**TECHNICAL APPENDIX**

Zonghao Li (zl2613)

**1. Data Preparation:**

**a**. We first saved the data as a ‘csv’ file and cleaned the data by replacing missing values by ‘NaN’ and assigning dummy variables to the dependent variable column. Then we dropped columns and rows that contain all missing values and also columns that contain the same value for all companies like ‘Company Type’.

**b**. If a company doesn’t have value in any of the 9 financial metrics, this company would be dropped. But we found out that only 438/448 companies have record in ‘Free Operating Cash Flow’/‘Current Ratio’, so we filled missing value with column mean for this two columns in order to keep these two categories and not deleting too many records.

**c**. For each company, we assign the mean of the available records in each category to missing values within that category.

**2. Training and Out-of-Sample Datasets:**

After cleaning, we randomly separated the remaining 2509 companies into equal sized training sample and hold-out sample.

**3. Modeling:**

**a.** For logistic regression, we used all remaining 70 columns and treated them as independent to predict whether the company would default.

**b**. For decision tree, we set tree depth to 8, minimum units in each non-terminal node to 20, and R2 minimal improvement per split to 1% for the library algorithm.

**4. Evaluation:**

The test error rate for logistic regression is 0.125920 and the test error rate for decision tree model is 0.121014. So both models performed well and decision tree model was favored in this problem. For further analysis, we could collect more data to build a more accurate model since the huge amount of missing data in our dataset limited the performance of the two algorithms.