Questions

1. Please describe the central limit theorem and provide an example.

Given a series of independent and identically distributed random variables drawn from a population, with sample mean and sample variance , the central limit theorem states that the distribution of the sample means will be approximately normally distributed around the true population mean given enough samples (usually above 30), even if the population distribution is non-normal. The sample variance will also approximately represent the population variance.

A classic example are dice rolls. The average value, or population mean, of many dice rolls is around 3.5 given a dice between 1 – 6. If you roll a fair dice once, average the roll, and repeat 30 times, you would get a fairly uniform shape since the probability of any given roll is equal. However, if you roll the dice 10 times, average this roll, and repeat 30 times, the distribution of these averages will start to represent a normal distribution around the population mean. If you keep increasing the number of rolls and times you repeat, the distribution will start to converge and peak towards 3.5.

1. Describe a classification algorithm that you have previously put into production and why it was chosen.

For our advertising modeling, I had worked with another DS to put an XGBoost algorithm into production. Prior to starting the company, the existing modeling process was using logistic regression as it was fairly quick and easy to train. However, often times we did not need interpretability in our models and some clients were asking for increased performance for our advertising targeting. After going through the model production pipeline, XGBoost outperformed all other models (logistic techniques and bagging techniques) on the out of sample test set by 5-8% on our various datasets, and the production runtime was only a few hours to predict on 250MM individuals. This boost in performance was worth the tradeoff in computation time.

1. Describe the difference between bagging and boosting methods, and when to use one or the other.

Bagging methods use many weak classifiers trained on different features of the data chosen randomly without replacement. The average of all of the predictions (for regression) or the majority vote (for classification) becomes the final prediction. A good example of a bagging method is Random Forest.

Boosting methods create a collection of weak classifiers, where each classifier learns from the errors of the prior classifier. The final classifier would then be the best perofrming model.

1. Describe 2 regularization techniques for a random forest model