The Factor Drive People Buy Travel Insurance

Yuqing Dai(✉)

W. P. Carey School of Business, Arizona State University, Tempe 85281, USA
kotolidai@gmail.com

Abstract. In the century of the rise of the travel industry, many related industries are also on the rise, and the insurance industry is one of them. The analysis of tourists or customers to predict or determine whether they will buy insurance is the general context of this paper. Predictions can be found everywhere in life. This article focuses on travel insurance forecasting. Eight non-potential factors are used as independent variables to examine the influence on whether to purchase travel insurance. The paper starts with comparative analysis, factor analysis, and logistic regression for analysis as well as prediction. It was found that there was a relatively high positive correlation between whether a traveler purchased travel insurance and the traveler's historical flight factors. Whether or not the traveler had flown abroad was also a significant influencing factor.

Keywords: Insurance · non-potential factors · regression

1 Introduction

Since the beginning of 2020, the world has been affected by COVID-19 and millions of people have died from the pandemic. The COVID-19 has had an impact on the economy, people's quality of life, and even politics. Helliwell et al. reported happiness in 2021 based on people's responses to quality of life, economic insecurity, and anxiety in 2021 reflecting severe social welfare losses and a lack of happiness [1]. Even without the effect of COVID-19, people's happiness index is still in a state of decline in recent years. This is due to work stress, anxiety about quality of life, etc. In this situation, the need to improve self-well-being is a trend topic within people.

Kwon et al. found that Subjective well-being had risen 15 days before travel and lasted for about 1 month after travel [2]. Due to this reason, travel will become the norm and the purchase of travel insurance will become an option for the client, and also will become an opportunity for the insurance company, so it is necessary to analyze and forecast whether the client buy travel insurance.

In recent years, there has been no shortage of research on travel insurance, for example, insurance claim forecasting, purchase forecasting, etc. Natural disasters or man-made terrorist attacks can be a travel risk. And travel insurance is the way to reduce the risk. In their study, Igor Sarman et al., found that underlying structures (perceptions, motivations, emotions, attitudes, perceptions) also influence the willingness of individuals to purchase travel insurance. The model they propose are more psychological in nature and predict user behavior at a deeper level [3].

J. Yen et al. (Eds.): ICBIS 2023, AHCS 14, pp. 217–227, 2023.
https://doi.org/10.2991/978-94-6463-198-2_25

In this study, the goal of research was to analyze the influence of non-potential factors of individuals on the purchase of travel insurance. Logistic regression is often used for travel related topics, such as travel insurance claims. Dadang Amir Hamzah created logistic regression from travel insurance claims data as well as features. And used to generate predicted claims data which prediction results are close to the actual data [4]. Li et al., further using the cluster-based logistic regression model to analyses the way to travel[5]. The prediction of travel behavior is also a relatively well known area of research. Sönmez et al., not only did they use logistic regression, but they also used cross-tabulations to analyze people with travel experience to determine the likelihood of continuing to visit. They found that both perceived risk and safety were better predictors of avoidance areas than planning to visit them [6].

National policies also have an impact on whether people travel, especially when people travel abroad. Arita et al., applying a fixed-effects estimation model to analyses how Approved Destination Status(ADS), which is a negotiation result by China with 120 countries early 1990s, affected outbound tourist travel from China. They found that ADS has resulted in significant increases in arrivals from China, averaging 52% over three years based on various model [7].

Not only logistic regression, but also linear regression is often used as a way to predict travel demand or travel insurance. V. R. Rengaraju and V. Thamizh Arasan, used multiple linear regression to analyze calibration data, cross-validation and backward prediction methods to create a model to forecast future air travel demand [8].

Factor analysis is also often used in analytical as well as predictive work to assemble similar variables for further analysis. Chen et al., Using the factor analysis to provide clearer dimensions of travel motivation of Taiwanese senior, and divide to two dimension, psychological factors and Socio-demographic factors. They found external interactions have a different impact on the willingness of senior to travel domestically than they do abroad [9]. Joseph N.Prashker combined factor analysis and indscal to conduct a perception of urban travel mode choice [10].

