Predicting a Company's Exit using Data Science

van Zelst, Marco C.

Award date:
2020

Predicting a Company's Exit using Data Science

Marco C. van Zelst, 1226238 (m.c.v.zelst@student.tue.nl)
Eindhoven University of Technology
Department of Mathematics and Computer Science
5612 AZ Eindhoven, The Netherlands

Abstract

On average, 100 million start-ups are founded each year worldwide (Bosma et al., 2020). Some reach a net worth of millions while other start-ups fail after a few months. However, all start-ups eventually end their 'start-up' phase and perform a so-called exit. This exit differs between start-ups and cannot be known beforehand. Using a data set with ex post data on the exits of start-ups, the influencing exit factors are predicted. This was done by creating a random forest and a logistic regression to see what could influence the exit decision. However, both internal and external effects influence the decision to go for an exit. External effects were not included within this data set. Future research can try to incorporate this external effect into the model and perform this research on a larger scale.

Keywords: Exit; Start-ups; Random Forest; Logistic regression

Introduction

A company's exit to the market can be regarded as the endphase of the start-up and is sometimes seen as the ultimate goal of building a venture (Pisoni & Onetti, 2018). However, there are different types of exits. Literature distinguishes five types of exits (Da Rin & Hellman, 2020). These exit types are: an initial public offering (IPO), a merger and acquisition (M&A), a secondary sale, a redemption and a write-off. An IPO means that the company goes public with their shares. A M&A means that another company buys all shares. A secondary sale means shares of a stockholder outside of the company are sold without interference of the company. Redemption means that the shares are bought back by the company from the shareholders, usually after five years with interest. A write-off means that the made costs are seen as a loss and the company stops existing. These exit types are accompanied by different benefits and drawbacks. These benefits and drawbacks can be found in Table 1 in Appendix 1.

Exits are important since they are critical to understanding the entrepreneurial process as a whole (DeTienne, 2010). The certainty and impact of an exit influences the entrepreneur since he should prepare for an exit with an appropriate exit plan (Dahl, 2005). Besides this, the economy is also influenced by companies undergoing an exit (Van Ewijk, 1997). However, a significant amount of research has been put into other parts of the entrepreneurial process (DeTienne, 2010). For example, team formation (Chowdhury & Sanjib, 2005), (Forbes, Borchert, Zellmer-Bruhn, & Sapienza, 2006), how ideas are generated and improved by entrepreneurs (Gemmell, Boland, & Kolb, 2012) and how investors can be attracted (Busenitz, Fiet, & Moesel, 2005), (Clark, 2008). As said, less attention has been given to the exit processes. "The lack of research on entrepreneurial 'exits' is striking when compared with the attention that has been given to entrepreneurial start-ups" (Mason & Harrison, 2006).

Research that can be done regarding the exits of companies is for example looking at the impact of an exit to the firm's financial situation (Shen & Cannella, 2002) or to predict factors which can influence the exit of a start-up. For this research, the focus will be on factors influencing the decision for a certain type of exit. These factors are both internal and external, which means that over some factors founders and investors do have control and over some they do not. An example of such an external effect are developments in the market (Blum, 2015).

Literature review

Previous research

Research regarding predicting the influencing factors for an exit of a start-up has been done before. In 2003, researchers demonstrated that the firm's industry, its age, its stage at the time of financing and the valuation at times of financing influenced the probability of an exit (Das, Jagannathan, & Sarin, 2003). The firm's industry heavily influences the probability of a type of exit, for example while an IPO is fairly common for the computer industry it was almost uncommon for the instruction industry. The probability of a company exiting through an IPO showed higher for older companies. The valuation and stage of the firm at the time of financing also influenced the exit decision where an higher valuation and a later stage at the time of financing would increase the chance for an IPO.

Other research found that the identity of the first investor is an important factor in classifying exits (Bhat & Zaelit, 2011). The reason for this is that besides monetary value some investors, like venture capital also have an important monitoring and mentoring role to the companies they finance which could eventually lead to more profitable exits like an IPO (Hellmann & Puri, 2000).

Human capital of a firm and the fear of failure also differ across exit routes (Wennberg, Wiklund, DeTienne, & Cardon, 2010). Human capital is described as the entrepreneur's experience, age and level of education. Failure-avoidance is described as entrepreneurs having for example a job besides their start-up or whether they put the revenue from the start-up back into the start-up. Having a side job or not reinvesting revenue lowers the risk for the entrepreneur but also lowers the probability of a successful exit.

Previous research hypothesized that candidates for a M&A are less homogeneous than those for an IPO (Giot &

Schwienbacher, 2007). This means that companies performing an IPO are generally more similar to each other compared to companies undergoing M&A's. They also mention the existence the external market conditions having an influence on the type of exit, meaning that the status of the market in which the company is active partially determines the type of exit. Other research by Francis et al. also pointed out that external factor. They described that firm fundamental factors influence the decision between an IPO and a M&A as well. These fundamental factors are the age, size and liquidity of a company (Francis, Hasan, & Siregar, 2009). However, the data set for the previously mentioned study contained information on banks and not on different types of companies. Other research that did include different types of companies found that four factors play a role in the choice between IPO and M&A (Brau, Francis, & Kohers, 2003). These four factors are industry, market-timing, deal-specific, and to a lesser degree demand for funds. Industry is related to what industry the company is active in alongside with the average debt for that industry. Market-timing is an external factor relating to overall trends within the market. For this research they have looked at whether performing an IPO at that moment in time is popular or not. Or if for example M&A's are more popular, which has been recorded in the past (Mitchell & Mulherin, 1996). Deal-specific factors relates to the transaction costs related to the exit type and the ownership intentions of the entrepreneur. An IPO for example may put off smaller firms due to the high amount of costs involved with that transaction (Ritter, 1987). Furthermore, if the owner wishes to maintain ownership after the exit, an IPO is for example preferable (Bebchuk & Zingales, 2000). Demand for funds means how much funding is wanted by the firm. If this is high, an IPO is more likely to be able to provide these funds. However, no full correlation between demand for funds and exit types was found.

This research

This research consists of two parts which both look at internal factors related to possible exit types.

Part 1 The first part tries to build upon the research of (Das et al., 2003) and (?, ?). This part tries to predict the factors that can influence the decision to pursue a certain exit. Leading to the following research question:

- Research question 1:
  What factors influence the exit decision of a company?

