VIDYASAGAR UNIVERSITY

DETERMINING THE IMPORTANT FACTORS FOR CHOOSING RARE SUBJECT AS CAREER IN UG

PROJECT SUBMITTED FOR PARTIAL FULFILLMENT OF BACHELOR'S DEGREE IN SCIENCE IN STATISTICS HONOURS

REGISTRATION NO.: 1160501 OF 2020-21

ROLL: 1126116 NO.: 200141

DEPT. OF STATISTICS, HALDIA GOVT. COLLEGE

ACKNOWLEDGEMENT

The success and final outcome this project required a lot of guidance and assistance from many people and I am extremely privileged to have got this all along with the completion of my project. All that I have done is only due to their supervision and assistance and I would not forget to thank them. I respect and thank for providing me an opportunity to do the project work under the department of statistics, Haldia Government College and giving you all the support and guidance that I required, which made me complete the project duly on time. I owe my deep gratitude to our project guides Dr. Shyamsundar Sahoo, Mr. Sibsankar Karan, Mr. Tanmay Maity, Mr. Bijitesh Halder, who took keen interest on my project work and guided me all along, till the completion of our project work by providing all the necessary information for developing a good system. I would not forget to remember my parents, and also my friends for their encouragement and more over for their timely support and guidance till the completion of our project work.

CONTENTS

INTRODUCTION

After completing HS, students are very confused about choosing their career. Most of the students opt Engineering and Medical stream. On one hand choosing a general subject as a career option is very rare. Above that choosing a rarer subject such as Statistics,

Geology, Anthropology, etc., is once in a blue moon. Specifically in rural areas, students are more unaware about these subjects as compared to urban students. So, in this project, we would try to analyze various factors responsible for making the students opt for such rarer subject by using some statistical tools and draw a conclusion on this study.

OBJECTIVE

The objectives of my project are simple, these are

- To determine the important factors for choosing rare subject as a career in UG.
- To check the association between the categorical variables using Pearson's Chi-Square Test.

- To check effects of the various factors on the decision of choosing rare subject using logistic regression.
- To identify the most important factors effecting the decision of choosing rare subject using stepwise regression.

DATA COLLECTION

For my Project, I have collected the data from the students who are studying in the colleges situated in the rural areas via both, online and offline mode. For collecting the data in online mode, I have made a google form (questionnaire), which I have attached in the 'Appendix'. And for collecting the data via offline mode, I have done a survey in my college in several departments. From this I have experienced and learned a lot about data collection.

DATA DESCRIPTION

There are 12 factors in my data. These are 'Availability of the subject in HS', 'Getting interest in the subject', 'Parents' preference', 'Own decision', 'Teacher's guidance', 'Friend's or relative's influence', 'Job opportunity', 'Not get any chance in other course', 'Sex', 'Father's education', 'Mother's education' and 'Family income'. And the dependent variable is 'Target'.

Size of the data = 150, where the no. of rare subject = 75 and the no. of rare subject = 75

❖ Factors descriptions:

1. Availability of the subject in HS: If the student had the subject in his/her HS level. (Categorical) [Yes/No]
2. Getting interest in the subject: If the student is getting interest in the subject. (Categorical) [Yes/No]
3. Parents preferences: If the student has taken the subject for his/her parents' preference. (Categorical) [Yes/No]
4. Own decision: If the student has taken the subject by own decision. (Categorical) [Yes/No]
5. Teacher's guidance: If the student has taken the subject for any teacher's guidance. (Categorical) [Yes/No]
6. Friend's or relative's influence: If the student has taken the subject for his/her friend's or relative's influence. (Categorical) [Yes/No]

7. Job opportunity: If the student has taken the subject for job opportunity. (Categorical) [Yes/No]
8. Not get any chance in other course: If the student has taken the subject not for getting chance in other course. (Categorical) [Yes/No]
9. Sex: The sex of the student. (Categorical) [Male/Female]
10. Father's education: The qualification of father of the student. (Categorical) [Up to matriculation, Higher secondary/Graduation/Post Graduation]
11. Mother's education: The qualification of mother of the student. (Categorical) [Up to matriculation, Higher secondary/Graduation/Post Graduation]
12. Family income: Family income of the student. (Numerical)

- Target: The target is defined as "the subject is rare or not."

❖ Index of the Factors:

- For the factors 'Availability of the subject in HS', 'Getting interest in the subject', 'Parents' preference', 'Own decision', 'Teacher's guidance', 'Friend's or relative's influence', 'Job opportunity' and 'Not get any chance in other course', Yes = 1 and No = 0
- For the factor 'Sex', Male = 1 and Female = 0.
- For the factors 'Father's education' and 'Mother's education',
  Up to matriculation = 0, Higher secondary = 1, Graduation = 2 and Post Graduation = 3.
- For the factor 'Family income', 0-9999 = 0,
  10000-19999 = 1,
  20000-49999 = 2
  and 50000 and above = 3
- For the variable, 'Target', Rare = 1 and Not rare = 0.

METHODOLOGY

Pearson's chi-square for independence:

The Chi-square test of independence checks whether two variables are likely to be related or not. We have counts for two categorical or nominal variables. We also have an idea that the two variables are not related. The test gives us a way to decide if our idea is plausible or not.