In this paper, there are many reasonable and interesting phenomena have been found when analyzing insurance data for prediction. In addition to the traditional factor analysis and logistic regression-based analysis, this paper uses comparative analysis for a more thorough analysis. Many non-potential factors can influence or drive whether people buy travel insurance or not. This can shed some light on predictive models for the travel insurance industry. The paper in organize as follow: Sect. 2 the basic logistic regression analysis; Sect. 3 factor analysis to reorganize the variable; Sect. 4 further study about the crosstabs. Section 5 conclusion.

2 The Basic Logistic Regression Analysis

A travel company offers insurance packages to its customers. The data was obtained from Tejashvi, Kaggle database. The data describes the responses to the company's insurance offer from 2000 customers who received it in 2019. This response is used as the dependent variable in the data[11]. Some of the customer profiles and family information described in this data are age, employment type, graduate or not, annual income, family members, chronic diseases, frequent flyer and ever travelled abroad as

the dependent variable. Where chronic diseases is used as a nominal variable to represent whether the client suffers from a major disease, such as asthma, hypertension, which 0 means never have disease history, 1 means have one of these disease history. Frequent flyer describes customers who booked at least four times from 2017 to 2019, which 0 means not frequent, 1 means frequent. Employment type describes the sector in which the customer is employed, and this dependent variable is divided into two, 0 for private sector/self-employed, and 1 for government sector. The purpose is predicting whether a customer buys travel insurance through theses Independent variables. If the customer buy the travel insurance, the dependent variable "travel Insurance" mark as 1, if not, mark as 0.

This data contains a large number of discrete variable and a small number of continuous variables (age, annual income, family member). A logistic regression is the most basic approach to initially process this data. Whether this regression model is meaningful is the first step that take when conducting the analysis. The Table 1 denotes that the p-value of Omnibus Tests of Model Coefficients in the model row is less than 0.05, so the model is considered to be significant overall.

When observing at the Table 2 in the logistic regression model, it denotes that the model has a certain degree of explanatory power and accuracy in prediction, with an accuracy of 69.3%.

The model has a good predictive power. In Table 3, several variables that have significant effects on the dependent variable are known, age, annual income, family members, frequent flyer, and ever travelled abroad. However, the employment type, graduate or not, and Chronic Diseases variables did not have a particularly strong predictive effect on the model.

Aging has always been an issue in our current society, and as our social structure becomes more complete, people are more inclined to enjoy it. According to Newgard et al., it is predicted that by 2050, the number of people over 60 years old in the world

Table 1. Omnibus Tests of model Coefficients

	Chi-square	df	Sig.
Model	554.871	11	.000

Table 2. Classification Table

		Predicted Travel Insurance		Percentage correct
		0	1	
Observed Travel Insurance	0	886	391	66.8
	1	210	500	69.3
Overall Percentage				67.7

Table 3. Variables in the Equation

	Sig.
Age	.000
Annual Income	.000
Family Members	.000
Frequent Flyer	.001
Ever Travelled Abroad	.000

will reach more than 2 billion, accounting for 22% of the total population [12]. This denotes that there is some relevance to this issue in terms of travel insurance projections. Not just because the aging population is rising and therefore this variable has some predictive significance. Today's Gen-Z, as life becomes more stressful and there is a shift in thinking, they may be more willing to spend on themselves, which is one of the reasons for aging. Which spending on oneself is not only limited to buying travel insurance. Therefore, the aging of Gen-Z will have a greater impact on prediction model of travel insurance future.

Annual income as an important indicator of consumption level has always shown a correlation with the willingness to purchase goods. There is no exception in this model, which simply means that those who have money are more willing to spend it (buy travel insurance).

It is reasonable to assume that the number of family members has an impact on whether or not to purchase family insurance packages and such packages are less costly than single or few-person coverage. The most important point of purchasing psychology in the market is price, so in this prediction model, family members are actually linked to cost. This can be judge as an indicator of cost of buying travel insurance.