I hypothesize that the sector of the economy will help to predict a company's exit since the sector a company is active in relates to their industry who Das et al. prove to be an indicator for different exit types. Furthermore, the influence of age and valuation at time of investing will also be looked at which according to the same study could influence the type of exit. Bhat et al. described the identity of the first investor as an important factor, to extend upon this research the last investor type will be looked at. This means that the type of investor which invested the last time before the company's exit will be investigated for its influence on the exit. This is hypothesized to also have an effect on the exit type since they can also provide a monitoring and or mentoring role just like the first investor. Most researches are done on a national scale, to be able to extrapolate the findings the founding location of a company will be looked at to see whether this could influence the exit decision. This might influence the exit type because the markets differ across nations so its hypothesized that this will influence the exit decision for a company. Unfortunately, data on the entrepreneur like human capital and failure avoidance was not included in the data and therefore the research of Wennberg et al. cannot be extended.

Part 2 The second part of this research zooms in upon the most interesting and statistically different categories IPO and M&A (Smith, Pedace, & Sathe, 2011). It tries to see if the fundamental factors (age, size and liquidity) do really influence the decision between an IPO and a M&A and to what extend the four factors (industry, market-timing, deal-specific and demand for funds) from Brau et al. influence the decision between an IPO and a M&A. Leading to the following research question:

- Research question 2:
  What factors differ for companies pursuing either an IPO or an M&A?

I hypothesize that the firm fundamental factors can be extrapolated to multiple industries meaning that these factors will help predicting IPO's and M&A's. Furthermore, the influence of the four factors from Brau et al. are expected to hold for the data set since it looks at the same problem. However, these four factors are substantiated from different variables in this study than for that original study meaning some differences in effects can occur.

Methods

Data

The data used for the analysis was retrieved from the Pro version of Crunchbase Queries (*Crunchbase Queries*, 2020). To get a sample from the data population of Crunchbase, several parameters were used to query on the data. To simplify the analysis only active, profitable companies with one exit where chosen since predicting multiple exits would make this process more difficult. These are the parameters of the query.

When performing the query with the said parameters the data was received. The data consisted of 399 companies with 46 variables. Each company in the data set is called an instance. After collecting the data, pre-processing was done to improve the quality of the data. Pre-processing included adding missing data with open source material if possible and increasing coherence within the data, which included for example getting date data in the same format.

Variables

After pre-processing the data the variables within the data set were examined. Appendix 2 contains plots on the variables to see their division in the data set. Appendix 3 takes a closer look at what of these variables can, in theory, affect the type of exit. When visualizing the variables, a lot of variables were unusable due to them being similar to other variables or having no explanatory power. These obsolete variables were removed, reducing the number of variables in the final data set to twenty. Appendix 4 shows what variables were removed and which were kept with an explanation on each variable.

After removing the obsolete variables in the data set, there were still missing values within the data set. This can be solved in four possible ways: the rows with missing values can be deleted, the missing values can be predicted using estimation techniques, the missing values can be replaced by the mean or the missing values can be replaced by minus infinity (Graham, 2009). The latter three options were not possible for this data set since some variables were other types than integers, like strings or categories. Therefore, the rows with missing values were deleted. So only complete instances within the data set were kept. After this, 256 instances were remaining with 20 variables.

The remaining variables were the organizational name, the total amount of funding received in United States Dollars (USD), the number of employees working for the company, the exit of the company, the amount of funding received in the last funding round of the company, the type of funding received in that last funding round, the total number of funding rounds, the number of investors, the location of the company, the age of the company at exit, the total number of investments, the number of website visits on average per month, the number of granted patents, and the sector in which the company was active.

As stated in the introduction, there are five categories of exits (Da Rin & Hellman, 2020). However, for the used data set the exit variable was divided into four categories instead of five predefined categories since write-offs were not included in the data set because Crunchbase does not collect this data. This is partly because this data is hard to gather since not all start-ups are registered and if non-registered start-ups fail it is not seen back in the data (Williams & Kedir, 2016). Furthermore, predicting start-up failure is another research area (Cantamessa, Gatteschi, Perboli, & Rosano, 2018), which is outside of the scope for this research. Thus, the remaining exit categories are an IPO, M&A, a secondary sale and redemption.

The founding location of the company location was divided into three categories. These locations of the companies were either in the United States, in the European Union or in the rest of the world. It was chosen to group the data by these locations since there was too little data to not group it. The division was made like this since the categories are approximately the same size and previous, similar research also looked at the differences between the United States and the European Union. This drew the conclusion that distinguishing those categories can be done since they approach the exits of start-ups differently (Pisoni & Onetti, 2018).

The age of the company at exit was calculated by subtracting the founding date from the date of exit. After doing this, seven instances of the age were negative meaning the date of exit was before the date of founding in the data set. Because this is faulty data these instances were removed to increase the accuracy of the data set.

The sector of the company was divided by the normal sectors of the economy. These are the primary sector which mainly concerns raw materials, the secondary sector which focuses on production and the tertiary sector which focuses on intangible goods and services to consumers (Wolfe, 1955). The data set only contained companies which are active in either the secondary or the tertiary sector.

Research Methods

To predict a company's exit, the exit has to be the independent y variable. The other data was used as the dependent variables. The chosen prediction model to predict a company's exit is a random forest. To make the random forest the data was divided in a training set and a validation set. The data was split randomly, with for each data point 25 % chance on being in the validation set and 75 % chance on being in the training set. The training data consisted of 192 instances and the validation data consisted of 64 instances.

To make this random forest, ten decision trees were made using the train data set and they were combined into one random forest. For this the CART (Classification And Regression Trees) algorithm was used. This algorithm makes use of a binary tree (Wu et al., 2008). Therefore, all rules within the decision tree will have a binary split. The algorithm looks at every node which rule makes the next nodes less impure. The Gini index was used to see the impurity of a node. This index goes from 0 to almost 1. If the index is 0, the node is pure and there is just one class in the node. The Gini index is calculated as follows:

$$G(N) = 1 - \sum_{j} [p(j|N)]^2 \tag{1}$$

N equals all the data points in the node, j is the classification and p is the probability of the classification (Parrillo & Volscho, 2012). So for each classification you calculate the squared probability that a single data point within the node has this classification. After this, the sum of these squared probabilities is extracted from one. So for every possible rule the average Gini index of the next two nodes gets calculated. The rule with the lowest Gini index is used as the leading rule. This CART algorithm does this for every rule within the decision tree. This ensures that the most important rules come first. However, you need different decision trees to combine them into one random forest. Therefore, it is impossible to make all decision trees on the same train and test data set since this yields equal decision trees. To prevent this, bootstrapping was used. This algorithm picks random data

points with replacement out of the training data. After this, the CART algorithm uses this new data set to train a decision tree. This happens 10 times. Now 10 different decision trees are created, but only 1 classification per new data point is required. To combine the decision trees an ensemble method was used. All 10 trees will classify this new data point and the majority of these different trees decide the classification of the data point. After training the random forest, the test data was used to test the random forest. A confusion matrix was used to see the classification outcomes (Figure 2).

