MODELLING A MACHINE LEARNING CLASSIFIER FOR PREDICTING STUDENT'S ENTREPRENEURIAL INTENTIONS

Mr. S.N. Vivek Raj

Research Scholar, College of Management, SRM Institute of Science and Technology,
Kattankulathur- 603203, Chengalpattu District, Tamil Nadu, India
viveksct@gmail.com, vn3984@srmist.edu.in

Dr. S.K. Manivannan

Associate Professor, College of Management, SRM Institute of Science and Technology,
Kattankulathur-- 603203, Chengalpattu District, Tamil Nadu, India
manivans@srmist.edu.in

Abstract

Entrepreneurship study and research is an important aspect in booming economy like India because entrepreneurs provide necessary impetus and act as a catalyst in bringing economic growth. In this study we have designed a logistic regression based predictive classification algorithm to predict entrepreneurship affinity to start a new business. The Logistic Model uses a binary categorical variable i.e., interest towards starting a business in near future (Sure or Not Sure) as dependent variable and socio-economic factors, behavioural traits i.e., Risk bearing capacity, Creativity, Decision Making capacity, Leadership ability, Ease of Communication, Self-confidence, and Willingness to enter unfamiliar territory as independent variables. The Model also compares various feature selection methods in improving predictive accuracy. Data is collected from 321 students using a structured questionnaire and model predicts the significant factors that impact the entrepreneurship decision and probability of students having positive attitude towards starting a new business. Findings from the study revealed that the Gender of the respondents, attitude towards entrepreneurship and risk bearing capacity as the significant factors impacting the student's intent towards starting a new business in near future.

Keywords: Predictive Model, Classification Algorithm, Logistic Regression, Feature Selection, Entrepreneurship, Machine Learning.

1. Introduction to study

Entrepreneurship is often linked with economic growth of the state and hence more measures must be taken by the government and concerned stake holders to boost entrepreneurship intent in any growing economy [7]. Before starting this paper, it is to fair to be noted that entrepreneurship is not only economically beneficial, but also socially beneficial to the country in many ways [13]. There are many instances where entrepreneurs have worked towards social up-lifting of the downtrodden. To add further Entrepreneurs who are engaging into the production of high-end products can be a great value to the country by creating skilled job and labour [6]. Entrepreneurship is not only about profits it is also about empowerment, especially women can be empowered if they are involved home based ventures [2].

Given the importance of his entrepreneurship which is beneficial to the social and economic transformation of the country, this study has been conducted to identify the significant predictors impacting the entrepreneurial intent among the students. Binary Logistic regression is used to predict the entrepreneurship intent among the students since the entrepreneurial intent is coded as binary variable i.e., Sure or Not sure to venture into a business in near future[11].

The present study contributes in different ways, the first one is important because the study probes the attitude of students about entrepreneurship by relying on the primary data collected from the students, thus contributing vital data for fostering entrepreneurship among young students. Next the study contributes to understand the application of one of the classical machine learning algorithm logistic regression in predicting the intentions of students and

selection of prominent features impacting the study. Finally, this study compares various feature selection methodologies in logistic regression and there by finding out the differences between the approaches.
The Study is vital since the study contributes in two significant problem domains one is entrepreneurship study and the other is application of machine learning i.e. classification algorithm. This research aims to achieve significant inroads in entrepreneurship by identifying significant factors affecting the student's entrepreneurial intent. The results thus would contribute significantly to the domain of entrepreneurship and would be a boost for the country's economic development.

