Will the Start-up Succeed? : The Impact of Skills on the Success or Failure of a Start-up Company

Aybüke Altuntaş, Gözde Akkaya, Gülin Oğuz

Middle East Technical University

{aybuke.altuntas, gozde.akkaya, gulin.oguz}@metu.edu.tr

1. Abstract

Nowadays, start-ups are trending all over the world. Even though not all of them succeed and become a long-term business, the amount of start-ups that succeed should not be underestimated. After examining the dataset, it was decided that skill factor and its subtopics were in the forefront. Thus, in order to make inferences about the success of a start-up, the contribution of skills, such as entrepreneurship, operations, engineering, data science and etc. were analyzed. Regression models and residual analyses were conducted to investigate research aims further. It was concluded that the model is a good fit but it should not be used for future predictions.

2. Introduction

The topic of this paper is the success or failure status of start-ups depending on different elements of skills that constitute for the skill factor all together. The skills, contributions of which were included in the model, are data science skills, domain skills, leadership skills and engineering skills. There exists a controversy about which skill is more important and is more needed in order to succeed for a start-up. As a result of this, it was decided to construct models and make use of stepwise selection. Even though the assumptions were not satisfied in the model building phase, treatment phase was skipped and it was directly proceeded to model validation. During the model validation phase, hypothesis testing and confusion matrix were used to determine accuracy and sensitivity.

3. Literature Review

Being able to access and possess knowledge, and therefore improve one's skills is an extremely important element for a company to start-up successfully. (Clercq & Arenius, 2006). To make a more general statement, the levels of skills of an individual contribute directly to the performance of that person, which is advantegous in terms of any kind of competition. Since start-ups are also in some sort of competition about securing their place in the market, it is aimed to investigate the skills of entrepreneurs to

make inferences about the success of their start-ups.

4. Data

The data was gathered from GitHub under the title of "AcqWire – Predict whether a start-up will succeed or not." The data contains 234 rows and a wide range of 50 variables. In this research, most of the variables about companies are not used, but instead, variables about the skills of founders are the main interest, since the research question was generated accordingly. Entrepreneurship, data science, engineering, operations, domain were some of the variables that are used in this research, detailed analysis of these skills will be discussed in the further parts of this paper.

Research Question

Danish Business Authority suggests that entrepreneurs come to entrepreneurship with different levels of skills and therefore each entrepreneur requires a different 'game plan' for developing his or her skills. As a result of this, main concern here is to investigate the skills levels of founders whether they have an impact on the success of startup or not.

5. Methods

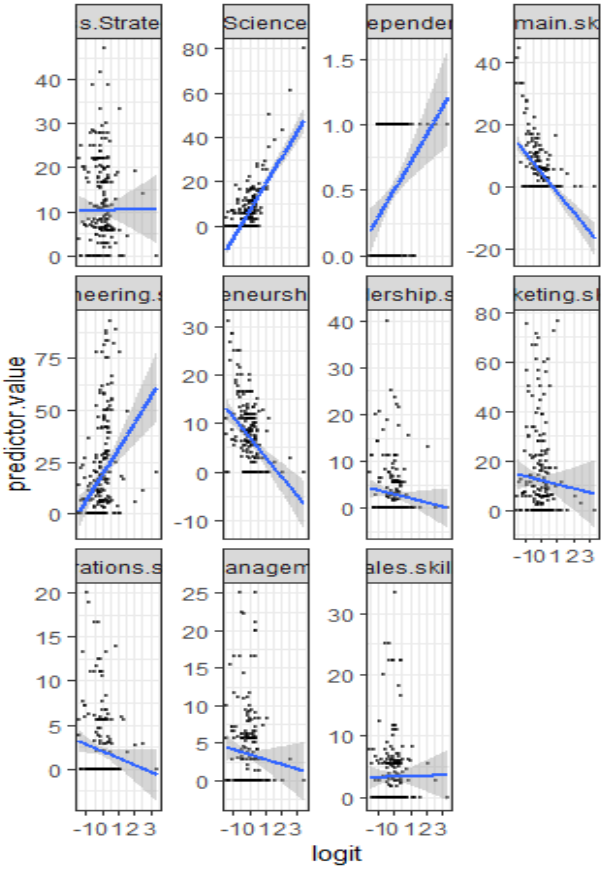
In order to investigate the relationship between startup's success and the skills levels of founders, a logistic regression model with binomial family were fitted to the data because of the binary outcome variable. After the model building, logistic regression assumptions were checked and since the main assumptions of the logistic regression model are not fully satisfied, variable selection techniques such as stepwise selection and forward selection were applied to model. After finding the best representative variables for the data, logistic regression assumptions were checked and model validation and accuracy check were also done.