The big downside of using a decision tree is that it easily overfits the data. To prevent overfitting two measures were taken. First multiple (10) decision trees were made and combined into one random forest. Secondly, pruning was used. Pruning means that the size of a decision tree gets limited by removing irrelevant sections (Ren, Cao, Wei, & Sun, 2015). The maximum size of each decision tree was ten.

After the random forest a logistic regression was made to view IPO and M&A in a more detailed way. To create a logistic regression all five of its assumptions must be met (Peng, Lee, & Ingersoll, 2002). The dependent variable exit must be binary, the observations must be independent of each other, there is no multicollinearity among the independent variables and the independent variables are linearly related to the log odds. These assumptions hold for the data set. The last assumption regarding the sample size holds but barely. Following the Green's rule of thumb to estimate the sample size an estimated sample size of 300 data points would suffice. (Wilson Van Voorhis & Morgan, 2007).

To increase the sample size a SMOTE algorithm was used (Synthetic Minority Oversampling Technique) (Chawla, Bowyer, Hall, & Kegelmeyer, 2002). This creates new samples from the data and uses k-nearest-neighbors to create similar but new observations. K-nearest-neighbors is computed as follows:

$$\sqrt{\sum_{i=1}^{n}(Q_i - P_i)^2} \tag{2}$$

Where n equals the number of dimensions the formula takes; in this case the variables put into the model. P_i subtracted from Q_i describes the Euclidean distance between two data points. The data point which is located at the minimum distance from the test point is assumed to belong to the same class. (Cunningham & Delany, 2007).

To optimize the parameters for each variable, a Recursive Feature Elimination (RFE) was used which constructs models and choose the best features over and over again (Zhu & Hastie, 2004). This process repeats itself until all features in the data set are used. Then the significance level of each variable gets calculated and removed if necessary. To evaluate the logistic model the accuracy was tested and a confusion matrix (Figure 4) and ROC curve (Figure 6) were created.

Eventually these two analyses can answer both research questions respectively since they show on what factors the exit decisions were based.

Results

A random forest consisting of 10 decision trees was created. To explain the working of such a tree a zoomed in part of such a decision tree is visualized in Figure 1. This decision tree is fully plotted in Appendix 6. The tree begins at the top, at the root node. Each node has a rule, a Gini index, a total amount of samples it contains (using the data set that is created during the bootstrapping), how many data points there are for every class (using the original training data), and the largest class in this node. The colour of the nodes have the following meaning: orange means that the exit gets classified as an IPO, green as a M&A, blue as a secondary sale and purple as redemption. Every decision node has two edges (the lines between the nodes). Every edge that goes to the left means that the rule of the decision node holds for the data points that follow this path, and if the data points follow the path to the right then it means that the rule does not hold for these points. For Figure 1, the split at the root node means that if there are more than one investment you will go to the right and for less than one investment you will go to the left of the decision tree.

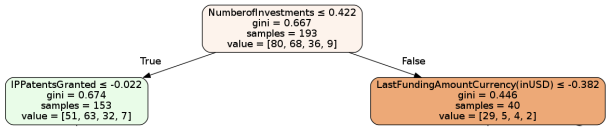


Figure 1: One decision tree used for the random forest, the full decision tree is shown in Appendix 6

Figure 2 shows the confusion matrix of the classification model. This matrix shows that most instances get predicted as IPO and M&A. It is also the most accurate there however also the most data points are in those categories. The entire classification process has an accuracy of 46.2 %.

Predicted Class

Actual Class	IPO	M&A	SecondarySale	Redemption
IPO	14	10	5	0
M&A	6	13	3	1
SecondarySale	3	4	3	0
Redemption	2	1	0	0

Figure 2: Confusion matrix of the predictions of the random forest

Figure 3 visualizes the importance of the dependent variables. The sum of all important values equals 1, this means that you can see the importance value as a proportion. This

means that the amount of funding received in the last round and the number of web visits each classify around 25 % of the data, therefore, those variables are the most important attributes in the classification model.

The hypothesized effects of sector and location where found to help predicting the exits but did not have a major influence in the decision process itself with an importance of 9.3 % and 2.8 % respectively. The age of the company had no effect when making the decision rules just as the identity of the last investor. The valuation at the time of investing was omitted from the random forest due to its high correlation with the amount of funding received in the last round which did have a lot of predicting power for the type of exit.

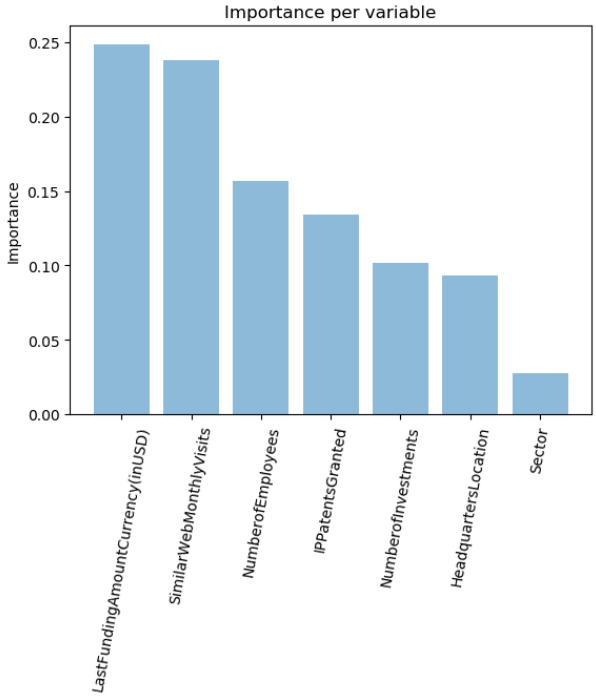


Figure 3: Importance of variables

The logistic regression wanted to view the effects of firm fundamental factors (age, size and liquidity) on the differences between IPO and M&A as well as the effects of industry, market-timing, deal-specific and demand for funds. The logistic regression was made for the following variables age at exit, total amount of funding to look at the effect of the firm fundamental factor size and the demand for funds and sector to look at the effect of industry. There was no data in the data set to look at the effects of market-timing and deal specific factors.

