**Data Science Interview Questionnaire**

This questionnaire will be divided into a few sections exploring different areas of data science work and background knowledge.

**Programming:**

1. What are some differences between sets, lists, and tuples in Python?

**Answer**:

1. What is a class?

**Answer:**

1. Can you explain the benefits of using a Version Control System?

**Answer:**

1. Why is it important to build shared coding standards within a team?

**Answer:**

**Data Science:**

1. What is the Curse of Dimensionality?

**Answer:**

1. Can you tell us some ways of reducing dimensionality?

**Answer:**

1. Can you explain what bagging is?

**Answer:**bagging is an ensemble learning method. It sample subsets from the entire data set with replacement, and then train the same model with the same hyper- parameters from different subsets, and produce different trained weak models, and finally average the predictions of these models to produce the final prediction. This methodology is suitable for the models with big variants, like decision tree.

1. **True or False?** Statistical significance of a factor or effect on a process affecting a variable Y implies it will have a big impact on changing Y:

**Answer:** False. Statistical significance is a measure of how confidence we are about the association between X and Y, not how big impact on the change of Y. How big the impact is X on the change of Y is measured by the coefficient, not the p-value.

1. **True or False?** Standardising predictor variables generally changes the predictions of statistical models.Explain your answer.

**Answer:**

1. **True or False?** A random forest model can extrapolate a gradient between a predictor variable and target variable beyond the limits of the predictor variable in its training data. Explain your answer.

**Answer:** True. Over fitting can be avoided by random forest, thus it can extrapolate beyond its training set. The reason is random forest use only a sub-set of the predictor variables to split the training set while constructing a node, thus it will not be over fitted to some variables. This gives the model the ability to generalize beyond the limits of the predictor variable.

1. **True or False?** You build a GLM with 2 predictor variables. The more correlated your predictor variables are with each other, the less likely your regression coefficients will change drastically when you train the model on new data. Explain your answer.

**Answer:**Flase. The correlated predictor variables leads to the problem of multicollinearity. Multicollinearity Interpretation will makes model coefficients swing wildly, some times even the signs can be inverted. Thus the p-value of the coefficients are not reliable anymore.