2. Review of literature

There have been prominent studies regarding use of Machine Learning algorithms to predict and classify respondent's attitude and intent. Liang, Yang, Chen, & Chung, [8] have used the logistic regression analysis to understand the attitude of respondents who consume organic food towards specific promotional programs and found that the consumers have likings towards some of the promotional programs like discounts and free give away. Aguirre [1] has applied Standardized ordinal regression on a random sample of 370 respondents to identify the predictors of coffee culture among the students in a private university in Coasta Rica. Cicatiello, De Rosa, Franco, & Lacetera,[5] has used logistic regression-based classifier to analyze the attitude of consumer in accepting insects as food by using a exploratory study among 200 consumers above 14 years of age and found that education as one of the significant variable affecting consumer willingness to try insect based food.
Rahman, Khanam, & Nghiem,[12] has constructed a logistic based regression model to predict the impact of micro finance in women empowerment in Bangladesh by properly accounting selection bias in their model. Bekoe, Owusu, Ofori, Essel-Anderson, & Welbeck [4] have conducted a Cross section study among 457 students using binary logistic regression to predict the attitude of students towards accounting profession and their willingness to follow an accounting degree. It was found that previous exposure to accounting and aspiration to accounting qualification are the important variable affecting the student's intent.

Aluko, Daniel, Shamsideen Oshodi, Aigbavboa, & Abisuga,[3] compared the prediction accuracy of Support vector machine and logistic regression using a sample size of 102 architecture students for predicting the performance of students in academics and found that Support vector machine classifier is better than logistic regression in predicting students' academic performance. Umer, Susnjak, Mathrani, & Suriadi,[14] have used three Machine learning classification algorithms to predict the performance of students by recording their weekly performance in MOOCS environment. The Study revealed that combining process mining with the standard machine learning techniques improved the accuracy of the models.
Peker, Kocyigit, & Eren,[10] have applied five machine learning algorithms logistic regression, support vector machine, neural networks, random forest and decision trees to analyze 2 million purchase transactions to predict the purchase behaviour of individual customers. The Results showed that a proposed hybrid approach performs better than the individual and segment level approaches.
Owusu, Obeng, Ofori, Ossei Kwakye, & Bekoe,[9] have used a cross sectional study among 641 Ghanian students to predict the factors affecting students' attitude to pursue a professional qualification in accounting using a logistic regression model and found that students preferences and beliefs is one of the significant factors affecting student's intention. Vohra & Soni,[15]have used Logistic regression model to predict the factors affecting the behaviour of children in retail stores by collecting data from 473 mothers of children and the results verified that frequency of visit, age of parents, education of father as the significant determinants of children behaviour. Following the cue from most of the studies, Logistic regression analysis is used in this study to predict the student's attitude.

3. Research methodology

Logistic Regression model has been deployed to predict the entrepreneurial affinity of the students because the dependent variable in the study is coded as binary i.e (Sure or Not Sure). A Structured electronic questionnaire has been framed and it is distributed to students online. A total of 321 students attended the survey using convenience sampling method. Logistic Regression model can be framed as follows. Vohra & Soni,[15]

$$Xi = ln(Pi/(1 - Pi) = f\,(Y1, Y2, Y3, Y4, Y5 \ldots\ldots Yi)\; where, i\; 1,2,3 \ldots . k \qquad (1)$$

Xi=Binary Dependent Variable
Yi = Dependent Variable both categorical and continuous
Pi= Likelihood of occurrence of an event
1-Pi = Likelihood that the event would not occur.

The independent variables of the study include both the socio-economic status of the students and personality traits measured through a five-point Likert scale. The Summary independent variables are given in Table1.

Binary Logistic Regression Model is formed with a total of 15 independent variables out of which Age, Gender, Education Level, Discipline Level, Family Support, Student's attitude towards entrepreneurship as a good career option, Student's opinion towards entrepreneurship comparing it to a Job, Student's opinion whether entrepreneurship fits only for family entrepreneurs are categorical variables. Personality Traits of the Respondents like Risk bearing capacity, Creativity.
Decision Making Capacity, Leadership ability, Ease of Communication, Self Confidence
Willingness to enter unfamiliar territory are coded as continuous variable. The Scales of independent variables are explained in Table 3. The Personality traits of the students are tested for reliability using Cronbach's alpha and test revealed that the input data is highly reliable. The Reliability test data is given in Table 2.