Figure 4 shows the confusion matrix of the classification model of the logistic regression model. The entire classification process has an accuracy of 62.2 %, predicting 28 out of the 45 data points correct in the test data set. This means that while the model has some predicting power it is not far off from random guessing (50 %). The errors mostly consist of classifying M&A's as IPO's.

	Predicted Class		
Actual Class		IPO	M&A
	IPO	14	5
	M&A	12	14

Figure 4: Confusion matrix of the predictions of the logistic regression

Figure 5 shows the sensitivity and specificity of the predictions. The low value for sensitivity means that it often defines IPO's which should have been predicted as M&A's. The higher value for specificity means that predicting M&A's is done better.

Accuracy	62.2 %
Sensitivity	53.9 %
Specificity	73.7 %

Figure 5: Sensitivity and specificity per class

Figure 6 shows the ROC curve of the classification. The macro average ROC has an area under the curve of 0.64. This value ranges from 0.5 until 1 where the closer to 1 the better it can distinguish between the two types of exits (Fawcett, 2006). A value of 0.64 means that the model is not able to properly distinguish between the exit category IPO and M&A.

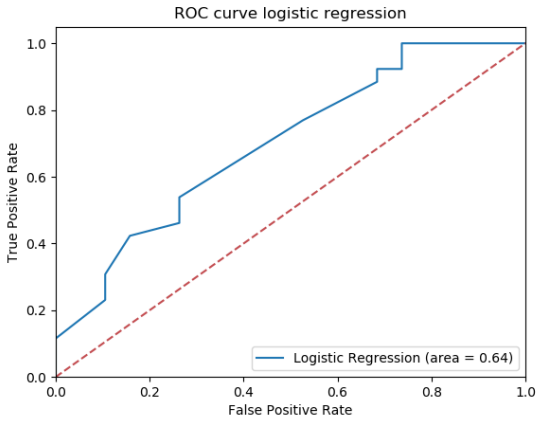


Figure 6: ROC Curve

Conclusion

Is it possible to predict a company's exit based on some of its characteristics? Creating a random forest based on data gathered from Crunchbase with information on companies an accuracy of 46.2 % was achieved when predicting four possible exit types. Namely an IPO, an M&A, a secondary sale

and redemption. This prediction depends mostly upon the following variables last funding amount, average web visits per month, number of employees and number of IP patents granted while sector and location only had a small effect on the prediction. The following variables had no significant effect on predicting exits: identity of the last investor and age of the company. The identity of the last investor is especially interesting for this research since this was expected to have prediction power over the data set. A possible explanation of this insignificant prediction power could be that the last investor has not enough time to exert a monitoring and mentoring role or this has already been fulfilled by other investors.

The accuracy of the random forest is with 46.2 % not the best prediction model available. Especially compared to the research of Bhat et al. were also exits got predicted which received an accuracy of 75.0 %. However, for this research the choice was made to not increase the accuracy at the cost of overfitting the model since this would inherently diminish the trustworthiness the results.

To zoom in upon IPO's and M&A's a logistic regression was made. This logistic regression has an accuracy of 62.2 % and tried to look at whether firm fundamental factors would influence the model as well as two out of four variables brought forward by Brau et al.. These are the industry and demand for funds, which where slightly changed compared to the study of Brau et al. due to not having the same data as that study.

Overall the logistic model found that the factors location, total amount of funding and age at exit could statistically identify these exit types. Total amount of funding and age at exit refer to the firm fundamental factors hypothesized to have predicting power. Total amount of funding also matches the demand for funds from Brau et al.. While age did not seem to have an influence for the random forest it was significant for the logistic regression. The following variable was thought to be statistically significant as well but was not was sector which was used as a substitute of the industry variable from Brau et al.. A reason for this might be that sector is too generic for a company while differences between industries themselves can be found. To examine the reliability of the logistic regression a robustness check was performed. The model was robust meaning that small changes in the model would not change the coefficients of the model drastically.

Overall two models where created trying to predict exit type accordingly. This study contributes to the research on exits since it defines possible predictor values for such exits. This could eventually help entrepreneurs and managers to prepare for an exit based on what they know about their firm (Dahl, 2005).

Discussion

Limitations

Although the findings from this study contribute to the understanding of factors influencing the decision for an exit, it has several limitations.

Firstly, the accuracy of both models could have been improved. This could be done in by performing this research on a larger scale or by improving upon the modelling techniques. Performing this research on a larger scale could improve the accuracy because currently the data set is narrow with only 256 instances. Increasing the number of instances would result in making better use of the data without overfitting the data.

Another limitation is the fact that much data was missing in the gathered data set. Out of the 399 original companies only 256 were left after pre-processing the data. This resulted in a loss in data of 35.8 %. This could in theory introduce bias into the data set since such a large part of the data is removed (de Noord, 1994).

Another limitation is the fact that Crunchbase did not make it possible to gather more variables than they currently have in their database. This limited the research that could have been done since not all wanted variables could have been included. This includes both internal and external variables. Internal variables that could have been included could be for example time between funding rounds and amount of funding gathered per funding round which could have been used to make a time-series model. Other internal variables which would have been nice to include in the data set were the ones discussed in the literature review. These include the identity of the first investor, human capital of the firm displaying information about the founders themselves and firm fundamental factors like the liquidity of the company. External effects that could have been included within the data set are trends in the economy which enables looking at the effect of market timing and overall market developments over time.

Future Research

Future research can build and improve upon this research by implementing the limitations mentioned previously. This can be done by gathering the data differently in such a way that it allows for the said variables to be gathered. This would help to improve the knowledge of the consequences of external effects to the exit decision.

In addition, gathering data in a different way might lead to less loss of data as a result to data pre-processing which can eventually lead to introducing no bias into the data set.

References

Bebchuk, L., & Zingales, L. (2000). Ownership Structures and the Decision to Go Public: Private versus Social Optimality. *NATIONAL BUREAU OF ECONOMIC RESEARCH*.