Frequent flyer can be simply explained as the frequency of flying. As declare above, if flying more than 4 times in 2 years, passenger are classified as frequent flyer, and if flying less than 4 times, passenger are classified as infrequent flyer. The frequency of flight is also reasonable in the interpretation of the model. On the other hand, the frequency of flight can be interpreted as the probability of accident, and the increase of accident probability will stimulate customers to buy travel insurance. This variable is similar in nature to the other variable "Ever travelled abroad", which can also be interpreted as the probability of an accident, but not in its entirety. In general, the distance traveled by international flights is greater than that of domestic flights, and the increase in distance also increases the probability of accidents. However, this is not a generalization, as international travel insurance policies may not provide the same coverage as domestic travel insurance policies, and this may also be a driving force for people who have traveled abroad to buy the travel insurance. Other factors may also be the policy, the international situation. All of these factors are likely to be drivers of whether someone with experience abroad will purchase travel insurance.

3 Factor Analysis to Reorganize the Variable

Factor analysis has good results in the integration of variables. Before performing factor analysis on data, it is necessary to make sure that there is some correlation in our data, and both the Kaiser-Meyer-Olkin (KMO) Measure of Sampling Adequacy and Bartlett's Test of Sphericity have been used in the research field as excellent test to determine whether the data are correlated or not. From Table 4, the KMO test result is 0.617, which is appropriate. p-value of Bartlett's test is less than 0.05. Overall, our data set is correlated and suitable for factor analysis.

Determining the interpretation of the common factor for each original variable can help us determine whether our common factor has good interpretation. Table 5 denotes that basically each of the original variables explains more than 0.7 of the common factors, which also indicates that the explanatory power of our common factor is very good.

For this dataset, there is a total of six common factors (a total of eight independent variables in the original data) was extracted, and the six original data had an explanatory value of 85.876% that shows in Table 6.

An interpretation of 85 percent is relatively an acceptable range within the research field. Table 7 denotes the first common factor explains mainly Frequent Flyer, ever travelled abroad and annual income after rotation. This common factor is consistent with the results determined in chapter 2, i.e., that the frequency of flying and the experience of flying abroad are essentially the same, but with the addition of annual income to the common factor, this can be also assumed that people with money will fly multiple times and fly internationally. Therefore, the first factor has interpreted as a historical flight factor. The rotation played a certain denoising role, and the remaining 5 factors

Table 4. KMO and Bartlett's Test

KMO test		.617
Bartlett's Test	Sig.	.000

Table 5. Communalities

Variable name	Extraction
Age	.987
Employment Type	.856
Graduate or not	.915
Annual Income	.731
Family Members	.999
Chronic Diseases	.991
Frequent Flyer	.803
Ever Travelled Abroad	.588

Table 6. Variance Explained

Component	Cumulative %
1	24.181
2	38.496
3	51.301
4	63.622
5	75.636
6	85.876

explained the independent variables more strongly: employment type, graduate or not, chronic diseases, age, and family, respectively.

It is also relatively meaningful to do logistic regression on these factors because there is an additional factor that is more meaningful: the historical flight factor. Table 8 shows logistic regression for the six common factors, some interesting phenomena emerged. The first is that our prediction accuracy improves by 1.3% (69.3 before) relative to the logistic regression model done previously.

Table 7. Rotated Component Matrix

	1	2	3	4	5	6
Frequent Flyer	.838					
Ever Travelled Abroad	.691					
Annual Income	.675					
Employment Type		.900				
Graduate or not			.949			
Chronic Diseases				.995		
Age					.991	
Family Members						.999

Table 8. Classification Table

		Predicted Travel Insurance		Percentage correct
		0	1	
Observed Travel Insurance	0	805	472	63.0
	1	209	501	70.6
Overall Percentage				65.7

Table 9. Variables in the Equation

	Sig.	Exp(B)
Historical Flight Factor	.000	2.762
Employment Type	.000	0.616
Graduate or not	.003	1.173
Age	.000	1.298
Family Members	.000	1.236

The second change is that the p-value of employment type is less than 0.05, reflecting the fact that this variable can explain the predictive model. The p-value of graduate or not becomes 0.03, slightly less than 0.05. The Not surprisingly, the historical flight factor had the highest Exp(B) at 2.762, reflecting the fact that people who fit this factor are about 2.8 times more likely to buy the travel insurance.

The Table 9 denotes that customer who work in government-related jobs (mark as 1) are less likely to buy travel insurance than those who work in other jobs (mark as 0). Government department workers may have government-provided travel insurance to the extent that they are less likely to purchase travel insurance from outside sources [13].

