Fast Growth Firms Prediction Model

By: Hasan Mansoor Khan

26.02.2023

INTRODUCTION

To carry out this task, I determined that a company would be classified as successful if its sales experienced a compound annual growth rate (CAGR) exceeding 30%. I specifically selected the years 2011-2012, with 2012 as the reference year, and calculated the CAGR change by computing the average annual sales rate between these years.

To reach any conclusion, it is important to narrow the scope of analysis and include specific variables. The report includes:

- Explaining the data set utilized, performing data cleaning, label and feature engineering
- Developing predictive models and selecting a model
- Generating probability predictions by utilizing models with increasing complexity
- Classification of findings by employing the loss function
- Creating a confusion matrix to assess the model's performance
- Conclusion: summing up the findings

DATA ANALYSIS

My first task is to know my data better. For this I explore the data source which can be found on this link: OSF Home. (A case study from Bekes & Kezdi's repository). The original data set comprised 287,829 observations and 48 variables, which encompassed all available information regarding company properties, balance sheets, profit and loss elements, and management information.

DATA CLEANING

For my analysis, it is essential to clean the data prior to performing any predictions or analysis. This is because the data set is vast and includes of various categories, many of which are not of interest. The cleaning begins by first selecting the years for my analysis. In this analysis, I filter for year 2011 and 2014. I then select year 2012, with 2011 as my base. I avoid using a two-year gap as most financial analysis is performed on a yearly basis. Alternatively, I could have also compared between 2010 and 2015, but as my aim is to provide investment advice with respect to high growing firms, a year on year is more accurate in terms of predictive analysis.

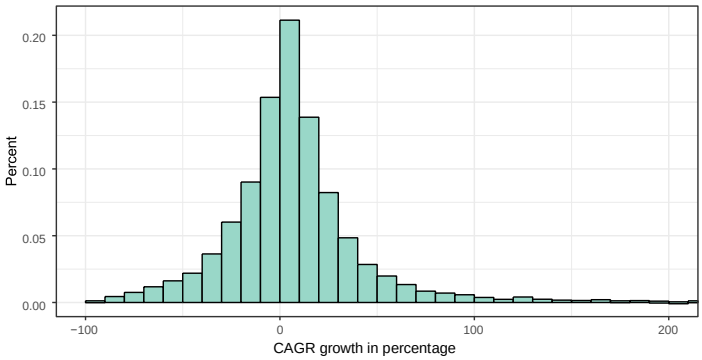
Apart from filtering for years, I also perform traditional cleaning tasks such as filtering out errors including negative sale values. I also remove companies which are non-alive and those with missing or NA values in sales. A detailed view on the cleaning process can be viewed in the cleaning file uploaded on the repository.

LABEL ENGINEERING

Next, I conducted filtering of the companies in two stages. Firstly, I excluded the firms with zero sales and marked them as inactive. Next, I filtered out companies with sales greater than 1000 euros but less than 10 million euros. Additionally, I generated a dummy variable named "fast growth" for companies with a compound annual growth rate (CAGR) exceeding 30%.

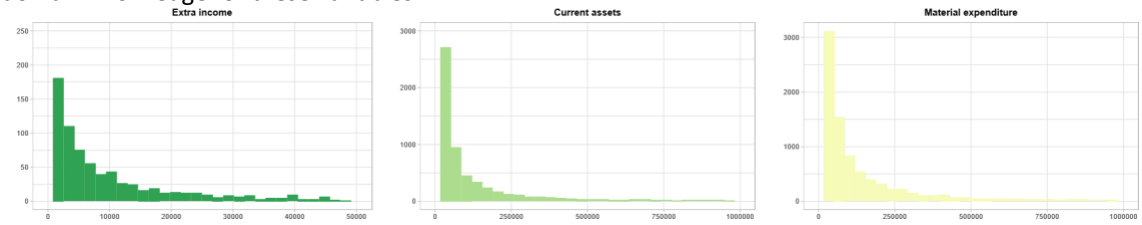
The distribution of the **CAGR** growth, which is the key variable in this study, is displayed below.

