STATISTICS SEMINAR

Michael Whitley
Master's Defense

Wednesday, November 18, 2015
Dickens Hall, Room 109, 10:00-11:00 am

USING STATISTICAL LEARNING TO PREDICT SURVIVAL OF PASSENGERS ON THE RMS TITANIC

When exploring data, predictive analytics techniques have proven to be effective. In this report, the efficiency of several predictive analytics methods are explored. During the time of this study, Kaggle.com, a data science competition website, had the predictive modeling competition, "Titanic: Machine Learning from Disaster" available. This competition posed a classification problem to build a predictive model to predict the survival of individuals on the RMS Titanic. The focus of our approach was on applying a traditional classification and regression tree algorithm. The algorithm is greedy and can over fit the training data, which consequently can yield non-optimal prediction accuracy. In efforts to correct such issues with using the classification and regression tree algorithm, we have implemented cost complexity pruning and ensemble methods such as bagging, random forests, and boosting. The decision trees and prediction accuracy of each method are presented and compared.