This is the motivation behind the hypothesis for the Chi-Square Test of Independence:

H_0: In the population, the two categorical variables are independent.

H_1: In the population, the two categorical variables are dependent

The Chi-Square test statistic is calculated as follows:

$$\chi^2 = \sum_{i=1}^{rc} \frac{(o_i - E_i)^2}{E_i} \quad \sim \chi^2_{(r-1)(c-1)}, \text{ under null}$$

hypothesis.

where o_i = the observed frequency

E_i = the expected frequency

r = no. of rows

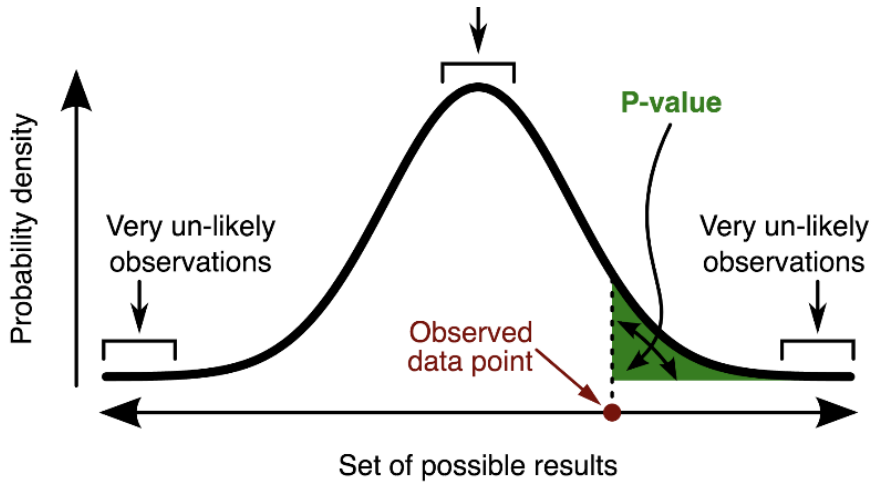
c = no. of columns.

Under the null hypothesis and certain conditions, the test statistic follows a Chi-Square distribution with degrees of freedom equal to $(r-1)(c-1)$, where r is the number of rows and c is the number of columns. We leave out the mathematical details to show why this test statistic is used and why it follows a Chi-Square distribution. As we have done with other statistical tests, we make our decision by either comparing the value of the test statistic to a critical value (rejection region approach) or by finding the probability of getting this test statistic value or one more extreme (p-value approach).

P-Value:

In statistics, the p-value is the probability of obtaining results at least as extreme as the observed results of a statistical hypothesis test, assuming that the null hypothesis is correct. The p-value is used as an alternative to rejection points to provide the smallest level of significance at which the null hypothesis would be rejected. A smaller p-value means that there is stronger evidence in favour of the alternative hypothesis. We Know that P-value is a statistical measure, that helps to determine whether the hypothesis is correct or not. P-value is a number that lies between 0 and 1. The level of significance(α) is a predefined threshold that should be set by the researcher. It is generally fixed as 0.05.

P-value	Decision
P-value > 0.05	The result is not statistically significant and hence don't reject the null hypothesis.
P-value < 0.05	The result is statistically significant. Generally, reject the null hypothesis in favor of the alternative hypothesis.
P-value < 0.01	The result is highly statistically significant, and thus reject the null hypothesis in favor of the alternative hypothesis.



Logistic Regression:

- Logistic regression is a supervised technique to find the probability of dependent variable.
- Logistic regression uses functions called the 'logit' functions, that helps derive a relationship between the dependent variable and independent variables by predicting the probabilties.
- The logistic functions (also known as the Sigmoid functions) convert the probabilities into binary values which would be further used for predictions.
- It's a classification algorithm, that is used where the response variable is categorical.
- The idea of logistic regression is to find a relationship between features and probability of particular outcome.
- Logistic regression is used for clasification problems in machine learning.
- Usually there are two types of supervised machine learning problems
  i) Linear regression where prediction value is continuous.
  ii) Classification where predicted value is categorical.
- It establishes the relationship between a categorical variable and one or more independent variables.
- This relationship is used in machine learning to predict the outcome or more independent variables.
- This is used in many different fields such as the medical field, trading and business, technologgy and many more.

➢ Types of Logistic Regression:

1. **Binary logistic regression:** The dependent variable has only two possible outcome/classes.
   Ex.: Yes/No.
2. **Multinomial logistic regression:** The dependent variable has only three or more possible outcome/classes without ordering.
   Ex.: Red, Green, Blue
3. **Ordinal logisic regression:** The dependent variable has only three or more possible outcome/classes with ordering.
   Ex.: Movie rating 1 to 5.

➢ Assmptions:

It assumes that there are minimal, or no-collinearity among the independent variables.

The best way to check the presence of multi-collinearity is to perform VIF (Variance Inflation Factor).

- The dependent variable must be categorical in nature.

