Feedback on the Project Proposal of Jack Wong and Daniel Fu

In this project, the authors propose to analyze financial market data to assess the likelihood that an individual company outperforms the S&P 500 ETF. The selected dataset provides an excellent example for demonstrating the modeling techniques covered in this course. The exploratory data analysis provides insightful guidance for model building. For frequentist methods, the authors propose using logistic regression, which is appropriate and well suited to the problem setting.

Please refer to my itemized comments below, and address these points carefully in your revision. When submitting the revised proposal, please include a cover letter that addresses each of these points item by item and explains how they have been incorporated into the revision.

(1) On page 2, Figure 1, "annual stock return" is a derived variable. Please clearly describe how this variable is defined and calculated.

(2) On page 2, Figure 2 shows the fraction of years in which each firm outperforms SPY. Please clarify how the bins on the x-axis (outperformance rate) were chosen. Given that the dataset spans only ten years, should the outperformance rate take values only in $\{0, 0.1, 0.2, \ldots, 0.9, 1\}$?

(3) On page 3, Figure 3 shows the relationship between the "Fraction of outperformance" and "Weekly volatility." Since the response variable is defined based on annual returns, please clarify why volatility is measured at the weekly level. How is weekly volatility defined, and how does it align with the annual scale of the outcome variable?

(4) On page 5, Figures 6 and 7 use averaged values of dividend yield and P/E ratio. Averaging these variables over time may attenuate the relationships of interest. It would be informative to instead present side-by-side boxplots, where one boxplot summarizes the (non-averaged) dividend yield for company-years in which the return beats SPY, and the other summarizes the dividend yield for company-years in which the return does not beat SPY.

(5) The relationship between the outperformance rate and the predictors considered here may vary depending on other factors. For example, the firm's GICS sector may play an important role. You may gain additional insights by repeating the exploratory data analysis stratified by sector.

(6) On page 6 (Section 3.1), Models 1 and 2 are effectively specifications of the same logistic regression model and should be combined into a single model description. In addition, the indicator variable is currently denoted as Y; please include appropriate subscripts to index firm and year, for example Y_{it}. Correspondingly, the Bernoulli success probability should be denoted as p_{it}. Finally, since the P/E ratio and dividend yield vary by firm and year, the logistic regression model should be specified as

$$\text{logit}(p_{it}) = \beta_0 + \beta_1 \log(PE_{it}) + \beta_2 \, \text{DivYield}_{it}.$$

(7) It would be interesting to incorporate company-specific random coefficients, potentially stratified by sector. For the next submission, I would like to see separate logistic regression

models fitted within each sector, along with a comparison of their performance against the overall logistic regression model that ignores sector information. Please also consider fitting a mixed-effects logistic regression model (e.g., with firm-level random intercepts and/or slopes, possibly allowing these effects to vary by sector).

(8) Feedback regarding model fitting and diagnosis will be provided in the next round of review.