Abstract/Summary (3-5 pages)

## Missing data

When first looking at the data it became apparent rather quickly that missing data would require a lot of consideration – since there is a considerable number of NAs in the data. However, this is usually the case with ‘real’ data so it didn’t come as a surprise.

Therefore, we spent some time researching how to tackle this issue. Our research came up with some papers, as well as blog entries surrounding the topic from which we took inspiration – some of which are listed in the appendix.

Albeit some of the proposed methods were too involved for our understanding and the scope of this project – as a conclusion of our research regarding this topic we’ve come up with some methods which seem to be suited for our problem.

Some basic methods include mean, median or most frequent value (mode) but these approaches seemed to be rather simplistic for our purposes. Therefore, we settled for more elaborate measures, such as KNN, MissForest and iterative imputer, the latter of which is still experimental according to the documentation page.

Before we got started with implementing these methods, we conducted some EDA to see what we were dealing with. One thing that caught our eye was the fact that quite some columns had a ton of NAs. Since imputing, with whatever method, does not seem reasonable for such factors we had to decide which ones we wanted to drop. Eventually, we settled for a benchmark of 1’500 NAs per column (out of roughly 4000-5000 values per year) and year.

After finding these high-NA variables we wanted to check whether some of them could be useful to assign stocks to one of the predefined labels (buy/hold/sell). We achieved this by creating a classification tree.

Since at that point of time we were not actually going to make a prediction based on the results, but were only concerned with finding factors that may be ‘important’, we used the predefined variable ‘Class’ as our response variable.

From this we got eight variables which were used in a decision tree for any of the years, which had a high NA-count. We then looked at each of them individually to determine whether to drop them or not. For this decision we took into account the difference between the actual NA-count and the predefined threshold. This, of course, was based on highly subjective judgement.

What is more, we have found that there were also a number of rows which had a rather high NA-count, which can be seen in the missing value matrix at the beginning of chapter: “Missing Values/ NAs”. For this we’ve used a threshold of 150 missing values per row, which accounted for approximately 10% of all firms (where a company is counted twice if it appears in two years, three times if it appears thrice, etc.). We then went on and also dropped these rows out of our dataset. We were able to see a clear improvement regarding the total NA-count (this is made visible in a graphic in the subsequent code-chunk).

## Zeroes

Another problem that had to be addressed were the zero-values found throughout the data set. This seemed to be a rather delicate matter, since we were and still cannot be sure whether these zero-values represent actual zeroes, NAs or something else which we weren’t considering.

After some plotting and further exploration of the data it showed that a lot of zeroes were connected to either dividends, R&D or things like Revenue from discontinued Operations, short or long term investments which are metrics that are not entirely unlikely to be zeroes if taken in an economical sense (see appendix for a graph).

Eventually, we decided to leave the zero-values untouched, by the reasoning that we would possibly do more harm than good by changing these values. Another point in favour of this course of action one might add is that, according to our research, changing (or even imputing) values in financial data is avoided as much as possible.

## Outliers

A next step in approaching our main task was the handling of outliers, which we met with a method that is widely used according to scientific literature: the IQR-approach. As an alternative we’ve looked at the Winsorization method to deal with the outliers which is very close to the IQR-approach, given the fact that both use percentiles and set abnormally large or small values on predefined boundaries

Some of the problems that we’ve encountered for this part of the project are firstly, the fact that a lot of methods assume normality which is violated for the given data as we’ve checked with the Henze-Zirkler-test. Secondly a lot of the literature is concerned with univariate data which is obviously not given either so finding methods that can be applied for multivariate data was rather challenging.

Again, one might add that this might bring more misinformation than good to the table so another way would be to simply leave the outliers alone. As one user has mentioned on the Kaggle webpage some values e.g., for revenue have not been converted to USD but rather left in some currency which has a multiplier of 100 such as the Japanese Yen. If one were to be extremely invested one could perform an extensive EDA and correct ‘obvious’ mistakes. Since we’ve got a rather tight schedule, we refrain from such actions.

## Decision Tree/Random Forest

decision tree test accuracy with grid search for unbalanced data: 0.6031  
decision tree test accuracy with grid search for balanced data: 0.6729

randomized parameter search:

for the unbalanced data we have 0.6390 (randomized search) vs. 0.6262 (no parameter specification)

for the balanced data we have 0.7548 (randomized search) vs. 0.7532 (no parameter specification)

## Difficulties and Problems

* One of the main problems that we have encountered throughout the project was the lacking processing power, i.e., at times we have spent quite some time waiting for python to execute code. This was especially a problem for the imputation of missing values and fitting of models (? – Lorena)
* Although we all had some experience in coding, the adaption to Jupyter Notebooks needed some time, especially for the handling of import statements, which sometimes had to be imported via the terminal first before we were able to access it in the Notebook
* We frequently used the documentation of different tools and functions to get a more in-depth understanding of functions and their respective parameters, this however, often was quite complex.
* One problem that did not have a big impact but was still somewhat annoying is the labelling of plots. It seemed as though sometimes the same settings gave different results with respect to the size of the axis labels.

## Further Actions

Iff we had more resources (mainly time and processing power) some further steps that we would take are the following:

* We have created different datasets, one for each different imputing method, and an additional one for the Winsorization method (see chapter on outliers), but we’ve only fitted our models on the dataset obtained via KNN imputer, and the IQR-method, as a future step we’d also fit the models on the other datasets (i.e., other imputing and outlier-handling methods)
* Create new features, complementing our inflation and momentum terms
* Rest the significance of added variables such as e.g., momentum or the added interaction terms
* Check the validity of the models with more/ different performance metrics

Appendix:

**References**

Armina, R., Mohd Zain, A., Ali, N. A., & Sallehuddin, R. (2017). A Review On Missing Value Estimation Using Imputation Algorithm. *Journal of Physics: Conference Series*, *892*, 12004. https://doi.org/10.1088/1742-6596/892/1/012004

Gautam, C., & Ravi, V. (2015). Data imputation via evolutionary computation, clustering and a neural network. *Neurocomputing*, *156*(4), 134–142. https://doi.org/10.1016/j.neucom.2014.12.073

John, C., Ekpenyong, E. J., & Nworu, C. C. (2019). Imputation of Missing Values in Economic and Financial Time Series Data Using Five Principal Component Analysis (PCA) Approaches. *Central Bank of Nigeria Journal of Applied Statistics*(Vol. 10 No. 1), 51–73. https://doi.org/10.33429/Cjas.10119.3/6

Ping, X.-O., Lai, F., & Tseng, Y.-J. (Eds.). (2014). *Evaluation of Imputation Methods for Missing Data and Their Effect on the Reliability of Predictive Models*. IARIA. http://www.thinkmind.org/index.php?view=instance&instance=BIOTECHNO+2014

Schmidt, P., Mandel, J., & Guedj, M. A Comparison of six Methods for Missing Data Imputation. *Journal of Biometrics and Biostatistics*, *6*(224), 1–6.

*Scikit Learn - Imputation of Missing Values*. https://scikit-learn.org/stable/modules/impute.html

*Tutorial on 5 Powerful R Packages used for imputing missing values*. https://www.analyticsvidhya.com/blog/2016/03/tutorial-powerful-packages-imputing-missing-values/

Graph showing the variables with the highest percentage of zero values