➢ Build a Logistic regression model :

The logistic function is given by the following formula:

$$\sigma(z) = \frac{e^z}{1+e^z}$$

or, $$\sigma(z) = \frac{1}{1+e^{-z}}$$

Now for simple linear regression,

$$z = a + bx$$

The sigmoid function for simple linear regression will be,

$$\sigma(z) = \frac{1}{1+e^{-(a+bx)}}$$

And for multiple linear regression,

$$z = \alpha + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n$$

Hence, the Sigmoid function will be

$$\sigma(z) = \frac{1}{1+e^{-\left(\alpha+\sum_{i=1}^{n}\beta_i X_i\right)}}$$

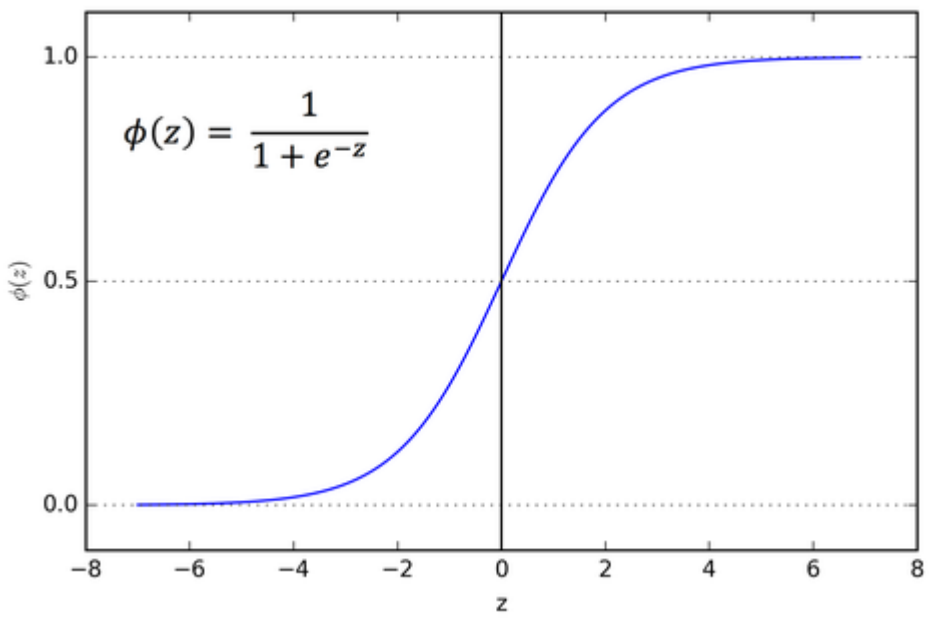


Fig: Logistic regression (sigmoid) curve.

Suppose, $\{y_i, X_i\}$, i = 1(1)n, contains independent samples.

For the Single observation the logit model is

$$P\{y_i = 1|X_i\} = \frac{e^{(x_i^T\beta)}}{1+e^{(x_i^T\beta)}}$$

$$P\{y_i = 0|X_i\} = 1 - \frac{e^{(x_i^T\beta)}}{1+e^{(x_i^T\beta)}}$$

$$= \frac{e^{(x_i^T\beta)}}{1+e^{(x_i^T\beta)}}$$

The likelihood of a single observation (y_i, X_i) is

$f(y_i, X_i \mid \beta) = p^{y_i}(1-p)^{(1-y_i)}$

$$= \left\{\frac{e^{(x_i^T\beta)}}{1+e^{(x_i^T\beta)}}\right\}^{y_i} \left\{\frac{1}{1+e^{(x_i^T\beta)}}\right\}^{(1-y_i)}$$

Hence, the joint likelihood is

$$\mathcal{L}(\beta; y, X) = \prod_{i=1}^{n} \{p_i^{y_i}(1-p_i)^{(1-y_i)}\},$$

where $p_i = \frac{e^{(x_i^T \beta)}}{1+e^{(x_i^T \beta)}}$

MLE of β:

$$\hat{\beta} = \underset{\beta}{argmax}\mathcal{L}(\beta; y, X)$$
$$= \underset{\beta}{argmax} ln\mathcal{L}(\beta; y, X)$$
$$= \underset{\beta}{argmin}[-ln\mathcal{L}(\beta; y, X)]$$

Stepwise Regression:

- **Stepwise regression** is the step-by-step iterative construction of a regression model that involves the selection of independent variables to be used in a final model. It involves adding or removing potential explanatory variables in succession and testing for statistical significance after each iteration.
- **Forward selection** begins with no variables in the model, tests each variable as it is added to the model, then keeps those that are deemed most statistically significant—repeating the process until the results are optimal.

RESULTS & CONCLUSIONS

We've checked association between the dependent variable 'Target' and the other factors by using **Chi-square test**. The result is given as following:

Table 1: The results from Chi-square test

Factors	χ^2 value	d.f.	P-Value	Decision
Availability of the subject in HS	96.618	1	2.2e-16	Reject
Getting interest in the subject	1.1884	1	0.2757	Accept
Parents preferences	1.2401	1	0.2655	Accept
Own decision	7.0417	1	0.007963	Reject
Teacher's guidance	6.5041	1	0.01076	Reject
Friend's or relative's influence	0	1	1	Accept
Job opportunity	11.863	1	0.0005725	Reject
Not get any chance in other course	3.5877	1	0.05821	Accept
Sex	2.1916	1	0.1388	Accept
Father's education	9.8328	3	0.02004	Reject
Mother's education	6.5344	3	0.08832	Accept
Family income	11.581	3	0.008966	Reject

Here we can clearly see that the P-values, for the factors 'Getting interest in the subject', 'Parents preferences', 'Friend's or relative's influence', 'Not get any chance in other course', 'Sex', 'Mother's education', are greater than $\alpha = 0.05$. Hence, there is no association of these factors with the dependent variable, denoted as 'Target'.