Bhat, H. S., & Zaelit, D. (2011). Predicting Private Company Exits Using Qualitative Data.. doi: 10.1007/978-3-642-20841-6_33

Blum, D. A. (2015, 1). Factors Contributing To Independent Venture Capital Successful Exits. *Journal of Business & Economics Research (JBER)*, *13*(1), 1. doi: 10.19030/jber.v13i1.9074

Bosma, N., Hill, S., Ionescu-Somers, A., Kelley, D., Levie, J., & Tarnawa, A. (2020). *Global Entrepreneurship Monitor 2019/2020 Global Report AUTHORS*. Retrieved from http://www.witchwoodhouse.com

Brau, J. C., Francis, B., & Kohers, N. (2003, 10). The Choice of IPO versus Takeover: Empirical Evidence. *Journal of Business*, *76*(4), 583–612. doi: 10.1086/377032

Busenitz, L. W., Fiet, J. O., & Moesel, D. D. (2005). *Signaling in venture capitalist - New venture team funding decisions: Does it indicate long-term venture outcomes?* doi: 10.1111/j.1540-6520.2005.00066.x

Cantamessa, M., Gatteschi, V., Perboli, G., & Rosano, M. (2018). Startups' roads to failure. *Sustainability (Switzerland)*. doi: 10.3390/su10072346

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*. doi: 10.1613/jair.953

Chowdhury, & Sanjib. (2005). Demographic diversity for building an effective entrepreneurial team: is it important? *Journal of Business Venturing*, *20*(6), 727–746.

Clark, C. (2008). The impact of entrepreneurs' oral 'pitch' presentation skills on business angels' initial screening investment decisions. *Venture Capital*. doi: 10.1080/13691060802151945

Crunchbase Queries. (2020). Retrieved from https://www.crunchbase.com/organization/query

Cunningham, P., & Delany, S. J. (2007). K -Nearest Neighbour Classifiers. *Multiple Classifier Systems*. doi: 10.1016/S0031-3203(00)00099-6

Dahl, D. (2005). *A New Study Says Most Small Biz CEOs Lack Succession Plans — Inc.com.* Retrieved from https://www.inc.com/news/200502/exit.html

Da Rin, M., & Hellman, T. (2020). Fundamentals of Entrepreneurial Finance. In *Fundamentals of entrepreneurial finance* (p. 656).

Das, S. R., Jagannathan, M., & Sarin, A. (2003). *PRIVATE EQUITY RETURNS: AN EMPIRICAL EXAMINATION OF THE EXIT OF VENTURE-BACKED COMPANIES* (Vol. 1; Tech. Rep. No. 1). Retrieved from www.joim.com

de Noord, O. E. (1994). The influence of data preprocessing on the robustness and parsimony of multivariate calibration models. *Chemometrics and Intelligent Laboratory Systems*. doi: 10.1016/0169-7439(93)E0065-C

DeTienne, D. R. (2010, 3). Entrepreneurial exit as a critical component of the entrepreneurial process: Theoretical development. *Journal of Business Venturing*, *25*(2), 203–215. doi: 10.1016/j.jbusvent.2008.05.004

Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters*. doi: 10.1016/j.patrec.2005.10.010

Forbes, D. P., Borchert, P. S., Zellmer-Bruhn, M. E., & Sapienza, H. J. (2006). Entrepreneurial team formation: An exploration of new member addition. *Entrepreneurship: Theory and Practice*. doi: 10.1111/j.1540-6520.2006.00119.x

Francis, B., Hasan, I., & Siregar, D. (2009, 12). The choice of IPO versus M&A: evidence from banking industry. *Applied Financial Economics*, *19*(24), 1987–2007. doi: 10.1080/09603100903251262

Gemmell, R. M., Boland, R. J., & Kolb, D. A. (2012). The socio-cognitive dynamics of entrepreneurial ideation. *Entrepreneurship: Theory and Practice*. doi: 10.1111/j.1540-6520.2011.00486.x

Giot, P., & Schwienbacher, A. (2007, 3). IPOs, trade sales and liquidations: Modelling venture capital exits using survival analysis. *Journal of Banking and Finance*, *31*(3), 679–702. doi: 10.1016/j.jbankfin.2006.06.010

Graham, J. W. (2009). Missing Data Analysis: Making It Work in the Real World. *Annual Review of Psychology*. doi: 10.1146/annurev.psych.58.110405.085530

Hellmann, T., & Puri, M. (2000). The interaction between product market and financing strategy: The role of venture capital. *Review of Financial Studies*. doi: 10.1093/rfs/13.4.959

Mason, C. M., & Harrison, R. T. (2006, 2). After the exit: Acquisitions, entrepreneurial recycling and regional economic development. *Regional Studies*, *40*(1), 55–73. doi: 10.1080/00343400500450059

Mitchell, M. L., & Mulherin, J. H. (1996). The impact of industry shocks on takeover and restructuring activity. *Journal of Financial Economics*. doi: 10.1016/0304-405X(95)00860-H

Parrillo, V., & Volscho, T. W. (2012). Gini Coefficient. In *Encyclopedia of social problems.* doi: 10.4135/9781412963930.n237

Peng, C. Y. J., Lee, K. L., & Ingersoll, G. M. (2002). An introduction to logistic regression analysis and reporting. *Journal of Educational Research*. doi: 10.1080/00220670209598786

Pisoni, A., & Onetti, A. (2018). When startups exit: comparing strategies in Europe and the USA. *Journal of Business Strategy*. doi: 10.1108/JBS-02-2017-0022

Ren, S., Cao, X., Wei, Y., & Sun, J. (2015). Global refinement of random forest. In *Proceedings of the ieee computer society conference on computer vision and pattern recognition.* doi: 10.1109/CVPR.2015.7298672

Ritter, J. R. (1987). The costs of going public. *Journal of Financial Economics*. doi: 10.1016/0304-405X(87)90005-5

Shen, W., & Cannella, A. A. (2002, 8). Revisiting the Performance Consequences of CEO Succession: The Impacts of Successor Type, Postsuccession Senior Executive Turnover, and Departing CEO Tenure. *Academy of Management Journal*, *45*(4), 717–733. doi: 10.5465/3069306

Smith, R., Pedace, R., & Sathe, V. (2011, 12). VC Fund Financial Performance: The Relative Importance of IPO and M&A Exits and Exercise of Abandonment Options. *Financial Management*, *40*(4), 1029–1065. doi: 10.1111/j.1755-053X.2011.01170.x

Van Ewijk, C. (1997). Entry and exit, cycles, and produc-

tivity growth. *Oxford Economic Papers*. doi: 10.1093/oxfordjournals.oep.a028602

Wennberg, K., Wiklund, J., DeTienne, D. R., & Cardon, M. S. (2010, 7). Reconceptualizing entrepreneurial exit: Divergent exit routes and their drivers. *Journal of Business Venturing*, *25*(4), 361–375. doi: 10.1016/j.jbusvent.2009.01.001

