**Big Data Group Assignment - Spark SQL and Machine Learning**

Group 1: Omar Hanley, Zachary Rosen, Yannan Zhu, Leslie Ling, Todd Schaffer

**Business Problem**

In the United States, there are over 5300 different colleges and universities that prospective students can choose from. Colleges are now facing an even more challenging competitive climate with universities having the capabilities of offering the same educational content on an online platform device without requiring students to be on campus. This has made colleges adapt to the changes in the educational system and look for ways to differentiate themselves from not only regional competitors but global competitors as well. A linear regression model will be run on the student-faculty ratio and graduation rate to understand if this metric is important in reducing the churn rate that most colleges face. A regression decision tree model will also be run on the ten different metrics which include, student-faculty ratio, New students from top 10% of high school class, New students from top 25% of high school class, Number of full-time undergraduates, Number of part-time undergraduates, Instructional expenditure per student, Estimated student personal spending, out of state tuition price, room/board cost and cost of books to understand which features help provide the greatest influence to graduation rate. The higher the graduation rate for schools, the higher probability of keeping students for the entire life cycle. A csv file containing over 700 different colleges with these metrics will be used to help address this business problem.

**Implementation**

The data was read in and loaded using the sqlContext.read.format function in python. Both models then used the VectorAssembler function to subset the columns needed to create the feature x/’s and y variable. The data was then randomly split with 80% of the data in a training variable and 20% in a test variable. The decision tree was created using the DecisionTreeRegressor function and the linear regression model was run for 15 iteration cycles using the LinearRegression function within Pyspark.

**Model Results**

The linear regression model had training RMSE of 15.99 and test RMSE of 15.43 indicating that there was high variance within data. The training R2 was .07 and test R2 was .15 indicating that there does not seem to be a high correlation between graduation rate and the student-faculty ratio. The average graduation rate was 65% with a 17% standard deviation.

The regression decision tree model produced a similar test RMSE of 15.19. After using the function that extracts the most important features for a decision tree model, it was apparent that out-of-state tuition was the most influential feature in determining the graduation rate with 48%. The second most important feature was new students from the top 25% of their high school class with 19.8%. The other features had little effect on graduation with most being under 10% or 5%.

**Conclusion**

Both models found that the student to faculty ratio is not a hard determinant in the graduation rate for universities. The most influential metrics that affect graduation rate is bringing students who are in the top 25% in their high school class and the cost of out-of-state tuition. In the future, colleges and universities should strategically focus on targeting seniors who fall within the top 25% as these students will have a longer life-cycle at the school. They could also look at implementing different procedures for charging out-of-state tuition prices by either offering more scholarships or possibly allowing out-of-state students to pay in-state tuition price after being at the school for one year to increase the probability of these students staying.