Applied Quantitative Analysis
Available online at: www.researchsynergy.org/quant
Volume **1** Number **1** (2021) Article **613**

Research

Predicting travel insurance policy claim using logistic regression

Dadang A. Hamzah[1,*], Averia A. Kalambe[1], Lucky S. Goklas[1], Naufal G. Alkhayyat[1]
[1] School of Business, President University, Jakarta, Indonesia

Abstract

This paper analyzes the characteristics that influence the travel insurance claim based on existing data records. Using logistic regression, the dependent variable is the feature that determines whether there is a claim or no claim. On the other hand, the independent variables are analyzed using exploratory data analysis to identify which characteristic has the highest correlation with the dependent variable. Based on selected features, the logistic regression model is created and used to generate the prediction claim data. The predicted data gives an excellent approximation to the actual data.

Keywords: *logistic regression; travel insurance; binary classification; classification; data science*

INTRODUCTION

Uncertainty always occurs in our daily lives, especially when traveling. Incidences like lost baggage, delayed flights, public transportation accident, food poisoning or sickness, injuries or health problems are common to occur. To minimize their losses, people consider travel insurance. According to Investopedia [1], travel insurance is a type of insurance that covers the costs and losses associated with traveling. Travel insurance is designed to cover insurer trips for a limited duration. The insurance company will make more money from the insurer that never claims. Mostly, the amount of claim is higher than the amount of premium paid by the insurer. Therefore, knowing the features that cause the occurrence of a claim is highly desired by insurance companies.

In this paper, we are interested in analyzing the features that influence the occurrence of a claim made by the insurer. Data are taken from Kaggle [2] with 11 attributes or columns and 63,326 entries. The target attribute is the claim status column that contains 'Yes' or 'No' as entries. We use the logistic regression model to predict the claim status column. In statistics, the logistic regression model is used to model the probability of an event that has binary outcomes such as pass/fail, win/lose, yes/no, healthy/sick, *etc*. Logistic regression was applied in various studies. In Weissert *et al.* [3], logistic regression was used to compare the mobile nursing system established during 1977-1985 in the USA. Binary logistic regression was also used to correlate retirement age with age, sex, economics, and social status [4]. Furthermore, logistic regression analysis was applied to classify companies' bankruptcy in America [5]. Then, the classification of automobile accidents in America was analyzed using logistic regression based on several criteria [6].

RESEARCH METHOD

This study follows the definition by Berenson *et al*. [7] on logistic regression. It is referred to as a model used to predict the probability of a particular categorical response for a given set of

* Corresponding author(s)
dadang.hamzah@president.ac.id (D.A.H.)

DOI: 10.31098/quant.613

independent variables. This model uses the odds ratio, representing the probability of an event of interest compared to not having an event of interest. The odds ratio is expressed as

$$odds\ ratio = \frac{Probability\ of\ an\ event\ of\ interest}{1 - Probability\ of\ an\ event\ of\ interest} \quad \text{Equation 1}$$

The logistic regression model is based on the natural logarithm (ln) of the odds ratio. That is, if there are independent variables, the logistic regression model is:

$$ln(odds\ ration) = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_k X_{ki} + \varepsilon_i \quad \text{Equation 2}$$

where $\beta_n, n = 1,2, \ldots, k \rightarrow$ regression coefficients
ε_i = random error in observation i

In application, finding the actual natural logarithm of odds ratio is not possible, therefore, the logistic regression is developed. This equation is developed using the method called *maximum likelihood estimation*. The logistic regression equation is written as

$$ln(Estimated\ odds\ ratio) = b_0 + b_1 X_{1i} + b_2 X_{2i} + \cdots + b_k X_{ki} \quad \text{Equation 3}$$

Once the logistic regression equation is determined, using the natural logarithm property, we have the estimated odds ratio as

$$Estimated\ odds\ ratio = exp(b_0 + b_1 X_{1i} + b_2 X_{2i} + \cdots + b_k X_{ki}) \quad \text{Equation 4}$$

Finally, we use Equation 4 to estimate the probability of an event of interest as

$$Estimated\ probability\ of\ an\ event\ of\ interest = \frac{Estimated\ odds\ ratio}{1 + Estimated\ odds\ ratio} \quad \text{Equation 5}$$

The value determined from Equation 5 is an estimation for the value in Equation 1. In practice, to find Equation 3, all observations in each independent variable X_{ki} must be numerical. Therefore, if possible, the observation that is not numerical must be converted into numerical or removed from the analysis. The binary classification will be determined based on the value of Equation 5, that is, whether the value is greater than or lesser than.