Williams, C. C., & Kedir, A. M. (2016). BUSINESS REGISTRATION and FIRM PERFORMANCE: SOME LESSONS from INDIA. *Journal of Developmental Entrepreneurship*. doi: 10.1142/S1084946716500163

Wilson Van Voorhis, C. R., & Morgan, B. L. (2007). Understanding Power and Rules of Thumb for Determining Sample Sizes. *Tutorials in Quantitative Methods for Psychology*. doi: 10.20982/tqmp.03.2.p043

Wolfe, M. (1955, 8). The Concept of Economic Sectors. *The Quarterly Journal of Economics*, *69*(3), 402. doi: 10.2307/1885848

Wu, X., Kumar, V., Ross, Q. J., Ghosh, J., Yang, Q., Motoda, H., ... Steinberg, D. (2008). Top 10 algorithms in data mining. *Knowledge and Information Systems*. doi: 10.1007/s10115-007-0114-2

Zhu, J., & Hastie, T. (2004). Classification of gene microarrays by penalized logistic regression. *Biostatistics*. doi: 10.1093/biostatistics/kxg046

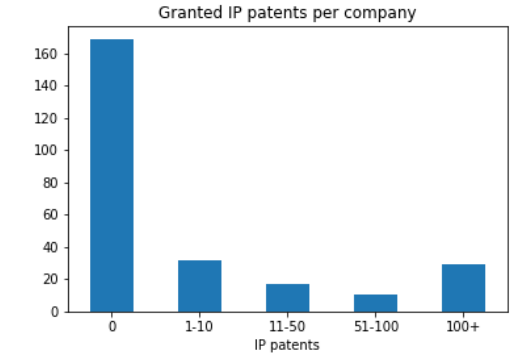
Appendices

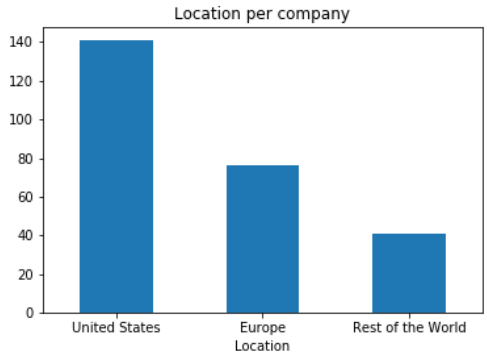
Appendix 1; Table 1.

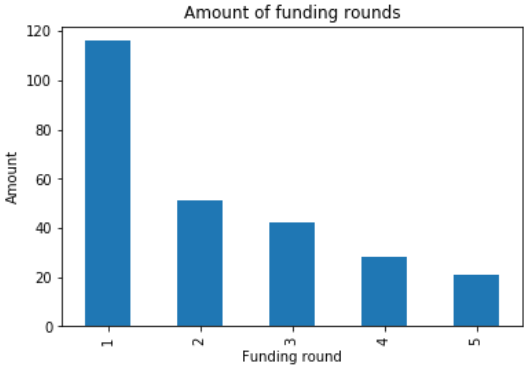
Exit:	Benefits:	Drawbacks:
IPO	Very profitable. Only happens to the best ventures. Gives reputation.	Only a partial exit. Remaining shares are subject to selling restrictions. High transaction costs. Time consuming.
M&A	Full exit, all shares are sold in one transaction. Lower costs than IPO.	Can give conflicts between founders. Lower value than IPO.
Secondary sale	Good for portfolio.	Lose share at future payouts.
Redemption	Get rights back.	Pay interest
Write-offs	No extra costs.	No chance to generate money.

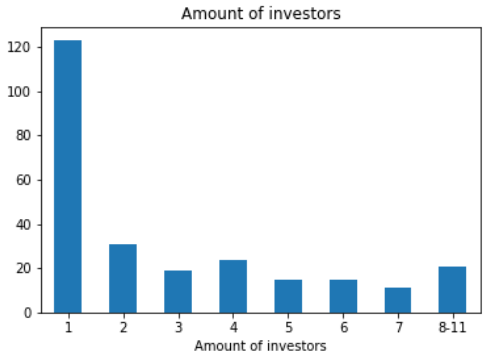
Table 1. Benefits and drawbacks for each exit type

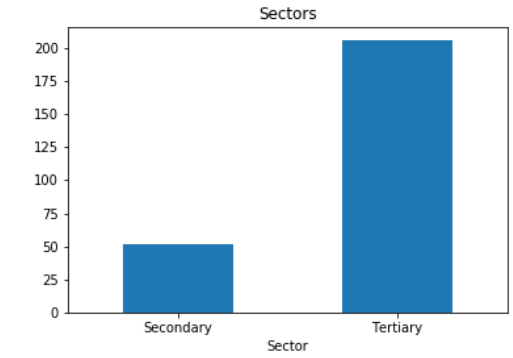
Appendix 2; Descriptive plots on the data