4 Further Study About the Cross-Tabulations

Cross-tabulations are frequently used in insurance forecasting related topics. For the discrete variables, the graphs after observing all the variables are as follows:

4.1 Discrete Variable

From the Table 10, it denotes that the independent variable "Chronic diseases", which was previously judged to be unrelated to the dependent variable in the logistic regression,

Table 10. Chi-Square Tests

	Asymptotic significance (2-sided)			
	PCS	CC	LR	LBLA
Employment Type	.000	.000	.000	.000
Graduate or not	.399	.436	.397	.399
Chronic Diseases	.417	.448	.418	.418
Frequent Flyer	.000	.000	.000	.000
Ever Travelled Abroad	.000	.000	.000	.000

Footer: PCS: Pearson chi-square, CC: Continuity Correction, LR: Likelihood Ratio, LBLA: Linear-by-Linear Association.

did not pass the test, proving once again that this dependent variable is not related to the independent variable. This is a reasonable explanation for the fact that passengers with specific diseases are not allowed to fly and airlines choose to refuse them. Therefore, it would not make sense for them to purchase travel insurance. Similarly, the independent variable "Graduate or not" also failed the test. In the logistic regression model, which has been done in chapter 3, the p-value for "Graduate or not" was 0.03, which is very close to 0.05. After doing the Chi-square test, this make more sense that this independent variable was not particularly related to the dependent variable.

The remaining discrete variables are all able to explain the dependent variable to some extent. Based on the cross-tabulation, The results derived the proportions of these independent variables in each situation. The number of people with an employment type of 0 is 1,417, 570 of them buy the travel insurance, which is 40.22% of them. In contrast, only 24.56% of those with employment type 1 buy travel insurance. This is consistent with our previous judgment. Government employees are even less likely to purchase travel insurance.

57.31% of frequent flyers would buy travel insurance and only 29.94% of infrequent flyers would buy travel insurance. 78.42% of those who have traveled abroad would purchase travel insurance, which is a very large number and a very strong independent variable. Conversely, only 25.64% of those who have not traveled abroad would purchase travel insurance. For a single discrete variable, it is easy to observe that people who had traveled abroad were more likely to purchase travel insurance.

4.2 Continuous Variables

In the present data, there are three sets of continuous variables. Age, annual income, and family member, respectively. After Kolmogorov-Smirnov analysis of travel insurance for these three data sets in Table 11, it was found that the p-value for all three data sets was less than 0.05, rejecting the assumption of normality.

Because of the rejection of the normality hypothesis, a nonparametric test for these three variables should be taken in the next step, the Hypothesis test in the nonparametric test rejected the original hypothesis that these three variables are equally distributed in travel insurance. This again demonstrates that these three continuous variables can influence the dependent variable "travel insurance".

Table 11. Tests of Normality

	Travel Insurance	Sig.
Age	0	.000
	1	.000
Annual Income	0	.000
	1	.000
Family Members	0	.000
	1	.000

Table 12. Crosstabulation

		Travel Insurance		Total
		0	1	
Family type	1	584	132	716
	2	205	94	299
	3	364	321	685
	4	124	163	287
Total		1277	710	1987

4.3 The Combination of Annual Income and Family Members

After conducting a comparative analysis of each variable, how the combination of annual income and number of household members would affect the dependent variable is a crucial question. Whether the behavior of purchasing travel insurance would change depending on different levels of consumption and different number of family members is also important. The Independent variable was named as family type, and it categorized low annual income and low family member as 1; low annual income and high family member as 2; high annual income and low family member as 3; and high annual income and high family member as 4. The classification of annual income and family member mainly uses the mean as an indicator. The better way to determine the effect of family type on the dependent variable is to perform a comparative analysis, using the cross-tabulation as before. This variable passed the Chi-Square test, and the Table 12 mainly shows specifically the proportion of each family type purchasing travel insurance.

According to the chart, 56.79% of families with family type 4 will buy travel insurance, while only 18.44% of families with family type 1 will buy travel insurance. This value decreases from family type 4 to family type 1, which family type 1 is the lowest. One of the more interesting things is that families with low incomes and large family sizes will be more likely to buy travel insurance than those with low incomes and low family sizes. This again explains the judgement in chapter 2 that with more family members, insurance companies may offer discounts, packages, etc. This also drives families with higher incomes and more family members to buy travel insurance because their annual income is higher.