APPLICATION AND RESULTS

In this section, we will construct the logistic regression model to predict travel insurance policy claims. We use the travel insurance data from Kaggle [2]. The construction of the model will be divided into three processes: data observation, data preprocessing, and model building and evaluation. Data observation is intended to determine independent and dependent variables and observe the content of the data, known as data type. Data preprocessing is intended to examine missing data, inconsistency, non-numerical observation, outlier data, etc. In this process, we make sure that the data are ready to analyze by eliminating all the mentioned situations. Once the data is

cleaned, we divide the data into data training and data testing. The final step is to build and evaluate the model. The resulted model is applied to training data and returns the estimated claim data. The resulting data are then compared with the actual data to examine the model.

Table 1. Travel insurance data observation

	Agency	Distribution Channel	Product Name	Claim	Duration	Destination	Net Sales	Commission (in value)	Gender	Age
0	CBH	Offline	Comprehensive Plan	No	186	MALAYSIA	-29.0	9.57	F	81
1	CBH	Offline	Comprehensive Plan	No	186	MALAYSIA	-29.0	9.57	F	71
2	CWT	Online	Rental Vehicle Excess Insurance	No	65	AUSTRALIA	-49.5	29.70	NaN	32
3	CWT	Online	Rental Vehicle Excess Insurance	No	60	AUSTRALIA	-39.6	23.76	NaN	32
4	CWT	Online	Rental Vehicle Excess Insurance	No	79	ITALY	-19.8	11.88	NaN	41