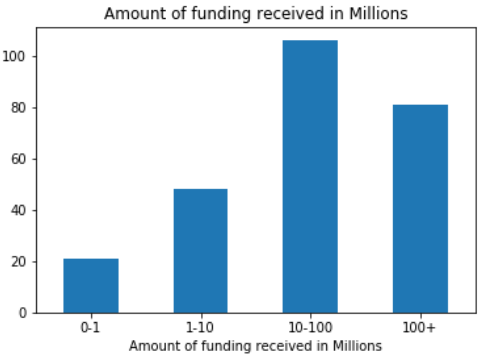


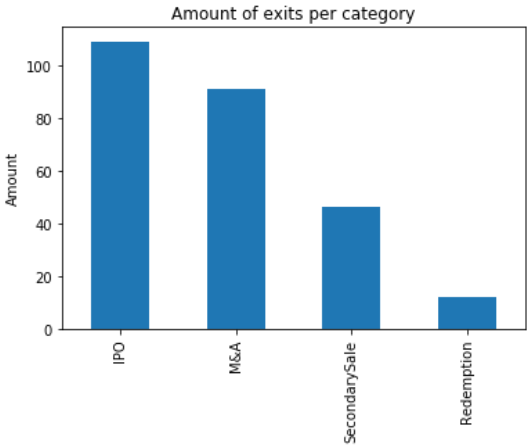






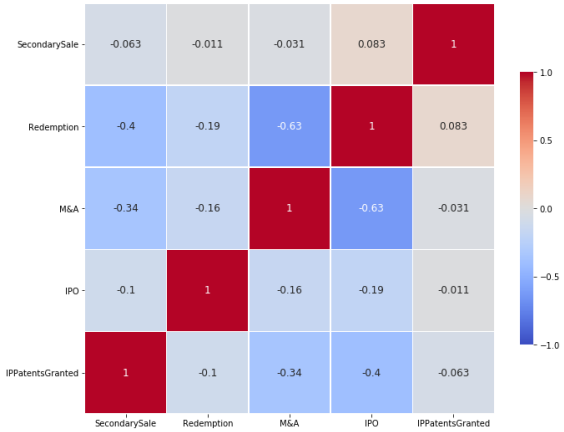




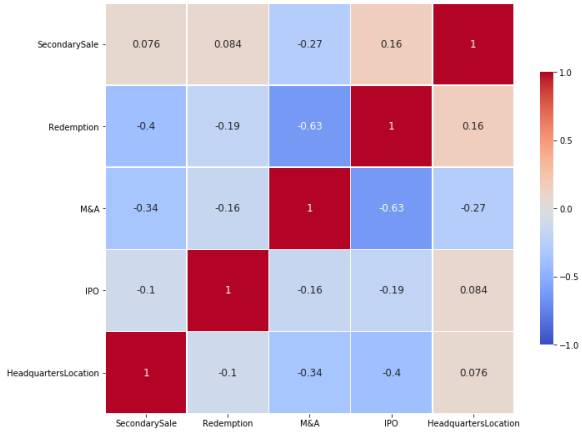


Appendix 3; Descriptive statistics on the data

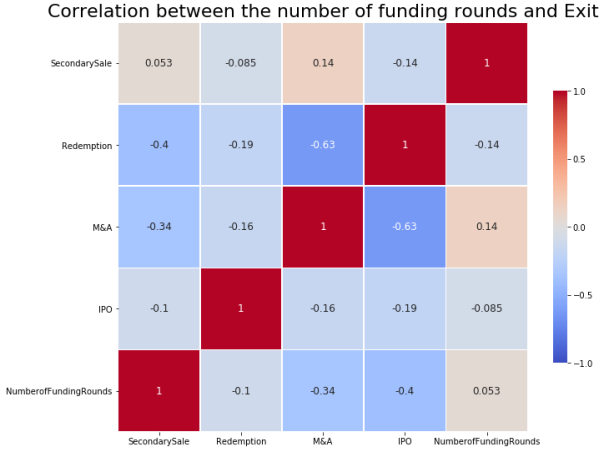
Correlation between Granted IP Patents and Exit



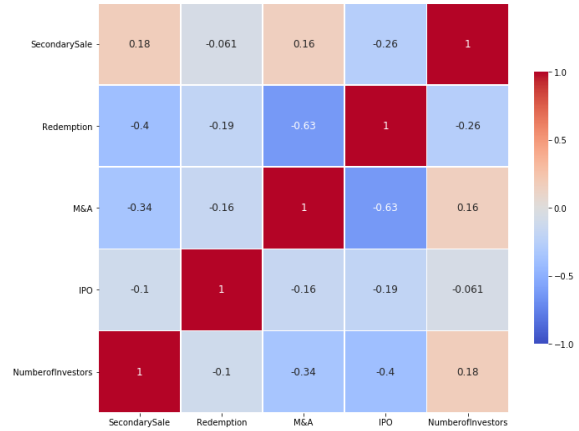
Correlation between the location and Exit



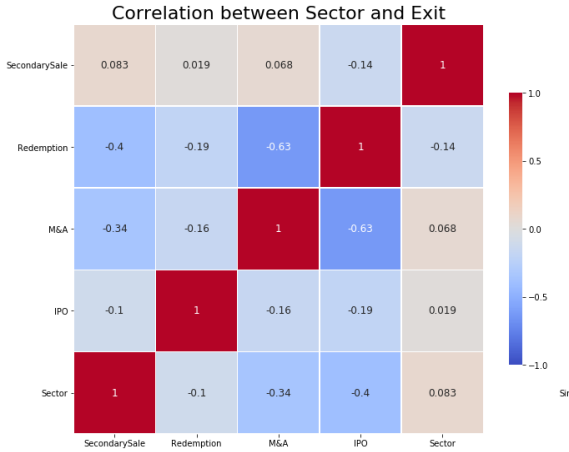
Correlation between the number of funding rounds and Exit



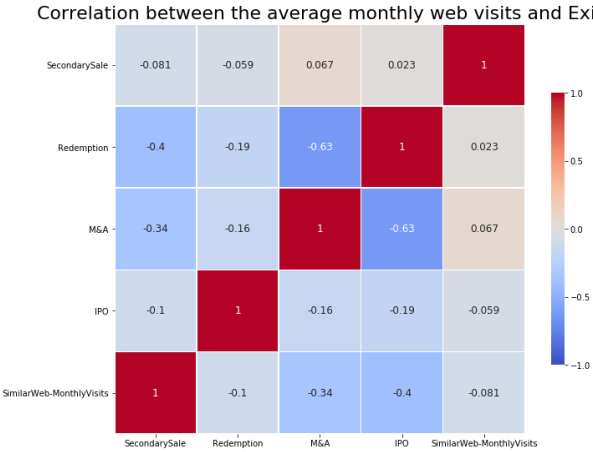
Correlation between the number of investors and Exit