Sl.No	Independent Variables	Type
1	Age	Categorical
2	Gender	Categorical
3	Education Level	Categorical
4	Discipline of Study	Categorical
5	Family Support towards career choices	Categorical
6	Students attitude towards entrepreneurship as a good career option	Categorical
7	Students' opinion towards entrepreneurship comparing it to a Job	Categorical
8	Students' opinion whether entrepreneurship fits only for family entrepreneurs	Categorical
9	Risk bearing capacity	Continuous
10	Creativity	Continuous
11	Decision Making Capacity	Continuous
12	Leadership ability	Continuous
13	Ease of Communication	Continuous
14	Self Confidence	Continuous
15	Willingness to enter unfamiliar territory	Continuous

Table 1. Summary of Independent Variables

Cronbach's Alpha	N of Items
.929	7

Table 2. Reliability Statistics

4. Data analysis

First Step in Logistic Regression modeling is the formation of null model. Null Model is framed only with intercept and with no independent predictor variables. The Table 4 gives the classification rate of the null model. If the final model is better than the null model it should give more accuracy in classification rate compared to the null model. Classification rate of the null model is 70.1 % but the problem is that it cannot predict the students who are not sure to venture to entrepreneurship.
The next step is formation of the model with all predictor variables. Before proceeding to the diagnostics of the model it is important to test whether the model to be formed is significant. The overall fit of the model is tested using omni bust test for coefficients which is given in Table 5. The model is found to be significant at α=0.01 since P Value of the Model is 0.000. Thus, the final model is significantly better than the null model.
Table 6 provides data about the classification rate of the final model. The Classification rate of the final model is 74.1 % i.e., the model correctly predicts the intent of the students towards starting a new business 74.1 percentage of times. There are significant improvements compared to the null model , first improvement is that classification percentage has increased from 70.1 to 74.1 % .Secondly the null model did not even classify a single student in the class 'Not sure' class, whereas the final model has classified 30.2 % of the students in the class 'Not sure'.
The next important diagnostic in any regression model is R squared value. R squared value provides the percentage variation in dependent variable explained by the predictor variables. Unfortunately, in Logistic regression model R squared value cannot be used directly and hence we use pseudo-R squared values to substitute R squared predictions. The Pseudo R squared values are given in the table 7. Percentage of variation in dependent variable explained by the independent variables ranges from 9.6 to 13.6 %.
Final Logistic regression model explaining the odds of students having positive intent towards entrepreneurship is given in Table 8. Of all the variables Gender(Female), Student's attitude towards entrepreneurship as a good career option(1), Student's opinion towards entrepreneurship comparing it to a Job(1), Risk bearing capacity are found to be significant variables.

Sl.no	Categorical Variables		Frequency	Parameter coding		
				(1)	(2)	(3)
1	Discipline	Arts and Science	13	1	0	0
		Engineering	138	0	1	0
		Management	164	0	0	1
		Others	6	0	0	0
2	Age	20-25	260	1	0	0
		26-30	13	0	1	0
		Greater than 30	6	0	0	1
		Less than 20	42	0	0	0
3	Entrepreneurship fits only for Family Business	Agree	83	1	0	
		Disagree	181	0	1	
		Neutral	57	0	0	
4	Gender	Female	119	1		
		Male	202	0		
5	Entrepreneurship is a Good Career Option	Not Sure	65	1		
		Sure	256	0		
6	Entrepreneurship is good doing a as doing a Job	No	68	1		
		Yes	253	0		
7	My Family supports my career choice	Not Sure	65	1		
		Yes	256	0		
8	Education	Less than PG	141	1		
		Post-Graduation	180	0		

Table 3. Summary of Categorical variable codes

Observed		Predicted		
		Venturing into new business		Percentage Correct
		Not Sure	Yes	
Venturing into new business	Not Sure	0	96	.0
	Yes	0	225	100.0
Overall Percentage				70.1

Table 4. Classification Percentage for Null Model

		Chi-square	df	Sig.
Step 1	Step	50.956	20	.000
	Block	50.956	20	.000
	Model	50.956	20	.000