Data Observation

We observe the data using the Python data analysis library or PANDAS [8]. The data consist of 11 attributes or columns with 63,326 entries. The attributes are claim status (Claim) that will be our target, name of agency (Agency), type of travel insurance agency (Agency Type), distribution channel of travel insurance agencies (Distribution Channel), name of travel insurance products (Product Name), travel destination (Destination), amount of sales of travel insurance policies (Net Sales), commission received for travel insurance agency (Commission) and gender of the insured (Gender). The overview of the first five rows of the data is in Table 1.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 63326 entries, 0 to 63325
Data columns (total 11 columns):
 #   Column                Non-Null Count  Dtype
---  ------                --------------  -----
 0   Agency                63326 non-null  object
 1   Agency Type           63326 non-null  object
 2   Distribution Channel  63326 non-null  object
 3   Product Name          63326 non-null  object
 4   Claim                 63326 non-null  object
 5   Duration              63326 non-null  int64
 6   Destination           63326 non-null  object
 7   Net Sales             63326 non-null  float64
 8   Commision (in value)  63326 non-null  float64
 9   Gender                18219 non-null  object
 10  Age                   63326 non-null  int64
dtypes: float64(2), int64(2), object(7)
memory usage: 5.3+ MB
```

Figure 1. Travel insurance data type

Table 1 implies that not all types of data are numerical. Figure 1 indicates the type of data in each column. From 11 columns, there are 7 columns that are not numerical. These columns are Agency, Agency Type, Distribution Channel, Product Name, Claim, and Destination. The data type which is not numerical will be converted into numerical. Every response in each column will be converted into a value based on several unique responses in the corresponding column. The number of unique responses in each column is presented in Table 2. The only column that has

missing data is the Gender column, with 45,107 entries. This value takes about 71.23% of the total entries. Therefore, we remove this column from our analysis.

Table 2. Number of individual responses in object data type column

Column Name	Agency	Distribution Channel	Product Name	Claim	Destination	Gender
n-Unique	16	2	26	2	149	2

Data Preprocessing

In this step, first, we convert the column with object data type into a numerical data type. The resulted table can be seen in Table 3. Next, we remove several columns that are not suitable. We choose the columns for the model construction based on the correlation value between each independent variable and the target variable. We select all columns that have a positive correlation with the target variable. The cross-correlation value between each variable in the data can be seen in Figure 2. From Figure 2, we can see that the variables highly correlated positively with the claim column are 'Duration', 'Net sales', 'Commission (in value)' and 'Destination'. We do not include the column 'Product name code' because the correlation value is too small.

Table 3. Converted travel insurance data

	Duration	**Net Sales**	**Commission (in value)**	**Age**	**Product Name**	**Claim**	**Agency**	**Destination**	**Agency Type**	**Distribution Channel**
0	186	-29.0	9.57	81	12	0	3	79	1	0
1	186	-29.0	9.57	71	12	0	3	79	1	0
2	65	-49.5	29.70	32	16	0	6	4	1	1
3	60	-39.6	23.76	32	16	0	6	4	1	1
4	79	-19.8	11.88	41	16	0	6	61	1	1

In the 'Claim' target column, we observe an imbalance in the number of responses between 'Yes' and 'No'. There are approximately 1.46% of 'Yes' answers, while 98.54% are 'No'. This situation will affect the precision of the model in predicting the 'Yes' answer since the 'No' answer dominates the target column. Therefore, we reconstruct the data such that the percentage of 'No' is 61.80% and the 'Yes' answer is 38.20%.

Table 4. Confusion matrix

Predicted	Actual	
	Yes	No
Yes	138	240
No	28	565

Model Building and Evaluation

The logistic regression model is created using a Python package (*scikit learn*). The result is described as true positive (TP), true negative (TN), false positive (FP), and false negative (FN) values. TP is the occurrence when the actual is 'Yes' agreed with predicted as 'Yes'. FP is the occurrence when the actual is 'No' but predicted as 'Yes'. FN is the occurrence when the actual answer is a “Yes” but predicted as a “No” answer. TN is the occurrence when the actual is 'No' agrees with the prediction as 'No'. In our case, the '1' denotes 'Yes' and '0' denotes 'No'. Table 4 provides the confusion matrix of our observation that describes all those values.

