ANA625 – Categorical data analysis

Project:

Analysis of Investment Pattern in Small Businesses for USA and India

Report created by:
Aditi Bhujbal

Date : March 29th 2022.

Content

OBJECTIVE

To investigate whether there is any association between 'getting an investment for a small business' and different domains that the business belongs to, while considering the fact that the sample product was presented to investors and the country.

INTRODUCTION

It all started with Shark Tank..!!

Shark Tank is a **business reality series** where entrepreneurs make their business presentations to a panel of investors, who decide whether to invest in their company or not. These investors are called "Sharks", as they have redefined the meaning of entrepreneurship and business in today's world.

Around $4million (30 Crores INR) was invested during the first season of this show in India, and $143million worth of capital has been invested during shark tank seasons in USA since 2009.

This is a show where money talks, but what are the chances an entrepreneur gets a deal on this show? Let's find out!

DATA

This study focuses on the relationship between Getting an investment (Deal = Yes/No and domain/industry of businesses (Sector) in the United States and India. Specifically, the businesses which were presented at Shark Tank is the population of interest.

Dataset links:

1. https://www.kaggle.com/competitions/ban-502-shark-summer-2021/data
2. https://www.kaggle.com/datasets/sriomsubham/shark-tank-data-121pitches

These datasets are generated from all the episodes of Shark Tank series in India and summer series in the U.S.

They contain the information regarding ideas that were presented on Shark Tank in both the countries. Dataset from USA had around 550 records and from India had 121 records, making the total number of records together as 672. Out of those, 24 records had missing value for Sector variable which is 3.57% of total.

Feature Engineering:

1. New categorical variables created from existing variables:
    a. Created a new variable called **'country'** with values **'USA'** and **'India'** in the datasets for respective countries.
    b. Shark_tank_USA dataset:
        I. It had two binary variables as: Deal_yes & Deal_no, from them **DEAL** variable with values Yes and no was created.
        Yes : Deal_yes = 1 and No: Deal_no = 1.

II. From the dummy variables malepresenter ,femalepresenter and mixedgenderpresenter, a new variable called Sex is created which contains values male, female and mix.

III. Similarly, from the dummy variables which has binary values for categories of domain/industry such as Software_Tech, Health_Wellness, Fashion & beauty, a new variable called **SECTOR** is created with different values as Technology, Health, Lifestyle, etc.

IV. The records with missing value for sector were updated with value 'other'.

2. Finally, both the datasets were merged together to create a single dataset called **"Shark_Tank".**

The CONTENTS Procedure

Data Set Name	ANA625.SHARK_TANK	**Observations**	672
Member Type	DATA	**Variables**	6
Engine	V9	**Indexes**	0
Created	24/03/2022 00:13:34	**Observation Length**	80
Last Modified	24/03/2022 00:13:34	**Deleted Observations**	0
Protection		**Compressed**	NO
Data Set Type		**Sorted**	NO
Label			
Data Representation	SOLARIS_X86_64, LINUX_X86_64, ALPHA_TRU64, LINUX_IA64		
Encoding	utf-8 Unicode (UTF-8)		

Figure 1: Proc content output for dataset Shark_Tank

Graphical Representation of Deal vs Sector groups by Country:

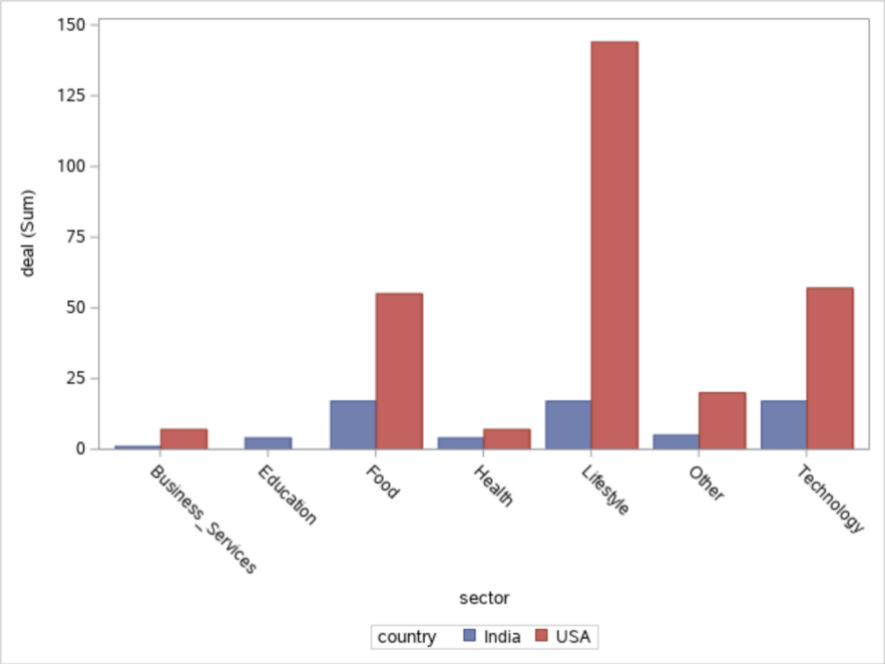

Figure 2: Bar Chart of Deal vs Sector (for USA and India)

MODEL

The research objective of this study, investigating the association between deal and the sector of a business, can be summarized by the following:

DEAL = f (SECTOR,COUNTRY, NOVELTIES)

Where DEAL represents whether an entrepreneur got a deal or not (yes/no); SECTOR represents industries/domains these businesses belong to and are divided into 7 different categories such as technology, food, health, lifestyle; COUNTRY has values USA and INDIA; and NOVELTIES variable indicates whether the actual product presented to the investors (1=yes, 0=no).

STATISTICAL ANALYSIS

The statistical analysis performed in this study consists of both tests of association and logistic regression. Pearson X^2 tests of association will be performed between the control variables and the exposure variable and are presented in Table 1. Similar tests will be performed between the control & exposure variables and the outcome variable, presented in Table 2. Logistic regression is used to estimate adjusted odds ratios and their 95% confidence intervals for the outcome variable (DEAL) with respect to the exposure (SECTOR) and control variables (COUNTRY & NOVELTIES), presented in Table 3.

With respect to the regression analysis, tests for confounding between the exposure and control variables are performed, goodness of fit statistics are reported, and interactions between the exposure and control variables are investigated. All statistical analysis is performed using SAS.

RESULTS

- First of all, let's find whether there is any association between exposure variable(Sector) and control variables (country and novelties).

Table 1. Characteristics of 672 entrepreneurs by sector.											
Variable		Total		Country				Novelties			
				USA		India		0 (No)		1 (Yes)	
		N	%	n	%	n	%	n	%	n	%
Population		672	100.00%	551	81.99%	121	18.01%	565	84.08%	107	15.92%
Sector (Exposure variable)	Business_Services	34	5.06%	30	5.44%	4	3.31%	30	5.31%	4	3.74%
	Education	6	0.89%	0	0.00%	6	4.96%	4	0.71%	2	1.87%
	Food	129	19.20%	96	17.42%	33	27.27%	96	16.99%	33	30.84%
	Health	20	2.98%	12	2.18%	8	6.61%	14	2.48%	6	5.61%
	Lifestyle	291	43.30%	261	47.37%	30	24.79%	269	47.61%	22	20.56%
	Other	81	12.05%	66	11.98%	15	12.40%	53	9.38%	28	26.17%
	Technology	111	16.52%	86	15.61%	25	20.66%	99	17.52%	12	11.21%
P-value*		---		<.0001				<.0001			
* p values based on Pearson chi-square test of association											

Of the entire population, 81.9% entrepreneurs belong to USA and 15.92% of all presented novelties. In USA, Lifestyle lead the small businesses with 47.37% of all, wherein In India, large number of entrepreneurs from Food sector (27.27%) came to present their ideas (p-value < .0001). Of all who presented their product to investors, a greater number of participants were from Food industry (30.84% with p-value<.0001).

- The demographic characteristics of this population are compared in Table 2 with respect to the outcome variable, Getting and investment (DEAL).

