous DV, but need to specify proper parameter for the R function. Confusion matrix cannot be applied.

**Model Execution/Evaluation**

* Descriptive model should be done before predictive models
* What are the parameters you set for your model?
* Need to provide a complete view of the decision tree.
* Is there performance comparison of these two predictive methods?
* Baseline ratio missing
* The accuracy rate is a little too low. What might be the reasons?