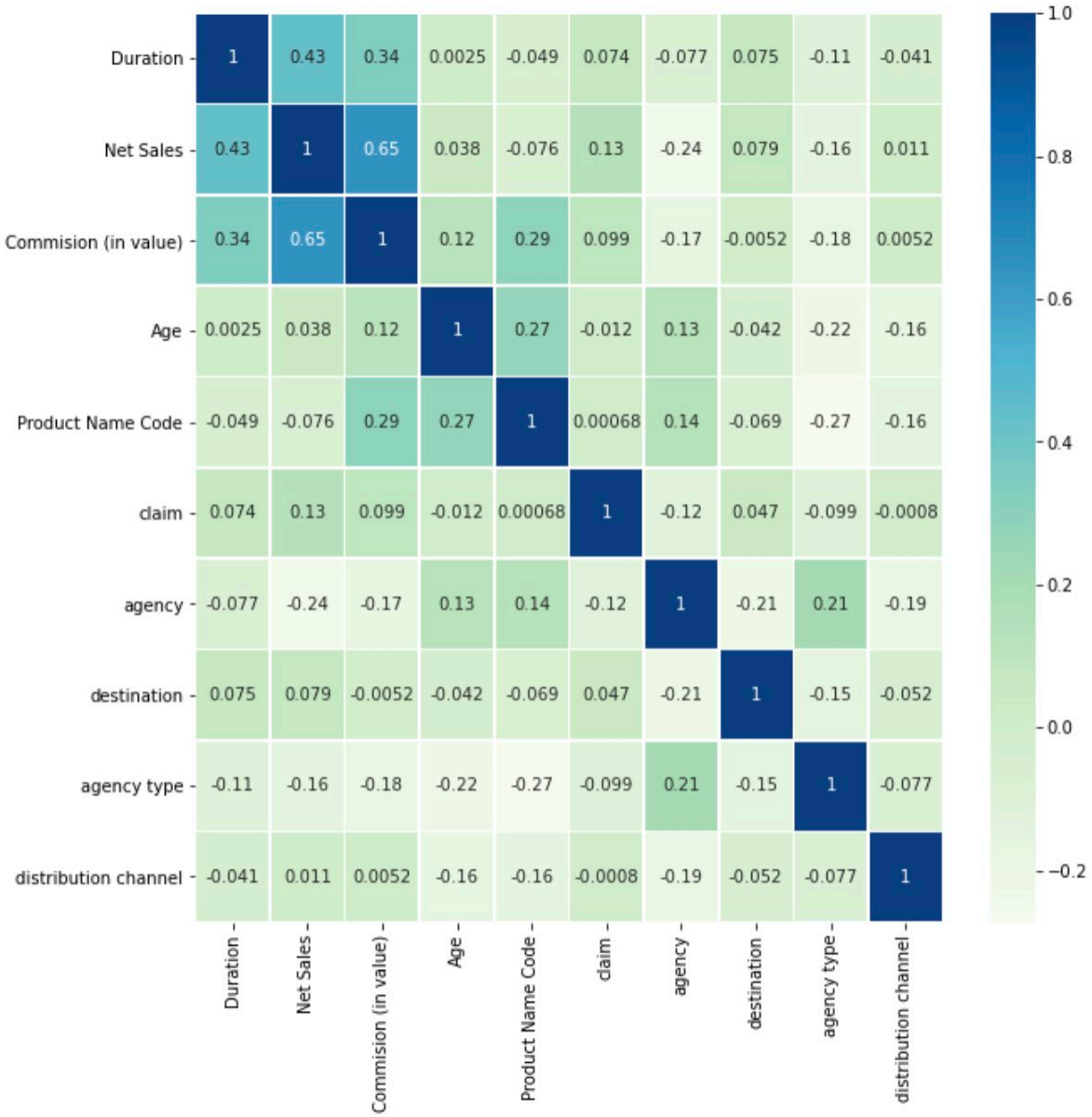

Figure 2. Cross-correlation matrix

From the confusion matrix (Table 4), we can evaluate the classification result of the report in Figure 3. The classification report in Figure 3 consists of several measures to assess the model. The measurements are precision, recall, f1 score, and accuracy. Accuracy is the most intuitive performance measure and is simply a ratio of correctly predicted observations to total observations. Precision is the ratio of correctly predicted positive observations to the total predicted positive observations. Recall the ratio of correctly predicted positive observations to all observations in the actual class - yes. The F1 score is the weighted average of precision and recall. This score takes into account both false positives and false negatives. Intuitively, it is not as easy to understand as accuracy, but F1 is usually more valuable than accuracy, especially if we have an uneven class distribution. Accuracy works best if false positives and false negatives have similar

costs. If the cost of false positives and false negatives is very different, it is better to look at both precision and recall. Our model gives the accuracy of 72%, 70% precision of no answer and 83% of yes answer, 95% recall of no answer and 37% recall of yes answer, and finally 81% f1 score of no answer and 51% of yes answer.

	precision	recall	f1-score	support
0	0.70	0.95	0.81	593
1	0.83	0.37	0.51	378
accuracy			0.72	971
macro avg	0.77	0.66	0.66	971
weighted avg	0.75	0.72	0.69	971

Figure 3. Classification report

CONCLUSION

We have analyzed the attributes that influence the travel insurance claim based on the correlation coefficient value with the target attribute. These attributes are 'Duration', 'Net sales', 'Commission (in value) ', and 'Destination'. The imbalance response of the target attribute is resolved by reconstructing the data so that the percentage of 'No' is 61.80% and 'Yes' is 38.20%. From this arrangement, we can construct a logistic regression model that returns a good result, as described in Figure 3. We observe that, in addition to the rearrangement of the data, the process that influences the result in Figure 3 is the random process that occurs when we split the data into training and test data. In *scikit learn*, this random process is denoted by a random state. The random state is an integer value that can be chosen to control the shuffle of 'yes' and 'no' responses in both the training and the test data. The composition of the training and test data will influence the result since our model is constructed based on the training data. In our case, we split the training and test data by 60% for the training data and 40% for the test data. Different treatments to handle the imbalance and missing data could lead to a different result. Also, applying different methods and comparing their results would be interesting future work.

CONFLICT OF INTERESTS

The author(s) declares that they have no conflict of interest.

REFERENCES

[1] Kagan, J. (2021, May 28). *Investopedia: Travel insurance defined* [online]. Accessed on 28 May 2021.
[2] Nasrudien, Z. (2019, January 28). *Kaggle: Travel insurance* [online]. Accessed on 28 May 2021.
[3] Weissert, W. G., Elston, J. M., & Koch, G. G. (1990). Risk of Institutionalization: 1977-1985. *Prepared for Office of Assistant Secretary for Planning and Evaluation US Department of Health and Human Services, Grant*, (88ASPE206A).
[4] Williamson, J. B., & McNamara, T. K. (2001). Why some workers remain in the labor force beyond the typical age of retirement. *Boston College CRR Working Paper*, 2001-09. https://doi.org/10.2139/ssrn.290095
[5] Barniv, R., Agarwal, A., & Leach, R. (2002). Predicting bankruptcy resolution. *Journal of Business Finance & Accounting*, *29*(3-4), 497-520. https://doi.org/10.1111/1468-5957.00440
[6] Sullivan, K. (2003). Transportation & work: exploring car usage and employment outcomes in the LSAL data. *NCSALL Occasional Paper*. National Center for the Study of Adult Learning and Literacy.

[7] Berenson, M., Levine, D., Szabat, K. A., & Krehbiel, T. C. (2012). *Basic business statistics: Concepts and applications*. Pearson Higher Education Australia.
[8] McKinney, W. (2011). PANDAS: a foundational Python library for data analysis and statistics. *Python for High-Performance and Scientific Computing*, *14*(9), 1-9.