STATISTICAL ANALYSIS OF SUCCESS OF START-UP'S

A FINAL PROJECT REPORT SUBMITTED
IN FULFILMENT OF THE REQUIREMENTS FOR COURSE STAT 364 –
LINEAR MODELS II
DEPARTMENT OF STATISTICS OF
METU

BY

DİLAY GÜMÜŞ 2361301
İLAYDA YILMAZ 2361657
OKAN ÖZHAYAT 2361459

June 2022

TABLE OF CONTENT

1. Introduction

Recently, many start-ups are established, but their success rate is not that high. The success of start-ups is crucial and depends on many factors. Many entrepreneurs are looking for reasons for this and ways to achieve success. In this study, different start-ups were examined. During the study, several factors which are related with the start-ups' success status were investigated such as company crowdfunding and the company's location. In order to analyze the data, at the first step descriptive statistics were benefited. For further analysis, a logistic regression model, statistical tests such as the chi-square independence test, and graphics were used to reveal the factors that are related to the success status of the start-ups.

2. Literature Review

Starting a business may be mentally and physiologically challenging. Recognizing typical startup issues and major success criteria, on the other hand, may assist entrepreneurs to get started on the right process. Early studies found that if family concerns are valued, women are more likely to join the industry and are less likely to become a man's subject. Nevertheless, it appears that it is much more challenging for females to become successful in work than it is for males who lack education and abilities (Hazudin et al.,2015). Another study uncovered that the key factors for the success of a start-up are innovation (R&D), internal market openness, government, and financing. They stated that expanding government assistance for startups and improving the marketplace plays a significant role in assuring the development and success of startups (Okrah et al.,2018).

3. Data

The data includes 234 observations about start-ups. We used company business model, company crowdfunding, company mobile app, founder's popularity, founder's industry exposure, company top angel VC funding, and company location as categorical variables. In the graphics below, the counts of each of these variables based on companies' success used are given.

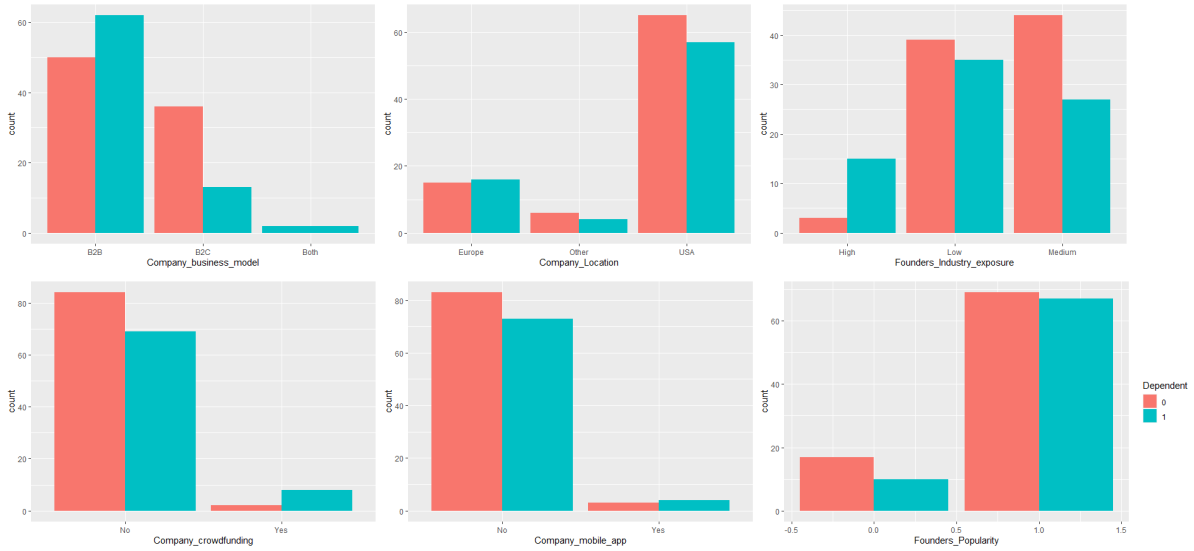

Figure 1

We used company senior team count, company analytics score, company investor count seed, founders' domain skills score, and founder's sales skills score as numeric variables.