Table 2. Characteristics of 672 entrepreneurs based on whether they got the Deal or not.								
Variable		Population		Deal				p value *
				0 (No)		1 (Yes)		
		N	%	n	%	n	%	
Total		672	100.00%	304	45.24%	368	54.76%	
Sector	Business_Services	34	5.06%	26	8.55%	8	2.17%	0.0011
	Education	6	0.89%	2	0.66%	4	1.09%	
	Food	129	19.20%	57	18.75%	72	19.57%	
	Health	20	2.98%	9	2.96%	11	2.99%	
	Lifestyle	291	43.30%	130	42.76%	161	43.75%	
	Other	81	12.05%	43	14.14%	38	10.33%	
	Technology	111	16.52%	37	12.17%	74	20.11%	
Country	USA	551	81.99%	250	82.24%	301	81.79%	0.8816
	INDIA	121	18.01%	54	17.76%	67	18.21%	
Novelties	0 (No)	565	84.08%	256	84.21%	309	83.97%	0.9317
	1 (Yes)	107	15.92%	48	15.79%	59	16.03%	
* p values based on Pearson chi-square test of association								

Overall, 54.76% of the all participants had a deal after they presented their idea/business to investors. The largest number of participants who got the deal were from Lifestyle sector (43.75%; p<0.0001). Almost 84% who got the deal had presented their product to investors (p-value = 0.9317) and 81.79% of participants to receive the deal were from USA(p-value = 0.8816). Since P-value for both the control variables (COUNTRY and NOVELTIES) is greater than 0.05, their relationship with outcome variable DEAL is not statistically significant.

- Adjusted odds ratios for DEAL with respect to the exposure and control variables obtained from the logistic regression are presented in Table 3.

Table 3. Logistic regression analysis comparing the adjusted odds ratio of getting a Deal when entrepreneurs from different sectors are compared with each other after controlling for Sector and Novelties.

Variable		Deal = 0 (No)		Deal =1 (Yes)		OR*	95% CI	
		n	%	n	%		lower Limit	Upper limit
Total		304	45.24%	368	54.76%			
Country	USA	250	82.24%	301	81.79%	---	---	---
	India	54	17.76%	67	18.21%	0.692	0.322	1.485
Sector	**Business_Services**	26	8.55%	8	2.17%	**0.147**	**0.060**	**0.358**
	Education	2	0.66%	4	1.09%	1.217	0.200	7.420
	Food	57	18.75%	72	19.57%	0.599	0.350	1.027
	Health	9	2.96%	11	2.99%	0.601	0.227	1.586
	Lifestyle	130	42.76%	161	43.75%	**0.699**	**0.376**	**0.953**
	Other	43	14.14%	38	10.33%	**0.388**	**0.201**	**0.748**
	Technology	37	12.17%	74	20.11%	---	---	---
Novelties	0 (No)	256	84.21%	309	83.97%	---	---	---
	1 (Yes)	48	15.79%	59	16.03%	1.518	0.641	3.594

*** 95% confidence intervals are for reported odds ratios.**

Odds of entrepreneurs from Business_Services category getting a deal was 19.6% lower when compared to those from Technology sector after controlling for Country and Novelties (OR=0.804; 95% CI = 0.062-0.364). Similarly, Lifestyle businesses had odds ratio of 0.607 with 95% CI = 0.383 – 0.964 and Other categories of sector had odds ratio of 0.412 with 95% CI = 0.223 – 0.762 when compared with Technology. Odds ratio for remaining categories of Sector along with Country and Novelties are not statistically significant as their 95% CI include 1.

LOGISTIC REGRESSION MODEL

❖ Main Effect model:

$$DEAL = ß_0 + ß_1 (SECTOR) + ß_2 (COUNTRY) + ß_3 (NOVELTIES)$$

Outcome variable : DEAL
Exposure Variable : SECTOR
Controlled for : COUNTRY & NOVELTIES

➢ Null hypothesis:

H_0 = There is no association between DEAL and SECTOR
i.e. $ß_1 = ß_2 = ß_3 = 0$.