6. Statistical Results

A logistic regression model with binary outcome Dependent and the covariates is fitted. The summary of logistic regression model with binomial family:

Coefficients	Estimate	SE	Pr(>|z|)
Intercept	-0.156595	0.273851	0.5674
Entrepreneurship.skills	-0.043721	0.027372	0.1102
Operations.skills	-0.029618	0.037533	0.4300
Engineering.skills	0.006853	0.006217	0.2703

Marketing.skills	0.005048	0.008357	0.5458
Leadership.skills	0.007066	0.027483	0.7971
Data.Science.skills	0.042684	0.017815	0.0166
Business.Strategy.skills	0.011949	0.016394	0.4661
Product.Management.skills	0.001736	0.029126	0.9525
Sales.skills	0.021502	0.027727	0.4380
Domain.skills	-0.033025	0.019392	0.0886

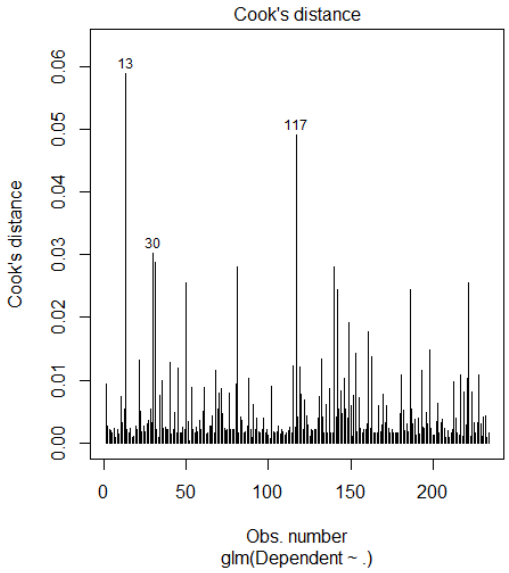


After model building logistic regression assumptions should be checked:

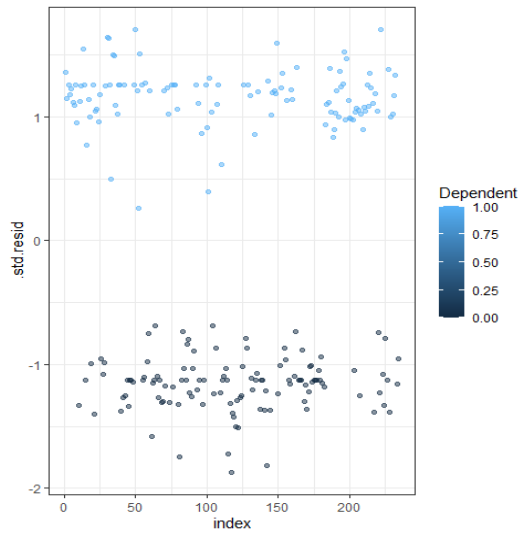
- The outcome variable, Dependent, is binary.
- Linearity assumption can be done visually inspecting the scatter plot between each predictor and the logit values.

As seen in the above plots, there is no linear relationship between the logit of the outcome and each predictor variables. Only the Data Science skills score, Domain skills score, Leadership skills score and Engineering skills score have linear relationship with the response.

- Influential values assumption can be examined by visualizing the Cook's distance values.



The Cook's distance plot shows 3 possible outlier values. However, not all outliers are influential observations. To check whether the data contains potential influential observations, the standardized residual error can be inspected. Data points with an absolute standardized residuals above 3 represent possible outliers.



As seen in the standardized residuals plot, there are no influential values.

- Multicollinearity assumption can be done by checking VIF values.

Entrepreneurships	Operations	Engineering	Marketing	Leadership	Data Science	Business Strategy	Product Management	Sales	Domain
1.855262	1.125114	1.161385	1.211039	1.245020	1.098742	1.489765	1.201632	1.246710	1.061075

As seen in the table above, there is no multicollinearity among the predictors. All variables have a value of VIF well below 5.

For this model, the main assumptions of the logistic regression are not fully satisfied and also from the summary of the model nearly all variables are insignificant, since all the p-values are higher than threshold 0.05.

Therefore, we should go over with variable selection methods to find the best fit for the observed data.