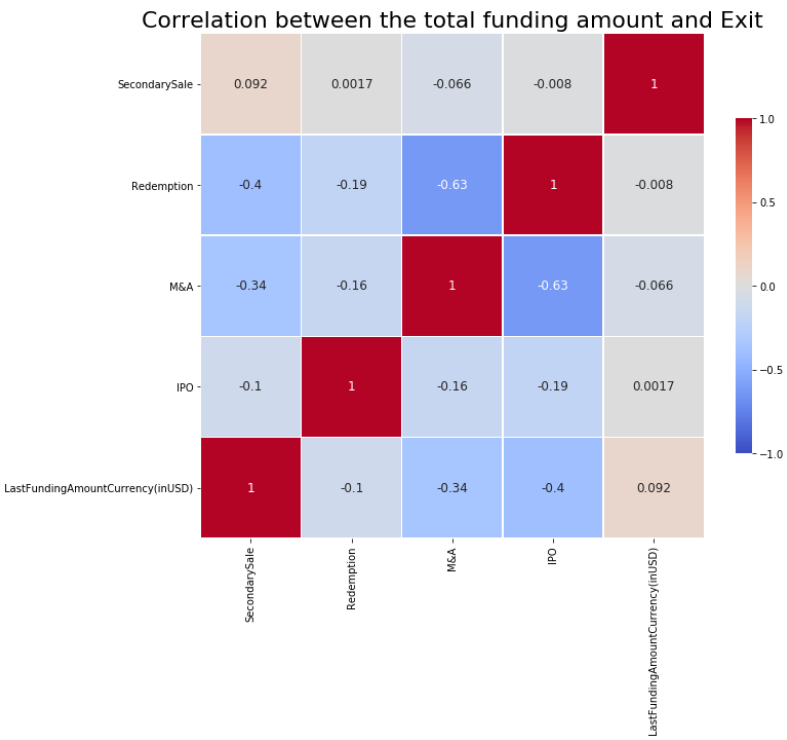


Correlation between Sector and Exit



Correlation between the average monthly web visits and Exit




Correlation between the total funding amount and Exit
SecondarySale
Redemption
M&A
IPO
LastFundingAmountCurrency(inUSD)
0.092
0.0017
-0.066
-0.008
1
-0.4
-0.19
-0.63
1
-0.008
-0.34
-0.16
1
-0.63
-0.066
-0.1
1
-0.16
-0.19
0.0017
1
-0.1
-0.34
-0.4
0.092
1.0
0.5
0.0
−0.5
−1.0

Appendix 4; Variables table

Variable:	Description:	Kept:
Organization Name	Name of the company	Yes
Total Funding Amount Currency (in USD)	Amount of money received (in USD) in all funding rounds	Yes
Number of Employees	Number of employees of the company	Yes
Exit	Type of exit the company performed; can be an IPO, a M&A, a Secondary Sale or Redemption	Yes
Last Funding Amount Currency (in USD)	Amount of money received (in USD) in the last funding round	Yes
Last Funding Type	Type of investors in the last round	Yes
Number of Funding Rounds	Total number of funding rounds	Yes
Number of Investors	Total number of investors	Yes
Location	Location of the company; divided in Europe, United States and rest of the world.	Yes
Founded Date	Founding date of the company	Yes
Number of Investments	Total number of investments the company received	Yes
Similar Web-Monthly Visits	On average how many people visit the company website	Yes
IP Patents Granted	Total number of Intellectual Property Patents granted to the company	Yes
Sector	Sector of the economy the company is active in	Yes
Number of Exits	Number of exits the company performed, for this data set it is always 1	Yes
Age at Exit	Age of the company at its exit in days	Yes
Exit Date	Date the company performed the exit	Yes
Operating Status	If the company is active or not, for this data the company is always active	Yes
Last Funding Date	Date last funding was received	Yes
Number of Founders	Number of founders of the company	Yes
Organization Name URL	Link to the company website	No
Industries	Industry the company operates in	No
Description	Description of the company	No
CB Rank (Company)	Crunchbase Rank showing the prominence of a company	No
Investor Type	Type of investor in last funding round	No
Total Funding Amount	Total amount of funding received in some currency	No
Total Funding Amount Currency	Currency type of the total amount of funding variable	No
Total Equity Funding Amount	Total amount of equity funding received in some currency	No
Total Equity Funding Amount Currency	Currency type of the total amount of equity funding variable	No
Total Equity Funding Amount Currency (in USD)	Total amount of equity funding calculated in USD.	No
Exit Date Precision	To what precision the exit date is certain	No
Company Type	Category company belongs to; used to determine sector	No
Acquisition Type	Was the acquisition an acquisition or a leveraged buyout	No
Investment Stage	Stage the last investment is made	No
Number of Founders	Number of founders of the company	No
Founders	Names of the founders of the company	No
Acquisition status	Has the acquisition happened already	No
Last Funding Amount	Amount received in the last funding round	No

Last Funding Amount Currency	Currency type of the last funding round	No
Money Raised at IPO	Money raised at the initial public offering	No
Money Raised at IPO Currency	Currency type of	No
Money Raised at IPO Currency (in USD)	Money raised at the initial public offering in USD	No
IPO status	Status of the IPO; public or delisted	No
Valuation at IPO	Valuation before initial public offering	No
Valuation at IPO Currency	Currency type of the valuation on the initial public offering	No
Valuation at IPO Currency (in USD)	Valuation before initial public offering in USD	No

Table 2. Variables with description

Appendix 5; Personal Evaluation

During my bachelor end project (BEP) I have gained multiple valuable insights regarding doing research and data science as a whole.

First of all, when doing research, you cannot simply start and find results. The layout of a research paper actually reveals this already. A proper literature study in which background research is performed helps to get a clearer overview of what you want to research. When disregarding this or performing this inadequately will eventually make it harder to create an ongoing story from your research while still providing valuable insights. I would say that I have done this background research inadequately which at the end of my BEP trajectory gave me problems creating a proper research paper.

Secondly, regarding planning. I thought that this would not be a major problem for me. I usually make plannings which are feasible and to which I keep myself. Most of my bachelor I have strictly held myself towards my own made plannings which increased my productivity. However, this period which includes BEP I have found it harder to keep myself to a planning while still having one. However, I do not think this is entirely related to the BEP on its own since I encountered this as well for the other courses I follow as well.

Thirdly with regards to data science, you are bound or limited by the data you have gathered. My initial research proposal covered different variables than the ones actually used for the study since the proposed variables were not obtainable. At first I thought this would not be a problem since I did gather data on which I could run a model. However, the effects of this only came visible later on when trying to make the best models possible and thinking about what would have been nice to include in the model as well. What I have learned from this is that gathering data is an important part of data science which I did not conclude beforehand. When your data is gathered in a way that it is hard to add data to the data set later on the initial data set should be good and be able to provide the information you want. For this again doing a fair amount of background research would help but you can never be certain that the data will provide what you want beforehand. Another solution to this problem is that you can try to create a dataset which can be updated and improved upon more easily. This would also help to enhance the reproducibility of the research.

Overall I have learned that gathering good data, having a good planning which is feasible and performing background research are all important aspects of writing a thesis. I will adhere myself to these points to improve them for the next time. Since no clear outline for this personal evaluation was given I will be able to reflect upon this evaluation more thoroughly at the final presentation or improve this version if necessary.

Appendix 6; Enlarged decision tree