Public Summary of the Capstone Project Report

SCORING MODEL DEVELOPMENT RESEARCH FOR A PRIVATE EQUITY FIRM

Gergő Ropka

2021

Central European University

Department of

Economics and Business

Master of Science in Finance

Introduction of the project

The client, with their main product (venture capital), is focusing companies, who offer business-to-business software-as-a-service solutions, and preferably at a late-seed or Series A stage. Also, they also provide secondary liquidity sources to early-stage investors and provide a venture debt product which focuses on Series B+ stage companies.

One of the key principles of the firm's investment approach is the date driven decision making. They use tailored scoring models for each strategy to make and evaluate their investments. Thus, their scoring models need a constant maintenance, it is a never-ending research and development cycle, to be able to preserve the competitive advantage.

The main role of the project was to provide insights (based on a data set acquired by the client) regarding possible new factors, which could be included in their scoring method, or statistical methods, which could give a better scoring or classification. Additionally, as a residual outcome, evaluating the quality of the data provider's service.

The data set, what I received from the client, was basically a sample of a service provided by a third-party database provider, focusing on essential business information of recently established, small businesses (start-ups). Gathering relevant business information about start-ups can be a labour-intensive process, especially if we would like to build a comprehensive and complete database. Thus, the firm was also interested in the quality of this product, if they can spare the data collection phase by purchasing the service.

The scoring model, what the client provided as an example for their scoring methods, was "hybrid" model, which contained assumptions based on statistical analysis, but also professional experience-based conditions. The determination of the weights and the score calculation itself was more towards the experience-based type at that stage. Thus, it gave me room to search for improvement opportunities based on statistical methods.

Main challenges

The data set included over 100 variables, but some of these variables were only containing URLs or other general information, after dropping the irrelevant columns, still over 80 potential factors left. However, the sample only contained 98 observations, what raised some challenges.

It increased the chance of overfitting, if too many predictors included in the same model. The small sample size also makes the statistical tests less accurate and less robust, so it is much more challenging to make realisable assumptions based on the analysis.

The greatest challenge regarding the data set was definitely caused by the high number of mining values. Usually, the records, which have missing information, are simply dropped from the sample to ensure the quality of the tests. On the other hand, if I deleted these observations, then only an insufficiently small sample would remain.

I used logistic regression (logit model) as the main method of modelling, because it is relatively easy to interpret the results, also, there are various ways to test the efficiency of the model. The limited number of observations constrained the model selection as well, because the more advanced classification methods work better on big-data applications.

Generally speaking, regressions cannot handle missing values, so to overcome this challenge, I created subsets of the data to be able to drop the unsatisfactory observations without losing a significant proportion of the sample. Also, I tested replacing the missing information with the median or statistical mode of the given predictor. Overall, there were just minor differences in the results, based on which data cleaning method was used.

Results of the analysis and modelling

I was able to create models, which had 90+% classification accuracy, when I predicted the original goal variable, provided by the client. However, none of the models were statistically efficient, the Akaike information criterion showed in every case that the null-model was better than the non-empty models. In my opinion this outcome is the result of the quality of the data, the small sample size combined with the numerous missing values made the models inefficient.

Thus, I was not able to provide a totally reliable insights, although I believe based on the models, I could show, which predictors could be more important and worthy to collect and test further.

The data set included some historical change information for some predictors. Some of these historical change information could be included in the main models, but most of them had to be left out due to large number of empty values.

I tested these variables in small subsets, where I could drop all the observations with missing values. The client suspected these growth ratios could be important to identify investment opportunities and I could confirm that these simple models had better accuracy then the random classification and also based on the AIC values they were better than the empty model. These variables can be scraped in-house by the client from different websites, so it definitely worth to invest some time to collect a larger and more complete sample.

Additionally, I created a new dependant (goal) variable, based on the client's recommendation, to run additional models. This prediction was a more successful in a sense that it was not just an accurate prediction, but also a statistically more efficient model, based various tests. This prediction unfortunately cannot be used directly for the original classification problem. However, it gives a good baseline for the more important variables, because based on my tests, the same predictors provide a fairly accurate classification for the original goal variable as well.

Conclusion

Overall, I believe the project was successful in a sense that I could provide meaningful insights about the data set and its quality for the client. I also provided recommendation about the key factors, which could improve their scoring method. For me it was also a great opportunity to learn, I gained valuable knowledge about an additional segment of the financial services industry, what I would not encounter otherwise. To overcome the challenges of task, I had to learn new concepts about programming and statistics. The project required a significant amount of coding in R, which was an excellent chance to practice and improve my skills.