On the other hand, the P-values for the factors 'Availability of the subject in HS', 'Own decision', 'Teacher's guidance', 'Job opportunity', 'Father's education', 'Family income', are lesser than $\alpha = 0.05$. And hence, these factors have association with the dependent variable, 'Target'.

Now, I have checked the effects of the factors on the dependent variable using **Logistic regression** as following:

Table 2: The results from Logistic regression

Factors	Estimate	Standard error	z value	Pr(>|z|)	Decision
Availability of the subject in HS	-5.22354	0.81299	-6.425	1.32e-10	Reject
Getting interest in the subject	0.32194	2.22907	0.144	0.8852	Accept
Parents preferences	-2.48777	1.14034	-2.182	0.0291	Reject
Own decision	-1.98451	1.08315	-1.832	0.0669	Accept
Teacher's guidance	1.24399	0.91048	1.366	0.1718	Accept
Friend's or relative's influence	0.21111	0.96375	0.219	0.8266	Accept
Job opportunity	0.37471	0.66852	0.561	0.5751	Accept
Not get any chance in other course	-0.68372	0.90409	-0.756	0.4495	Accept
Sex	-0.61074	0.79762	-0.766	0.4439	Accept
Father's education	0.18001	0.6272	0.287	0.7741	Accept
Mother's education	0.46728	0.82151	0.569	0.5695	Accept
Family income	0.04829	0.48303	0.1	0.9204	Accept

On applying logistic regression, we have arrived at the conclusion that the factors 'Getting interest in the subject', 'Own decision', 'Teacher's guidance', 'Friend's or relative's influence', 'Job opportunity', 'Not get any chance in other course', 'Sex', 'Father's education', 'Mother's education', 'Family income' have no effect on the dependent variable denoted as 'Target'. Only 'Availability of the subject in HS' and 'Parents preferences' have effects on the dependent variable, 'Target'.

Results from Stepwise Regression:

Table 2: The results from Stepwise regression

Factors	Estimate	Standard error	t-value	Pr(>|t|)
Availability of the subject in HS	-0.79155	0.04694	-16.862	< 2e-16
Teacher's guidance	0.16079	0.06047	2.659	0.00872
Mother's education	0.06822	0.03411	2.000	0.04735

As we know that stepwise regression is usually used to identify the most significant factor among the independent variables, in this case we have found that 'Availability of the subject in HS' is the most crucial factor in effecting the dependent variable. After that the second most important factor, in this case is 'Teacher's guidance' following which there is 'Mother's education'. So, there are these three most effecting factors which are identified by stepwise regression.

- After applying the three different methods in the project, we have arrived on a conclusion that the factor, 'Availability of the subject in HS', is the only common effective factor in the all three methods. Practically, we can observe if a subject is available in higher secondary, then the students are more likely to opt the subject in UG.

APPENDIX

Questionnaire of the data:

1. What is your **current UG course?** *

_____

2. Did you have this subject in your HS level? *

Mark only one oval.

- ◯ Yes
- ◯ No

3. Are you **getting interest in your subject?** *

Mark only one oval.

- ◯ Yes
- ◯ No

What are the reasons for choosing this subject as UG course?
Give the answers of the following questions which are the part of the above question. **Select 'Yes' for which factors you choose the subject, otherwise select 'No'.** (You can choice multiple reasons)

4. **Parents Preference:** *

Mark only one oval.

- ◯ Yes
- ◯ No

5. **Own decision:** *

Mark only one oval.

- ◯ Yes
- ◯ No

6. **Teacher's guidance:** *

Mark only one oval.

- Yes
- No

7. **Friend's or relative's influence:** *

Mark only one oval.

- Yes
- No

8. **Job opportunity:** *

Mark only one oval.

- Yes
- No

9. **Not get any chance for other course:** *

Mark only one oval.

- Yes
- No

10. ***Any other reasons?***

_____

Some Personal Queries:

11. Your **sex:** *

Mark only one oval.

- Male
- Female
- Others

12. Your **father's education:** *

Mark only one oval.

- ◯ Below matriculation
- ◯ Matriculation
- ◯ Higher secondary
- ◯ Graduation
- ◯ Post Graduation

13. Your **mother's education:** *

Mark only one oval.

- ◯ Below matriculation
- ◯ Matriculation
- ◯ Higher Secondary
- ◯ Graduation
- ◯ Post Graduation

14. Your **family income (Monthly in Rs.):** *

Data:

Avail in HS	Getting int	Parents pr	Own decis	Teacher's	Friend's or	Job Oppor	Not get an	Sex	Father's ec	Mother's e	Family inc	Target
1	1	0	1	1	0	0	0	1	3	2	2	1
0	0	0	0	0	1	1	0	1	2	0	0	1
0	1	0	1	0	0	1	0	1	1	1	0	1
0	1	0	1	0	0	0	0	1	0	0	0	1
0	1	0	0	0	0	1	0	1	2	2	3	1
0	1	0	1	0	0	1	0	1	0	0	2	1
0	1	0	1	0	0	1	0	1	0	0	1	1
0	1	0	1	0	0	0	0	1	0	0	0	1
0	1	0	0	0	1	1	0	1	2	2	3	1
0	1	0	1	1	0	1	0	1	2	0	3	1
0	1	0	0	1	0	0	0	1	1	0	2	1
0	1	1	1	0	1	1	0	1	0	0	1	1
0	1	1	0	1	0	1	0	0	1	0	0	1
0	1	0	1	0	0	0	0	1	1	0	2	1
0	1	1	0	1	0	1	1	1	1	0	0	1
0	1	0	0	0	0	1	0	1	2	2	1	1
0	1	0	0	0	0	1	0	1	2	1	2	1
0	1	0	0	1	0	1	0	0	1	0	2	1
0	0	0	1	0	1	0	0	1	0	0	1	1
0	1	0	0	0	0	1	0	1	3	3	2	1
1	1	0	1	1	0	1	0	0	3	2	1	1
0	1	0	1	0	0	0	0	1	3	2	3	1
0	1	0	1	1	0	1	1	1	2	0	1	1
0	1	0	1	1	0	1	0	0	1	1	1	1
0	1	1	0	1	1	1	0	0	3	2	3	1
0	1	0	1	0	1	1	0	1	1	1	2	1
0	1	1	1	0	0	1	0	1	2	2	3	1
0	1	0	1	0	0	1	0	1	0	0	0	1
0	1	0	1	0	0	1	0	1	0	0	1	1
0	1	0	1	0	0	0	0	1	0	0	0	1
0	1	0	0	0	1	1	0	1	1	0	1	1
0	1	0	1	0	0	0	0	0	3	2	0	1
1	1	0	1	0	0	0	0	1	0	0	0	1
1	1	1	0	0	0	0	0	0	0	0	0	1
0	1	0	1	0	0	0	0	1	1	1	1	1
0	1	0	0	0	0	1	0	0	2	1	0	1
0	1	0	1	0	0	0	0	0	0	0	1	1
0	1	0	1	0	0	0	0	0	0	0	1	1
1	1	0	1	0	0	1	0	0	0	0	0	1
0	1	0	1	0	0	1	0	1	2	1	2	1
1	1	1	0	1	0	0	0	0	3	2	3	1
0	1	0	1	0	0	0	1	1	0	0	0	1
0	1	0	1	0	1	1	0	0	1	0	0	1
1	1	0	1	0	0	1	0	0	1	1	1	1
1	1	0	1	1	0	0	0	0	1	0	1	1
0	1	0	1	1	0	0	1	1	0	0	0	1
0	1	0	1	0	1	0	1	1	1	0	1	1
0	1	0	1	1	0	0	0	1	0	0	1	1
1	1	0	0	0	0	1	0	0	0	0	1	1
1	1	0	0	0	0	1	0	0	2	1	2	1
0	1	0	1	1	0	1	0	1	0	0	1	1
0	1	0	0	1	0	1	1	0	0	0	0	1
0	1	0	1	0	0	1	0	1	0	0	1	1
0	1	1	0	0	0	1	1	1	1	0	1	1
0	1	0	1	1	0	1	0	0	0	0	0	1
0	1	0	1	0	0	0	0	0	0	0	0	1
0	1	0	0	1	0	0	1	1	0	0	1	1
0	1	0	1	0	0	1	0	1	0	0	0	1
0	1	0	1	0	0	0	1	0	0	0	1	1
0	1	0	1	0	0	0	1	1	1	0	0	1
0	1	0	1	0	0	1	0	0	1	0	0	1
0	1	0	1	0	0	0	0	0	0	0	1	1
0	1	0	1	0	1	1	0	1	0	1	1	1
0	1	0	1	0	0	1	1	1	0	0	0	1
0	1	0	0	1	0	1	0	1	2	0	1	1
0	1	0	1	0	1	0	1	1	1	0	1	1
0	1	0	1	0	0	1	0	1	0	0	1	1
0	1	0	1	0	0	0	0	1	1	1	0	1
0	1	0	0	1	0	1	1	0	1	0	1	1
0	1	0	1	0	0	0	0	0	2	1	2	1
0	1	0	1	0	0	0	0	1	0	0	1	1
0	1	0	1	0	0	1	0	1	1	1	1	1
0	1	0	1	0	1	1	0	0	1	0	1	1
0	1	0	1	0	0	0	0	1	0	0	0	1
0	1	1	1	0	0	1	0	0	3	2	3	1