Distribution of CAGR growth (2011 to 2012)



FEATURE ENGINEERING

After completing label engineering, the next step was feature engineering. I focused on the financial variables and assessed their significance. I examined the distribution of certain financial variables, as illustrated below. This step is crucial in order to prevent any skewed results when transforming the variables. As I can see from the figure, the distribution is skewed. To rectify this issue, I applied either a logarithmic transformation or winsorizing, depending on the type of variable. Both methods were used in the study, as can be seen in the model selection process. Some variables were standardized, and then the ratios were winsorized. This means that I chose a threshold based on my domain knowledge for these variables.



In addition, I included flagging variables for any errors in the balance sheet, such as negative values. I also generated category variables and factors for future use. Finally, I eliminated observations with more than 90% missing values. As a result, the final or clean data set contained 116 variables and 11,910 observations for analysis.

PREDICTION MODELS & MODEL SETUP

My objective was to forecast fast-growing companies, so I computed the compound annual growth rate (CAGR) for each company from 2012 to 2014. I established a threshold of 30%, whereby an increase in CAGR was regarded as a significant improvement and thus designated a company as fast-growing. Approximately 16% of companies in our data set met this threshold. This is the underlying assumption I make as a data analyst for my models and analysis.

fast_growth_f	Number of companies	Percentage
no_fast_growth	9954	84%
fast_growth	1956	16%

CLASSFICATION: LOSS FUNCTION

The loss function is a useful approach to determine the optimal threshold for classification. By converting predicted probabilities into classifications, I can identify the ideal threshold for each of my models. Ultimately, I can determine the best model for prediction based on the lowest average expected loss.

The objective is to forecast fast growth in companies. In this context, false negatives pose a greater concern, as missing out on an investment opportunity due to predicting that a company won't grow could result in significant losses. Conversely, false positives may lead us to invest in a company that appears to be growing rapidly but is not. However, the financial loss incurred in this case would be lower, as it only means that the growth rate is slower but not negative.

MODEL SELECTION

	Number.of.predictors	CV.RMSE	CV.AUC	CV.threshold	CV.expected.Loss
Logit X1	11	0.3582516	0.6523847	0.2662623	0.4315753
Logit X3	35	0.3531567	0.6882375	0.2691901	0.4062807
LASSO	49	0.3530023	0.6688684	0.2340256	0.4296101
RF probability	33	0.3517776	0.6993763	0.2801155	0.4050179

To select my model, I will take into consideration the RMSE, AUC and most importantly the expected loss. As seen in the summary table above, The expected loss is lowest for Random Forest and then secondly for Logit Model 3 (X3). In third place is LASSO, while the last is LOGIT X1. It is important to note that the Random Forest and LOGIT Model X3 have very similar expected loss. On the other hand, as RMSE and AUC is considered, the Random Forest clearly out performs Logit model 3. In my opinion, the Random Forest method does a slightly better predictive analysis when compared to Logit or LASSO models. However, there is always a trade off and therefore a final decision depends also on other factors including simplicity and computational power. In case simplicity and interpretation is of utmost importance, the LOGIT model 3 is a reasonable model which scores a significantly low RMSE also and expected loss. Hence, I will proceed with Logit Model 3.

CONCLUSION

The analysis aimed to determine if a company could achieve a 30% increase in sales within a year. Logit model X3 and Random Forest were used to evaluate expected loss, RMSE, and AUC, taking into account model complexity. To ensure external validity, the model should be applied to a broader time period, such as every year between 2005 and 2016, to determine if coefficients remain significant over the decade.

The Random Forest model was found to be the best suited for making predictions in the data, with the LOGIT model 3 also suitable for this specific data set. However, for more accuracy, different models can be created for different industries, as this data is across industries. Lastly, the definition of **"Fast growth"** is subjective and may vary for different industries. For some traditional industries, achieving such a high rate may not be possible, and a lower threshold can be used to define "high growth."

For a complete detailed analysis, please view the technical Report using this link.