For the variable selection, first forward selection method was applied and the following outcome was obtained:

AIC	Model
Start: 208.5	Dependent ~ 1
204.71	Dependent ~ Domain.skills
201.39	Dependent ~ Domain.skills + Data_Science.skills
197.57	Dependent ~ Domain.skills +

	Data_Science.skills + Leadership.skills
196.86	Dependent ~ Domain.skills + Data.Science.skills + Leadership.skills + Engineering.skills

And final model suggested by the forward selection algorithm contains the Domain skills, Data Science Skills, Leadership skills and Engineering skills.

To ensure the outcome of forward selection method, stepwise selection is also applied and the same result is obtained too.

AIC	Model
205.28	Dependent ~ Entrepreneurship.skills + Operations.skills + Engineering.skills + Marketing.skills + Leadership.skills + Data.Science.skills + Business.Strategy.skills + Product.Management.skills + Sales.skills +Domain.skills
203.34	Dependent ~ Entrepreneurship.skills + Operations.skills + Engineering.skills + Marketing.skills + Leadership.skills + Data.Science.skills + Business.Strategy.skills + Sales.skills + Domain.skills
201.43	Dependent ~ Entrepreneurship.skills + Operations.skills + Engineering.skills + Leadership.skills + Data.Science.skills + Business.Strategy.skills + Sales.skills + Domain.skills
199.71	Dependent ~ Entrepreneurship.skills + Engineering.skills + Leadership.skills + Data.Science.skills + Business.Strategy.skills + Sales.skills + Domain.skills
198.15	Dependent ~ Entrepreneurship.skills + Engineering.skills + Leadership.skills + Data.Science.skills + Business.Strategy.skills + Domain.skills
197.14	Dependent ~ Entrepreneurship.skills +

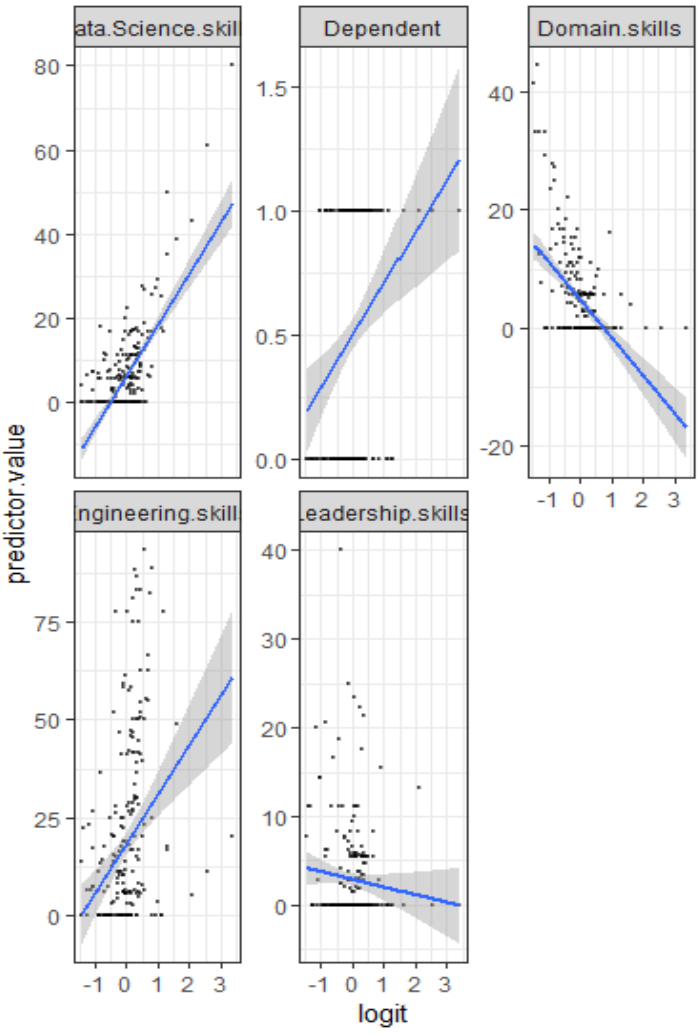
	Engineering.skills + Leadership.skills + Data.Science.skills + Domain.skills
196.86	Dependent ~ Engineering.skills + Leadership.skills + Data.Science.skills + Domain.skills

After variable selection the final model obtained and the summary of the model as follows:

Coefficients	Estimate	SE	Pr(>|z|)
Intercept	-0.1626	0.28461	0.5678
Data.Science.skills	0.0586	0.02374	0.0134 *
Domain.skills	-0.0540	0.02680	0.0438
Leadership.skills	-0.0816	0.03956	0.0391
Engineering.skills	0.0119	0.00738	0.1064