Table 5. Omnibus Test for coefficients

Observed		Predicted		
		Venturing into new business		
		Not Sure	Yes	Percentage Correct
Venturing into new business	Not Sure	29	67	30.2
	Yes	16	209	92.9
Overall Percentage				74.1

Table 6. Classification Rate of Final Model

Model Summary			
Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	365.258a	.079	.112
2	359.368a	.096	.136

Table 7. Pseudo R squared values

Sl.No	Variables	B	Sig.	Exp(B)
1	Age		.799	
2	Age (1)	-.081	.851	.922
3	Age (2)	.252	.769	1.287
4	Age (3)	.990	.421	2.692
5	Gender (1)	-.508	.078*	.602
6	Education Level (1)	.179	.728	1.196
7	Discipline of Study		.672	
8	Discipline of Study (1)	-.117	.923	.890
9	Discipline of Study (2)	-.852	.421	.427
10	Discipline of Study (3)	-.360	.738	.697
11	Family Support (1)	-.398	.222	.672
12	Students attitude towards entrepreneurship as a good career option (1)	-1.200	.000***	.301
13	Students' opinion towards entrepreneurship comparing it to a Job (1)	-.694	.038**	.500
14	Students' opinion whether entrepreneurship fits only for family entrepreneurs		.714	
15	Students' opinion whether entrepreneurship fits only for family entrepreneurs (1)	-.266	.528	.767
16	Students' opinion whether entrepreneurship fits only for family entrepreneurs (2)	-.013	.973	.987
17	Risk bearing capacity	.323	.026**	1.381
18	Creativity	-.078	.669	.925
19	Decision making capacity	-.162	.461	.850
20	Leadership ability	.270	.161	1.310
21	Ease of communication	.114	.537	1.121
22	Self Confidence	-.298	.129	.742
23	Willingness to enter unfamiliar territory	.127	.328	1.135
24	Constant	1.337	.284	3.809

*, **, *** Significant variables with α=0.1, 0.05, 0.01 respectively

Table 8. Logistic Regression Model Explaining the Students Entrepreneurial Intention

5. Comparison of feature selection methods

Feature selection methods are approaches of selecting significant features or predictors impacting the dependent variable. In this study two such feature selection methods are compared. One is the Forward stepwise selection and other is backward selection. Both the methods select the prominent features based on Wald criteria. Overall classification rate of the forward step wise selection is 73.5 % whereas the classification rate of the backward stepwise regression is 73.2 %.

Although there is no significant difference between the classification percentages of both the methods, the main difference arises in the feature selection. The Forward selection method has identified only two significant features in the final model i.e., Risk bearing capacity and attitude of students i.e. Whether entrepreneurship is a good career option. The Backward selection method is able to find four different features Gender, Risk bearing capacity and two attitude measuring features. While comparing both methods forward step wise selection method based on Wald criteria has slightly outperformed the backward selection in terms of the classification percentage, but the backward selection method has clearly outperformed the forward selection in terms of feature selection.

Sl.No	Feature Selection Method	Criteria	Classification Percentage	No of features	Features Selected
1	Forward Selection	Wald	73.5 %	2	Risk bearing capacity and attitude of students i.e. Whether entrepreneurship is a good career option
2	Backward Selection	Wald	73.2 %	4	Gender, Risk bearing capacity and two attitude measuring features