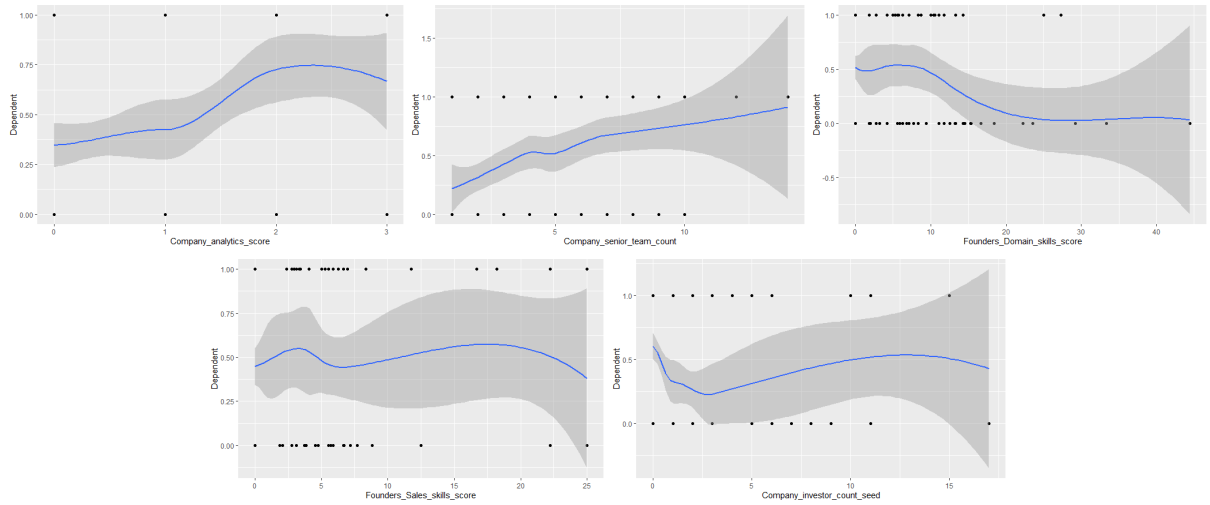

Figure 2

4. Methods

Chi-square test of independence: It shows whether the two categorical or nominal variables are most probably related or not.

Logistic Regression model: It is used for the prediction of the binary outcome e.g. 1 or 0, depending on the beforehand variables.

Hosmer Lemeshow Test: It is used for testing the goodness of fit for the logistic regression

models.

Anderson Darling Normality Test: It is used to understand whether the data are from a population with a certain distribution.

Plots: Scatter plots, bar plots and the correlation matrix plot are used.

5. Statistical Results

5.1 What are the variables that have a significant relationship with the company's success status?

	Estimate	Pr(>|z|)
Intercept	0.61473	0.5052
Company_senior_team_count	0.18860	0.0222
Founders_Operations_skills_score	-0.10382	0.0544
Company_competitor_count	-0.12509	0.0603
Company_1st_investment_time	0.01697	0.1549
Company_analytics_score	0.37182	0.0408
Company_business_modelB2C	-1.03368	0.0139
Company_business_modelBoth	1.70870	0.1473
Founders_Industry_exposureLow	-1.51398	0.0784
Founders_Industry_exposureMedium	-1.61724	0.0567
Company_crowdfundingYes	2.19297	0.0149

Figure 3

The logistic regression model was used to see the relationship between the independent variables and the response. The forward selection, forward stepwise regression, and LRT variable selection methods are used to find the best subset model. All the methods suggested the best model contains the variables that are number of top management employees, operational skill score of founders, number of direct competitors of the company, time in months to get 1st investment, analytics score of company, if company business model is B2B, B2C or both, industry exposure of founders, if the company is crowdfunding related.

5.2 According to different types of skill scores of founders, which type of skills are most related to whether start-ups are successful or not?

According to the established model, the founder's domain knowledge has a significant relationship with the start-up’s success. In addition, it was seen that the overall talent score, rather than the field skill scores, has a significant relationship with the success of the start-up.

5.3 Is there a significant relationship between it is funded by top Angel or VC funds whether the start-up is successful or not?