Analysis of Maximum Likelihood Estimates						
Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		1	0.7333	0.2100	12.1914	0.0005
sector	Business_Services	1	-1.9177	0.4546	17.7989	<.0001
sector	Education	1	0.1962	0.9225	0.0452	0.8316
sector	Food	1	-0.5121	0.2750	3.4684	0.0626
sector	Health	1	-0.5100	0.4957	1.0585	0.3036
sector	Lifestyle	1	-0.5128	0.2372	4.6762	0.0306
sector	Other	1	-0.9476	0.3351	7.9960	0.0047
country	India	1	-0.3686	0.3899	0.8937	0.3445
Novelties	1	1	0.4172	0.4399	0.8997	0.3429

Figure 3: MLE output table for main effects model

Odds Ratio Estimates			
Effect	Point Estimate	95% Wald Confidence Limits	
sector Business_Services vs Technology	0.147	0.060	0.358
sector Education vs Technology	1.217	0.200	7.420
sector Food vs Technology	0.599	0.350	1.027
sector Health vs Technology	0.601	0.227	1.586
sector Lifestyle vs Technology	0.599	0.376	0.953
sector Other vs Technology	0.388	0.201	0.748
country India vs USA	0.692	0.322	1.485
Novelties 1 vs 0	1.518	0.641	3.594

Figure 4: Odds Ratio Estimates generated by main effects model

➢ Global Chi-Squared and AIC test:

- **Global Chi-Square**:

  X_G^2 = -2LogL(Intercept) + -2LogL(Intercept + Covariate) = 892.378 – 867.935
  = **24.44**

  Since, P-value of likelihood ratio is 0.0019 < 0.05, we can reject null hypothesis and conclude that there is an association between DEAL and SECTOR when controlled for COINTRY and NOVELTIES i.e. at least one of the coefficients ($ß_1$, $ß_2$, $ß_3$) is not zero.

Testing Global Null Hypothesis: BETA=0			
Test	**Chi-Square**	**DF**	**Pr > ChiSq**
Likelihood Ratio	24.4426	8	0.0019
Score	23.8524	8	0.0024
Wald	21.6755	8	0.0056

Figure 5: Global null hypothesis testing

- **Akaike's Information Criterion (AIC):**

  Since $AIC_{(Intercept\ only)}$ > $AIC_{(Intercept + Covariates)}$, we can reject null hypothesis and can say that, main effects model i.e. DEAL = f(SECTOR, COUNTRY, NOVELTIES) is better fitting model than the intercept only model.

Model Fit Statistics		
Criterion	**Intercept Only**	**Intercept and Covariates**
AIC	894.378	885.935
SC	898.851	926.200
-2 Log L	892.378	867.935

Figure 6: Model Fit Statistic for main effects model

Hence, we can confirm that there is an association between getting an investment for a business and its sector/domain.

➢ Deviance Test:

H_0 : Coefficient on interaction term is 0

or

Current model fits the data better than saturated model.

Deviance statistic value = 6.1826, indicates that there's still a little variation in DEAL which is not explained by current model. However, P-value = 0.7215 which is > 0.05 indicates that this deviance is not statistically significant.

Therefore, we reject H_0 and conclude that current model or main effects model fits the data better than saturated model.

Deviance and Pearson Goodness-of-Fit Statistics				
Criterion	Value	DF	Value/DF	Pr > ChiSq
Deviance	6.1826	9	0.6870	0.7215
Pearson	4.9054	9	0.5450	0.8425

Number of unique profiles: 18

Figure 7: Deviance statistic for main effects model

❖ Full Model:

This model consists of all variables as below,

DEAL = ß0+ ß1(Sector) + ß2(Country) + ß3(Novelties) + ß4(Sex)

or

DEAL = $ß_0$ + $ß_1$(sector=Business_Services) + $ß_2$(sector=Education) + $ß_3$(sector=Food) + $ß_4$(sector=Health) + $ß_5$(sector=Lifestyle) + $ß_6$(sector=Other) + $ß_7$(Country=India) + $ß_8$(Novelties=1) + $ß_9$(sex=1) + $ß_{10}$(sex=2)

➢ Estimates generated by full model using SAS:

Analysis of Maximum Likelihood Estimates						
Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		1	0.6607	0.2196	9.0563	0.0026
sector	Business_Services	1	-1.8938	0.4553	17.3010	<.0001
sector	Education	1	0.0576	0.9303	0.0038	0.9506
sector	Food	1	-0.5432	0.2767	3.8541	0.0496
sector	Health	1	-0.5197	0.4958	1.0987	0.2945
sector	Lifestyle	1	-0.5157	0.2377	4.7085	0.0300
sector	Other	1	-0.9439	0.3366	7.8666	0.0050
country	India	1	-0.3867	0.3903	0.9817	0.3218
Novelties	1	1	0.4007	0.4394	0.8315	0.3618
Sex	1	1	0.1624	0.2025	0.6434	0.4225
Sex	2	1	0.2496	0.2173	1.3194	0.2507

Figure 8: MLE output table for Full model

- **Odds ratios and %change in OR:**
  % change in OR for DEAL with respect to the SECTOR and different combination of control variables COUNTRY, NOVELTIES and SEX are presented in Table 4.1 and 4.2:

Table 4.1 % change in OR for each model.

Variable		Full Model		Remove sex		Remove Novelties		Remove Country	
		OR	% Chg OR	OR	% Chg OR	OR	% Chg OR	OR	% Chg OR
Sector	Business_Services vs Technology	0.153	----	0.15	-1.96%	0.156	1.96%	0.157	2.61%
	Education vs Technology	1.007	----	1.121	11.32%	0.932	-7.45%	0.879	-12.71%
	Food vs Technology	0.597	----	0.612	2.51%	0.616	3.18%	0.61	2.18%
	Health vs Technology	0.6	----	0.605	0.83%	0.611	1.83%	0.597	-0.50%
	Lifestyle vs Technology	0.606	----	0.607	0.17%	0.614	1.32%	0.621	2.48%
	Other vs Technology	0.416	----	0.412	-0.96%	0.444	6.73%	0.439	5.53%

Table 4.2: % change in OR for each model.

Variable		Full Model		Remove Novelties & sex		Remove Country & sex		Remove country & Novelties	
		OR	% Chg OR	OR	% Chg OR	OR	% Chg OR	OR	% Chg OR
Sector	Business_Services vs Technology	0.153	----	0.153	0.00%	0.154	0.65%	0.157	2.61%
	Education vs Technology	1.007	----	1.038	3.08%	0.981	-2.58%	0.887	-11.92%
	Food vs Technology	0.597	----	0.632	5.86%	0.624	4.52%	0.616	3.18%
	Health vs Technology	0.6	----	0.616	2.67%	0.601	0.17%	0.604	0.67%
	Lifestyle vs Technology	0.606	----	0.616	1.65%	0.621	2.48%	0.62	2.31%
	Other vs Technology	0.416	----	0.441	6.01%	0.433	4.09%	0.446	7.21%

- Since, change in OR is less than 10% for all the models generated by removing variables Sex, Novelties, Country one at a time, we can say that the relationship between DEAL and SECTOR **cannot be confounded by** relationship between sector and Country or sector and Novelties or sector and Sex.

- From figure , we can see p-value for coefficients of Country, Novelties and sex is greater than 0.05, i.e. these variables are not statistically significant, hence they **cannot be retained in the model**.

CONCLUSION

The objective of this study was to identify whether there is any relationship between a deal of an investment in a business and sector of the business, while controlling for the country and the presentation of novelty. Based on the results of Global-chi squared and Deviance tests, we can conclude that there is an association between "DEAL" and "SECTOR", however, %change in OR proves that country and novelties are neither confounding variables nor they're controlled for in this model. This research also finds that the odds of getting an investment for a business in field of technology is higher than the others except for Education, wherein odds ratio for education is not significant.

FUTURE WORK

1. I would like to include continuous variables e.g. invested amount, equity into this research and analyze the investment trends in different industries or by investors.
2. I would also include the data for other countries to understand the world-wide investment patterns.

REFERENCES

1. Shark Tank deep dive: A data analysis of all 10 seasons: https://thehustle.co/shark-tank-data-analysis-10-seasons/
2. Shark Tank Ranked: https://www.mansworldindia.com/entertainment/shark-tank-india-startup-investment-sharks/