Table 9. Comparison of feature selection Methods

6. Findings from the study

Table 8 implies that there exists a significant negative relationship between student's attitude towards entrepreneurship and affinity towards venturing into a new business. It is clear from the study that students having negative attitude towards entrepreneurship have less chances of starting a new business. Secondly, there exists a negative relationship between Gender of the student and Entrepreneurial affinity at α=0.10. Female students have less affinity towards starting a new venture than the male students. Female students are 0.602 times unlikely to start a new business than the male students. Another variable which has significant impact towards entrepreneurial intent of students is student's opinion towards entrepreneurship comparing it to a Job. Students who consider entrepreneurship is not good as doing a job in a reputed organization is 0.5 time unlikely to start a new business. Of all the personality traits of the students, the Primary characteristic that is significantly influencing the entrepreneurship intent of the student is Risk Bearing Capacity. There exists a positive relationship between risk bearing capacity of the students and entrepreneurial affinity at 1% level of significance. i.e., Student who possess risk bearing trait is 1.381 times more likely to start a new business compared to who do not possess. Table 6 provides the classification rate of the final model created. It is to be noted that the classification percentage of the final model has improved from 70.1 to 74.1% comparing with the null model and the final model able to predict that 30.2 % students are not sure to venture into a new business while the null model has not even predicted even a single student having negative affinity. The Final model is clearly a better classifier compared to the null model.

7. Suggestions from the study

The Study used logistic based regression for predicting student's intention towards entrepreneurship as binary logistic regression was the method of choice for many researchers. The Logit models clearly explain the significant variables that are key determinants of student's intent towards entrepreneurship. Female students have found to be having less odd of having positive intent towards entrepreneurship. Educational institutions and Government bodies can take actions to promote entrepreneurship among girl students this would be a key aspect in promoting the nation's economy and empowerment of women. Another important take away from this study is that students who have negative attitude that entrepreneurship is not a good career option are having less odds of having positive intent towards starting a business. As students are future leaders of the country, government, colleges, and other stake holders must take necessary steps in building positive image about entrepreneurship. Moreover, colleges can introduce dedicated courses on entrepreneurship in helping students having positive intent towards entrepreneurship. Of all the personality traits, Risk bearing capacity is found to be the predominant predictor of entrepreneurial intent. This trait must be inculcated in students from early childhood.

8. Scope for further Research

There is further scope of research in the following areas. The present only employed logit-based classifier for predicting student's intent towards starting a new business. Further research can include more classification algorithms like naive bayes, decision trees, support vector machine, neural networks, random forest, and more sophisticated techniques. Secondly the present study compared only step wise feature selection methods, Future research can apply an exhaustive list of feature selection methods.

References

[1] Aguirre, J. (2017). A new coffee culture amongst Costa Rican university students. British Food Journal, 119(12), 2918–2931. https://doi.org/10.1108/BFJ-12-2016-0614

[2] Al-Dajani, H., & Marlow, S. (2013). Empowerment and entrepreneurship: a theoretical framework. International Journal of Entrepreneurial Behavior & Research, 19(5), 503–524. https://doi.org/10.1108/IJEBR-10-2011-0138

[3] Aluko, R. O., Daniel, E. I., Shamsideen Oshodi, O., Aigbavboa, C. O., & Abisuga, A. O. (2018). Towards reliable prediction of academic performance of architecture students using data mining techniques. Journal of Engineering, Design and Technology, 16(3), 385–397. https://doi.org/10.1108/JEDT-08-2017-0081

[4] Bekoe, R. A., Owusu, G. M. Y., Ofori, C. G., Essel-Anderson, A., & Welbeck, E. E. (n.d.). Attitudes towards accounting and intention to major in accounting: a logistic regression analysis. Journal of Accounting in Emerging Economies, 8(4), 459-475. https://doi.org/10.1108/JAEE-01-2018-0006

[5] Cicatiello, C., De Rosa, B., Franco, S., & Lacetera, N. (2016). Consumer approach to insects as food: barriers and potential for consumption in Italy. British Food Journal, 118(9), 2271–2286. https://doi.org/10.1108/BFJ-01-2016-0015

[6] Edoho, F. M. (2015). Entrepreneurship paradigm and economic renaissance in Africa. African Journal of Economic and Management Studies, 6(1), 2–16. https://doi.org/10.1108/AJEMS-11-2014-0086

[7] Hafer, R. W. (2013). Entrepreneurship and state economic growth. Journal of Entrepreneurship nd Public Policy, 2(1), 67–79. https://doi.org/10.1108/20452101311318684