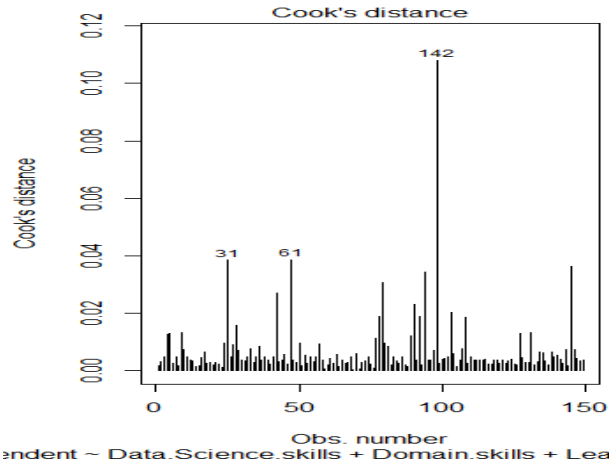
Again, logistic regression assumption should be checked for the final model.

- The outcome variable, Dependent, is binary.
- Linearity assumption can be done visually inspecting the scatter plot between each predictor and the logit values.

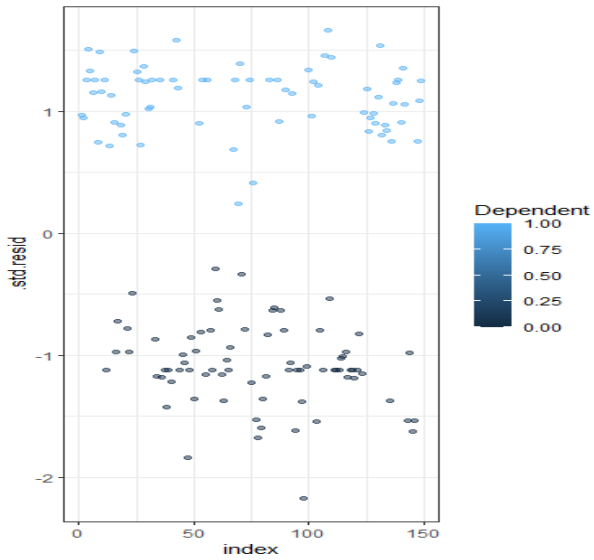


As seen in the smoothed scatter plots show that variables Data Science skills, Domain skills, Leadership skills and Engineering skills are linearly associated with the Dependent outcome in logit scale.

- Influential values assumption can be done by visualizing the Cook's distance values.



There seems to be 3 possible outlier values from the Cook's distance plot. However, not all outliers are influential points. To check this, the standardized residual error can be inspected. Data points with absolute standardized residuals above 3 represent possible outliers.



From the standardized residuals plot, there are no influential values.

- Multicollinearity assumption can be done by checking VIF values.

Data.Science.skills	Domain.skills
1.095234	1.007826
Leadership.skills	Engineering.skills
1.098678	1.011299

From the above table, there is no multicollinearity among the predictors. All variables have a value of VIF well below 5. For the final model, the main assumptions of the logistic regression are fully satisfied and also from the summary of the model nearly all variables are significant. To measure the quality of the fit of a given model and evaluate its performance in order to avoid poorly fitted models, The Hosmer and Lemeshow goodness of fit (GOF) test was applied. This test is generally used for grouped data but since we don't have any groups, we take g which is number of groups as 10 the default value.

H_0: Model is a good fit

H_1: Model is not a good fit

From the output, p-value was obtained as 0.524. That is larger than 0.05 so, we can't reject the null hypothesis and conclude that the model seems to fit well.

After checking goodness of fit, accuracy was checked by confusion matrix and following results are obtained.

Accuracy	0.5176
95% CI	(0.4066, 0.6274)
No Information Rate	0.5059
P-Value [Acc > NIR]	0.4570
Kappa	0.0381
Mcnemar's Test P-Value	0.1183
Sensitivity	0.3953
Specificity	0.6429
Pos Pred Value	0.5312
Neg Pred Value	0.5094
Prevalence	0.5059
Detection Rate	0.2000
Detection Prevalence	0.3765
Balanced Accuracy	0.5191

Since 95% CI for accuracy contains no info rate, this model shouldn't be used for further estimations. Also, for the other levels such as sensitivity, specifity, etc. desired value is to be higher than 0.5; but in this example we have 0.4 for sensitivity.

7. Conclusion

After all the statistical findings, obtained model is a good fit of the observed data, however it shouldn't be used for future predictions. Fixing the potential problems might improve considerably the usability of the model.

References

Clercq, D. D., & Arenius, P. (2006). The role of knowledge in business start-up activity. *International Small Business Journal: Researching Entrepreneurship*, *24*(4), 339–358. https://doi.org/10.1177/0266242606065507