1	1	0	1	0	0	0	0	1	0	0	0	0
1	1	0	1	0	0	1	0	1	0	0	0	0
1	1	0	1	0	0	1	0	1	1	0	0	0
1	1	0	1	0	0	0	1	1	0	0	1	0
1	1	0	1	0	0	0	0	1	0	0	0	0
1	1	0	1	0	0	1	0	1	0	0	0	0
1	1	0	1	0	0	0	0	0	0	0	0	0
1	1	0	1	1	0	1	0	0	2	0	0	0
1	1	0	1	1	0	1	0	0	2	0	1	0
1	1	1	1	0	0	0	0	0	2	1	2	0
1	1	0	1	0	1	0	1	1	2	1	1	0
1	1	1	0	0	0	1	0	0	2	2	2	0
1	1	0	1	0	0	0	1	1	1	0	0	0
1	1	0	1	0	0	1	0	0	1	1	2	0
1	1	0	1	0	1	0	0	0	0	0	0	0
1	1	0	1	0	0	0	1	1	0	0	0	0
1	1	1	1	0	0	1	0	0	2	1	2	0
1	1	0	1	0	1	0	1	1	1	0	1	0
1	1	0	1	0	0	1	0	1	0	0	1	0
1	1	1	0	1	0	0	0	0	2	1	2	0
1	1	0	1	0	1	1	0	1	1	1	1	0
1	1	0	1	0	0	1	1	1	1	0	1	0
1	1	1	1	0	0	1	0	0	2	1	2	0
1	1	0	1	1	0	1	0	1	0	0	1	0
1	1	0	1	0	1	0	1	1	2	1	2	0
1	1	0	1	0	0	0	0	1	0	0	0	0
1	1	0	1	0	0	0	0	0	0	0	0	0
1	1	1	0	1	0	1	0	0	2	2	2	0
1	1	0	1	0	1	1	1	1	1	0	1	0
1	1	0	1	0	0	0	0	0	0	0	0	0
1	1	0	1	0	0	0	1	0	1	0	1	0
1	1	0	1	0	1	0	0	0	0	0	1	0
1	1	0	1	0	0	0	1	0	0	0	0	0
1	1	1	1	0	0	0	1	1	1	0	1	0
1	1	0	1	1	0	0	0	0	0	0	0	0
1	1	0	1	0	1	0	1	1	1	0	0	0
1	1	0	1	0	0	1	0	1	0	0	0	0
1	1	0	1	0	0	0	1	0	1	1	3	0
1	1	0	1	1	0	0	0	0	1	0	2	0
1	1	0	1	0	0	0	0	1	0	0	0	0
0	1	0	1	0	0	0	0	1	0	0	0	0
1	1	0	1	0	1	1	0	1	1	1	1	0
0	1	1	0	0	0	1	0	1	2	2	2	0
0	1	0	1	0	0	0	0	0	0	0	1	0
0	1	0	1	0	0	1	0	1	0	0	0	0
1	1	0	0	0	0	0	1	0	1	0	0	0
1	1	0	1	0	0	0	0	0	0	0	0	0
1	1	1	0	0	0	0	0	0	0	1	0	0
1	1	0	1	0	0	0	0	0	0	0	0	0
1	1	1	1	0	0	1	0	0	0	0	0	0
1	0	1	1	0	0	1	0	0	2	0	0	0
1	1	1	1	0	0	1	0	0	0	0	0	0
1	1	1	1	0	0	1	0	0	0	0	0	0
1	0	0	1	0	0	0	1	1	0	0	0	0
1	1	0	1	0	0	0	0	1	0	0	1	0
1	1	0	1	0	1	0	0	0	1	0	0	0
1	1	0	1	0	0	0	0	1	0	0	1	0
1	1	1	0	0	0	0	0	0	2	1	3	0
1	1	0	1	0	0	0	1	0	0	0	0	0
1	1	0	1	0	0	0	0	1	0	0	3	0
1	1	0	1	0	0	0	0	1	1	0	0	0
1	1	0	1	0	1	0	1	1	0	0	1	0
1	0	1	0	0	0	0	1	0	0	0	0	0
1	1	0	1	0	0	0	0	0	0	0	0	0
1	1	0	1	0	0	0	1	1	1	1	2	0
1	1	0	1	0	0	0	0	0	0	0	0	0
1	0	0	1	0	0	0	1	1	0	0	0	0
1	1	0	1	0	0	0	0	1	0	0	0	0
1	1	0	1	0	0	0	0	0	1	0	1	0
1	1	0	1	0	0	0	1	1	0	0	0	0
1	0	0	1	0	0	0	1	1	0	0	0	0
1	1	0	1	0	0	0	1	0	0	0	0	0
1	1	0	1	0	0	0	0	1	0	0	0	0
1	1	0	1	0	0	0	0	0	1	0	0	0
1	0	0	1	0	0	0	1	0	0	0	0	0

Command in R:

```
setwd("C:\\Users\\Somnath Samanta\\Downloads")
d<-read.csv("ProWork2.csv")

chisq.test(d$Avail.in.HS,d$Target)
chisq.test(d$Getting.interest.in.the.subject,d$Target)
chisq.test(d$Parents.preference,d$Target)
chisq.test(d$Own.decision,d$Target)
chisq.test(d$Teacher.s.guidance,d$Target)
chisq.test(d$Friend.s.or.relative.s.influence,d$Target)
chisq.test(d$Job.Opportunity,d$Target)
chisq.test(d$Not.get.any.chance.for.other.course,d$Target)
chisq.test(d$Sex,d$Target)
chisq.test(d$Father.s.education,d$Target)
chisq.test(d$Mother.s.education,d$Target)
chisq.test(d$Family.income..monthly.in.Rs..,d$Target)

glmfit=glm(Target~Avail.in.HS+Getting.interest.in.the.subject+
Parents.preference+Own.decision+Teacher.s.guidance+
Friend.s.or.relative.s.influence+Job.Opportunity+
Not.get.any.chance.for.other.course+Sex+Father.s.education+
Mother.s.education+Family.income..monthly.in.Rs..,
data=d,family=binomial)
summary(glmfit)

f_model <- lm(Target ~ ., data = d)
nu_model <- lm(Target~1,data=d)
library(MASS, lib.loc = "C:/Program Files/R/R-4.3.0/library")
summary(stepAIC(nu_model,direction = "forward",scope =list(upper=f_model,lower = nu_model)))
```

