**Common Statistical Terms**

**Classification**

* Output is a set of categories. The statistical task of predicting which discrete category an observation or record belongs to

**Data Set**

* Contains the inputs and output, occurring as observations of "variables"

**Efficiency**

* The computational and statistical efficiency required of the problem space. Statistical - minimal number of records needed to achieve the desired statistical accuracy. Computational - how much time do you need to learn the relationship between inputs and outputs

**Feature Engineering**

* A way of maximizing the predictive value of inputs. Specifically, the creation of new random variables from those already observed in the dataset with the specific aim of achieving a higher degree of accuracy and/or greater precision for a predictive model.

**Loss Function/Error**

* A way to measure the quality of the outputs. Reflects the success of a statistical model. Ideally this is connected to the respective business goal or desired task performance.

**Machine Learning**

* Programming a computer to run a statistical model repeatedly (i.e., train it) in order to find the weighting/combination of inputs that optimally reduces or minimizes the loss/error function associated with your output.

**Overfitting**

* Learning a relationship between inputs and outputs that is idiosyncratic to that sample. Results with a lower likelihood of generalizing to a new set of records

**Prediction (or Confidence) Interval**

* Measuring the uncertainty of the outputs. Allows us to measure uncertainty in the output. A interval which ought to contain the "true" value of the parameter being estimated XX% of the time, assuming the model is correct.

**Predictive Model**

* A mathematical equation or a computer program that takes an input and generates an output

**Regression**

* Output is an on order of continuous magnitude. How much does X affect Y?

**Sampling**

* Means of gathering inputs and outputs by drawing a smaller number of units from some larger set, in such a way, that the smaller set has properties which represent the entire, larger set.  When done properly - helps ensure that you learn the correct relationship between inputs and outputs

**Supervised Learning**

* The Inputs. Statistical models in which the true answer (i.e., label, value) is explicitly observed in your data. Part of the input is the targeted output

**Unsupervised Learning**

* The Inputs. Statistical models in which the true answer (i.e., label, value) is NOT explicitly observed in your data (e.g., clustering)