5 Conclusion

This paper focuses on the analysis of the effect of different independent variables on the dependent variable and finds the extent to which different independent variables affect the dependent variable. Among these variables, there are also variables that do not explain the dependent variable "travel insurance", such as chronic diseases and graduate or not, and people with special diseases are not allowed to fly, which makes it meaningless for them to buy travel insurance. Graduate or not also means that the level of education is not a driving factor in the purchase of travel insurance. Among the

other driver variables in this paper, age shows that the aging population is concerned about their own safety, and based on the prediction in this paper, the Gen-Z ages, age will become a more stronger independent variable driving customers purchase insurance or not because of the shift in thinking and the increase in aging. Because government officials have government-provided benefits or workplace restrictions that cause them to not purchase travel insurance any more than people in other jobs. Annual income and the number of family members affect whether or not a customer buys travel insurance. Annual income is easier to explain why customers are willing to buy insurance, and for the variable of family member, It because that insurance companies offer more benefits or packages to families with more family members, which leads to such people being willing to buy insurance. In a further study, there are two variables has been combined into one and found that households with higher income and more family members were more willing to purchase travel insurance. Frequent flyer as well as ever travelled abroad as the two strongest variables in this data largely influenced the dependent variable. Having frequent flyer and ever travelled abroad experiences drive customers to purchase travel insurance, also because frequent frequency and long distance travel increases the probability of accidents and therefore increases the probability of purchasing travel insurance. For these two variables, the first factor (historical flight factor) in the factor analysis done explains most of these two variables as well as annual income. This factor was also found to have a very strong predictive power in the logistic analysis. The logistic regression model using the common factor after the factor analysis also had better predictive power.

References

1. Aknin, Lara B., et al. “Mental health during the first year of the COVID-19 pandemic: A review and recommendations for moving forward.“ Perspectives on Psychological Science (2021): 17456916211029964.
2. Kwon, Jangwook, and Hoon Lee. “Why travel prolongs happiness: Longitudinal analysis using a latent growth model.“ Tourism Management 76 (2020): 103944.
3. Sarman, Igor, Riccardo Curtale, and Homa Hajibaba. “Drivers of travel insurance purchase.“ Journal of Travel Research 59.3 (2020): 545-558.
4. Hamzah, Dadang Amir. “Predicting travel insurance policy claim using logistic regression.“ Applied Quantitative Analysis 1.1 (2021): 1-7.
5. Li, Juan, et al. “Cluster-based logistic regression model for holiday travel mode choice.“ Procedia Engineering 137 (2016): 729–737.
6. Sönmez, Sevil F., and Alan R. Graefe. “Determining future travel behavior from past travel experience and perceptions of risk and safety.“ Journal of travel research 37.2 (1998): 171-177.
7. Arita, Shawn, et al. “Impact of Approved Destination Status on Chinese travel abroad: an econometric analysis.“ Tourism Economics 17.5 (2011): 983–996.
8. Rengaraju, V. R., and V. Thamizh Arasan. “Modeling for air travel demand.“ Journal of Transportation Engineering 118.3 (1992): 371–380.
9. Ching-Fu, Chen, and Chine-Chiu Wu. “How motivations, constraints, and demographic factors predict seniors’ overseas travel propensity.“ Asia Pacific Management Review 14.3 (2009).
10. Prashker, Joseph N. “Scaling perceptions of reliability of urban travel modes using indscal and factor analysis methods.“ Transportation Research Part A: General 13.3 (1979): 203-212.

11. Tejashvi. “Travel Insurance Prediction Data.” Kaggle, 23 Aug. 2021, https://www.kaggle.com/datasets/tejashvi14/travel-insurance-prediction-data?resource=download.
12. Newgard, Christopher B., and Norman E. Sharpless. “Coming of age: molecular drivers of aging and therapeutic opportunities.“ The Journal of clinical investigation 123.3 (2013): 946-950.
13. “Employee Benefits.” Go Government, 28 Apr. 2022, https://gogovernment.org/all-about-government-jobs/employee-benefits/.