Option #1: Credit Data Mining

Scott Miner

Colorado State University – Global Campus

**Option #1: Credit Data Mining**

The assignment asks the student to submit what he will be doing for his portfolio project. I plan to pursue option #1. Option #1 asks the student to perform exploratory data analysis (EDA) on a file titled, ***GermanCredit.csv*** (*Module 8: Portfolio Project*, n.d.). Shumeli et al. (n.d.) describe the data as having 1000 records and 30 variables, each record corresponding to a prior credit application and is labeled either as “good credit” (700 cases) or “bad credit” (300 cases). I will use histograms to display the distributions of all numerical variables. The section titled *Distribution Plots: Boxplots and Histograms* in our text of Section 3.3 will help in coding these visualizations. I will use side-by-side boxplots to evaluate the potential of numerical predictor variables. Pairs most separated represent potentially useful predictors. Figure 3.3 in Section 3.3 of our text will be of assistance when coding these visualizations. Figure 3.4 in Section 3.3 of our text will aid in visualizing correlation tables via heatmaps.

After performing EDA on the ***GermanCredit.csv*** file, the student will divide the data into training and validation sets. The student will use logistic regression, classification trees, and neural networks to assess classification models and will describe the two most accurate classification techniques. Chapters 9, 10, and 11 of our text are dedicated to classification trees, logistic regression, and neural networks, respectively. Useful metrics and visualizations for determining model performance include lift charts, decile-wise lift charts, confusion matrices, ROC (Receiver Operating Characteristic) curves, and the AUC (Area Under Curve) metric. Section 5.2 of our text, titled *Evaluating Predictive Performance,* covers predictive accuracy measures and visualizations. The final document will include appropriate R comments and R Markdown headers and will be output to Microsoft Word using the *Knit to Word* function within RStudio. I do not have any questions regarding the assignment currently.

References

Module 8: Portfolio Project. (n.d.). Retrieved March 7, 2020, from https://csuglobal.instructure.com/courses/18808/assignments/382129?module\_item\_id=919911

Shmueli, G., Bruce, P. C., Yahav, I., Patel, N. R., Lichtendahl, K. C., & Jr. (n.d.). Data Mining for Business Analytics. Retrieved from <https://platform.virdocs.com/r/s/0/doc/503437/sp/21743566/mi/74416482?cfi=%2F4%2F2%2F10%2F20%2F12%2F4%2F9%3A408&menu=search&q=area%20under%20curve>