Output in R:

```
Console   Terminal ×   Background Jobs ×
R 4.3.0 · C:/Users/Somnath Samanta/Downloads/
> setwd("C:\\Users\\Somnath Samanta\\Downloads")
> d<-read.csv("ProWork2.csv")
>
> chisq.test(d$Avail.in.HS,d$Target)

        Pearson's Chi-squared test with Yates' continuity correction

data:  d$Avail.in.HS and d$Target
X-squared = 96.618, df = 1, p-value < 2.2e-16

> chisq.test(d$Getting.interest.in.the.subject,d$Target)

        Pearson's Chi-squared test with Yates' continuity correction

data:  d$Getting.interest.in.the.subject and d$Target
X-squared = 1.1884, df = 1, p-value = 0.2757
```

```
> chisq.test(d$Parents.preference,d$Target)

        Pearson's Chi-squared test with Yates' continuity correction

data:  d$Parents.preference and d$Target
X-squared = 1.2401, df = 1, p-value = 0.2655

> chisq.test(d$Own.decision,d$Target)

        Pearson's Chi-squared test with Yates' continuity correction

data:  d$Own.decision and d$Target
X-squared = 7.0417, df = 1, p-value = 0.007963

> chisq.test(d$Teacher.s.guidance,d$Target)

        Pearson's Chi-squared test with Yates' continuity correction

data:  d$Teacher.s.guidance and d$Target
X-squared = 6.5041, df = 1, p-value = 0.01076

> chisq.test(d$Friend.s.or.relative.s.influence,d$Target)

        Pearson's Chi-squared test with Yates' continuity correction

data:  d$Friend.s.or.relative.s.influence and d$Target
X-squared = 0, df = 1, p-value = 1

> chisq.test(d$Job.Opportunity,d$Target)

        Pearson's Chi-squared test with Yates' continuity correction

data:  d$Job.Opportunity and d$Target
X-squared = 11.863, df = 1, p-value = 0.0005725

> chisq.test(d$Not.get.any.chance.for.other.course,d$Target)

        Pearson's Chi-squared test with Yates' continuity correction

data:  d$Not.get.any.chance.for.other.course and d$Target
X-squared = 3.5877, df = 1, p-value = 0.05821

> chisq.test(d$Sex,d$Target)

        Pearson's Chi-squared test with Yates' continuity correction

data:  d$Sex and d$Target
X-squared = 2.1916, df = 1, p-value = 0.1388

> chisq.test(d$Father.s.education,d$Target)

        Pearson's Chi-squared test

data:  d$Father.s.education and d$Target
X-squared = 9.8328, df = 3, p-value = 0.02004

> chisq.test(d$Mother.s.education,d$Target)

        Pearson's Chi-squared test

data:  d$Mother.s.education and d$Target
X-squared = 6.5344, df = 3, p-value = 0.08832
```

```
> chisq.test(d$Family.income..monthly.in.Rs..,d$Target)

        Pearson's Chi-squared test

data:  d$Family.income..monthly.in.Rs.. and d$Target
X-squared = 11.581, df = 3, p-value = 0.008966
```

```
>
> glmfit=glm(Target~Avail.in.HS+Getting.interest.in.the.subject+Parents.preference+Own.decision+Teacher.s.guidance+Friend.s.or.relative.s.influence+Job.Opportunity+Not.get.any.chance.for.other.course+Sex+Father.s.education+Mother.s.education+Family.income..monthly.in.Rs..,data=d,family=binomial)
> summary(glmfit)

Call:
glm(formula = Target ~ Avail.in.HS + Getting.interest.in.the.subject +
    Parents.preference + Own.decision + Teacher.s.guidance +
    Friend.s.or.relative.s.influence + Job.Opportunity + Not.get.any.chance.for.other.course +
    Sex + Father.s.education + Mother.s.education + Family.income..monthly.in.Rs..,
    family = binomial, data = d)

Coefficients:
                                 Estimate Std. Error z value Pr(>|z|)
(Intercept)                       4.46821    2.57307   1.737   0.0825 .
Avail.in.HS                      -5.22354    0.81299  -6.425 1.32e-10 ***
Getting.interest.in.the.subject   0.32194    2.22907   0.144   0.8852
Parents.preference               -2.48777    1.14034  -2.182   0.0291 *
Own.decision                     -1.98451    1.08315  -1.832   0.0669 .
Teacher.s.guidance                1.24399    0.91048   1.366   0.1718
Friend.s.or.relative.s.influence  0.21111    0.96375   0.219   0.8266
Job.Opportunity                   0.37471    0.66852   0.561   0.5751
Not.get.any.chance.for.other.course -0.68372  0.90409  -0.756   0.4495
Sex                              -0.61074    0.79762  -0.766   0.4439
Father.s.education                0.18001    0.62720   0.287   0.7741
Mother.s.education                0.46728    0.82151   0.569   0.5695
Family.income..monthly.in.Rs..    0.04829    0.48303   0.100   0.9204
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 207.944  on 149  degrees of freedom
Residual deviance:  73.548  on 137  degrees of freedom
AIC: 99.548

Number of Fisher Scoring iterations: 6

>
```

```
>
> f_model <- lm(Target ~ ., data = d)
> nu_model <- lm(Target~1,data=d)