	Not Successful	Successful
No	71	64
Yes	15	26

Figure 4

By Chi-Square Independence Test, there is no statistically significant relationship between whether a company has been funded by top Angel or VC funds, and a start-up is successful or not.

Model Adequacy Check:

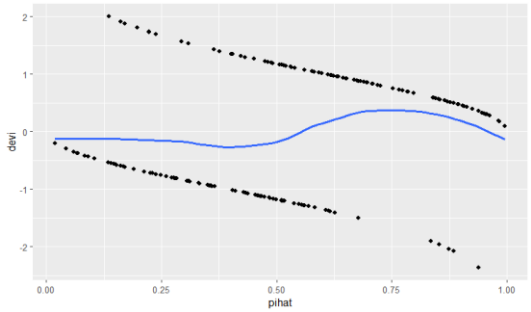

Figure 5

There does not appear to be a strong association between devi and estimated probability because the lowess curve has a 0 intercept and exhibits practically no pattern in Figure 5, therefore it appears to be a good fit.

Normality Check: Anderson-Darling normality test indicates that the deviance residuals are not from a normal distribution.

The Goodness of Fit Test: Hosmer and Lemeshow's goodness of fit (GOF) test gives the p-value = 0.4945. From this, it is concluded that the model is a good fit.

Multicollinearity Check: Multicollinearity seems to exist between 2 variables, but VIF values are checked. All VIF values are lower than 10. So, there is no multicollinearity problem.

Outliers and influential observations: By looking at the influence measure function, 10 points seem to be influential by cov.r. From the function, it seems that there are 2 highly influential observations. Also, there are 7 possible high leverage observations.

Model Validation: The accuracy of the test set is 0.7414 and the train set is 0.7614. So, they have the good accuracy. Also, the sensitivity of the test set is 0.7308 and the sensitivity of the

train set is 0.6556. The sensitivity value is good. The specificity of the test set is 0.75 and the specificity of the train set is 0.8721. Also, when we obtained the RMSE value as 1.64 in the test set and 1.41 in the train set. So, these values are close to each other.

	Accuracy	%95 CI	No Information Rate	Sensitivity	Specificity	Pos Pred Value	Neg Pred Value
Train Set	0.7614	(0.6914, 0.8223)	0.5114	0.6556	0.8721	0.8429	0.7075
Test Set	0.7414	(0.6096, 0.8474)	0.5517	0.7308	0.7500	0.7037	0.7742

Figure 6

6. Conclusion/Discussion

At the first research question, the main aim was to see which variables are the most related with the company's success status. All of the variable selection techniques resulted in the same model. As a result of this, the best subset model was found to predict the company's success status. For the second research question, the subset logistic regression model was formed to see which type of skills are the most related for the success status of the company. According to the logistic regression model result, the most important skill is the domain knowledge score of founders. It was emphasized that the total talent score is crucial in order to be a successful start-up. For the third research question, since it is desired to see whether the categorical response variable is related with the categorical regressor, the Chi-Square Independence Test was used. The test provided a p-value which was higher than 0.05. Therefore, the null hypothesis was not rejected, and concluded that whether start-up is funded by top Angel or VC has no significant relationship with company's success status. For the model validation part, we obtain close values in train and test sets. The accuracy and sensitivity values are satisfactory in both train and test sets. Also,RMSE values in train and test sets are close to each other. These suggest that the model used is valid. The model has a good fit from the result of the goodness of fit test. By the Anderson-Darling test, residuals are not normally distributed. There is no multicollinearity problem by looking at the VIF values. Also, by influence measure function, 8 points are influential. From research question 1, the company's own characteristics were found to be more relevant to the success of the startup. However, we think that the characteristics of the founders are related to the success of the startup. Therefore, more detailed research can be done using the data about the founders.

REFERENCES

Hazudin, S., Kader, M., Ishak, M. & Ali, R. (2015). Discovering Small Business Startup Motives, Sucess Factors and Barries: A Gender Analysis.https://www.sciencedirect.com/science/article/pii/S2212567115012186?via%3Dihub

Okrah,J.,Nepp,A.,Agbozo,E.(2018). Exploring the factors of startup success and growth. (2018, April). Retrieved June 14, 2022, from https://www.researchgate.net/publication/336642098_Exploring_the_factors_of_startup_success_and_growth