[8] Liang, A. R.-D., Yang, W., Chen, D.-J., & Chung, Y.-F. (2017). The effect of sales promotions on consumers' organic food response: An application of logistic regression model. British Food Journal, 119(6), 1247–1262. https://doi.org/10.1108/BFJ-06-2016-0238

[9] Owusu, G. M. Y., Obeng, V. A., Ofori, C. G., Ossei Kwakye, T., & Bekoe, R. A. (2018). What explains student's intentions to pursue a certified professional accountancy qualification? Meditari Accountancy Research, 26(2), 284–304. https://doi.org/10.1108/MEDAR-06-2016-0065

[10] Peker, S., Kocyigit, A., & Eren, P. E. (2017). A hybrid approach for predicting customers' individual purchase behavior. Kybernetes, 46(10), 1614–1631. https://doi.org/10.1108/K-05-2017-0164

[11] Peng, C.-Y. J., Lee, K. L., & Ingersoll, G. M. (2002). An Introduction to Logistic Regression Analysis and Reporting. The Journal of Educational Research, 96(1), 3–14. https://doi.org/10.1080/00220670209598786

[12] Rahman, M. M., Khanam, R., & Nghiem, S. (2017). The effects of microfinance on women's empowerment: new evidence from Bangladesh. International Journal of Social Economics, 44(12), 1745–1757. https://doi.org/10.1108/IJSE-02-2016-0070

[13] Thompson, J. L. (1999). The world of the entrepreneur – a new perspective. Journal of Workplace Learning, 11(6), 209–224. https://doi.org/10.1108/13665629910284990

[14] Umer, R., Susnjak, T., Mathrani, A., & Suriadi, S. (2017). On predicting academic performance with process mining in learning analytics. Journal of Research in Innovative Teaching & Learning, 10(2), 160–176. https://doi.org/10.1108/JRIT-09-2017-0022

[15] Vohra, J., & Soni, P. (2015). Logit modelling of food shopping behaviour of children in retail stores. Management Research Review, 38(8), 840–854. https://doi.org/10.1108/MRR-03-2014-0061

Authors Profile

.

S.N.Vivek Raj , is pursuing is Ph.D.(Part-time) in College of Management , SRM Institute of Science and Technology, Kattankulathur, Tamil Nadu, India. He is currently working as Assistant Professor in Centre of Analytics , KCT Business School, Kumaraguru College of Technology, Coimbatore . He has presented in international conferences and published in many international journals of repute . His research areas include Machine Learning Applications , Predictive Analytics and Technology Adoption. He believes in continuous learning and has been doing a lot of professional certifications throughout his illustrious career. He has received many awards and achievements from various institutions and is a National level Topper in five courses conducted by various IIT's in NPTEL platform. He has also qualified in UGC-NET eligibility test for Assistant Professor. He has delivered many guest lectures in various spheres of business analytics and organized many MDP's and FDP's in evolving areas of business analytics .

.

Dr. S.K.Manivannan is currently working as , Associate Professor in College of Management , SRM Institute of Science and Technology, Kattankulathur, Tamil Nadu, India. He has done his PGDM from Indian Institute of Management , Ahmedabad, and Ph.D from SRMIST , Kattankulathur. He has also qualified in UGC- NET examination in the year 2014. He has published many journals in many reputed international journals . Before Joining Academics, he has over 23 years of rich corporate experience as a practicing senior Manager in National and Multinational corporate in multiple domains . He has delivered many invited lectures and he is a Board of Studies member for many colleges. He was given Silver Award for Quality Circle Implementation in Teaching from Quality Circle Federation of India, Chennai Circle in the year 2014. He as an Industry Interaction Coordinator arranged multiple types of interactions with industry in the forms of MDPs, Guest lectures, Project valuations as well as Internships. He is recognized research supervisor in SRMIST, and he is currently guiding five research scholars in emerging domains of management research. He has also designed and implemented many new courses such TQM Certified Teachers program at SRM University which was designed and implanted in association with Tokai University, Japan.