> library(MASS, lib.loc = "C:/Program Files/R/R-4.3.0/library")
> summary(stepAIC(nu_model,direction = "forward",scope =list(upper=f_model,lower = nu_model)))
Start:  AIC=-205.94
Target ~ 1

                                    Df Sum of Sq    RSS     AIC
+ Avail.in.HS                        1   24.9665 12.534 -368.33
+ Job.Opportunity                    1    3.2550 34.245 -217.56
+ Own.decision                       1    2.0417 35.458 -212.34
+ Teacher.s.guidance                 1    1.9083 35.592 -211.78
+ Family.income..monthly.in.Rs..     1    1.6239 35.876 -210.58
+ Father.s.education                  1    1.4494 36.051 -209.86
+ Mother.s.education                  1    1.4373 36.063 -209.81
+ Not.get.any.chance.for.other.course 1    1.0853 36.415 -208.35
+ Sex                                1    0.6764 36.824 -206.67
+ Getting.interest.in.the.subject    1    0.5282 36.972 -206.07
<none>                                           37.500 -205.94
+ Parents.preference                 1    0.4464 37.054 -205.74
+ Friend.s.or.relative.s.influence   1    0.0128 37.487 -204.00

Step:  AIC=-368.33
Target ~ Avail.in.HS
```

```
                                    Df Sum of Sq    RSS     AIC
+ Teacher.s.guidance                 1   0.58457 11.949 -373.50
+ Father.s.education                 1   0.46724 12.066 -372.03
+ Mother.s.education                 1   0.34009 12.194 -370.46
+ Own.decision                       1   0.32338 12.210 -370.25
+ Sex                                1   0.20473 12.329 -368.80
+ Family.income..monthly.in.Rs..     1   0.20028 12.333 -368.75
+ Not.get.any.chance.for.other.course 1   0.17789 12.356 -368.48
<none>                                       12.534 -368.33
+ Job.Opportunity                    1   0.14521 12.388 -368.08
+ Getting.interest.in.the.subject    1   0.05207 12.482 -366.96
+ Friend.s.or.relative.s.influence   1   0.02259 12.511 -366.60
+ Parents.preference                 1   0.01314 12.520 -366.49

Step:  AIC=-373.5
Target ~ Avail.in.HS + Teacher.s.guidance

                                    Df Sum of Sq    RSS     AIC
+ Mother.s.education                 1   0.31865 11.630 -375.55
+ Father.s.education                 1   0.26853 11.680 -374.91
+ Not.get.any.chance.for.other.course 1   0.17076 11.778 -373.66
<none>                                       11.949 -373.50
+ Own.decision                       1   0.14109 11.808 -373.28
+ Family.income..monthly.in.Rs..     1   0.11175 11.837 -372.91
+ Sex                                1   0.08526 11.864 -372.57
+ Job.Opportunity                    1   0.07788 11.871 -372.48
+ Parents.preference                 1   0.03529 11.914 -371.94
+ Getting.interest.in.the.subject    1   0.02325 11.926 -371.79
+ Friend.s.or.relative.s.influence   1   0.00084 11.948 -371.51

Step:  AIC=-375.55
Target ~ Avail.in.HS + Teacher.s.guidance + Mother.s.education
```

```
                                    Df Sum of Sq    RSS     AIC
<none>                                       11.630 -375.55
+ Parents.preference                 1  0.142013 11.488 -375.40
+ Not.get.any.chance.for.other.course 1  0.087253 11.543 -374.68
+ Sex                                1  0.061294 11.569 -374.35
+ Own.decision                       1  0.042605 11.588 -374.10
+ Job.Opportunity                    1  0.024504 11.606 -373.87
+ Father.s.education                 1  0.017228 11.613 -373.78
+ Getting.interest.in.the.subject    1  0.007086 11.623 -373.64
+ Friend.s.or.relative.s.influence   1  0.002003 11.628 -373.58
+ Family.income..monthly.in.Rs..     1  0.000094 11.630 -373.55

Call:
lm(formula = Target ~ Avail.in.HS + Teacher.s.guidance + Mother.s.education,
    data = d)

Residuals:
     Min       1Q   Median       3Q      Max
-1.00856 -0.08056 -0.08056  0.12788  0.91944

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)        0.87212    0.04007  21.762  < 2e-16 ***
Avail.in.HS       -0.79155    0.04694 -16.862  < 2e-16 ***
Teacher.s.guidance  0.16079    0.06047   2.659  0.00872 **
Mother.s.education  0.06822    0.03411   2.000  0.04735 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2822 on 146 degrees of freedom
Multiple R-squared:  0.6899,    Adjusted R-squared:  0.6835
F-statistic: 108.3 on 3 and 146 DF,  p-value: < 2.2e-16

>
```

REFERENCES

1. https://www.wikipedia.org/
2. https://www.investopedia.com/
3. https://www.geeksforgeeks.org/
4. https://www.youtube.com/
5. https://docs.google.com/forms/
6. An Introduction to Probability and Statistics – Rohatgi and Saleh
7. Fundamentals of Statistics (Volume One) – Gun, Gupta, Dasgupta