Available online at: https://www.researchsynergy.org/aqa/
Applied Quantitative Analysis (AQA)
E-ISSN 2808-4934
Volume 1 Number 1 (2021): 18-24

Predicting Travel Insurance Policy Claim with Logistic Regression

Dadang Amir Hamzah, Averia A. Kalambe, Lucky S. Goklas, Naufal G. Alkhayyat
School of Business, President University, Indonesia

Abstract

This paper analyses the features that influence the travel insurance claim based on the 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 feature has the highest correlation with the dependent variable. Then, 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.

Keyword – *logistic regression, travel insurance, binary classification, classification, data science*

INTRODUCTION

Uncertainty always occurs in our daily life, especially when travelling. Incidents like lost baggage, delayed flights, public transportation accidents, food poisoning or sickness, injuries, or health issues have occurred. To minimize the loss, people consider travel insurance. According to Kagan (2021), travel insurance is a type of insurance that covers the costs and losses associated with travelling. Travel insurance is designed to cover the insured's trips for a limited duration. The insurance company will get more profit from the insurer who never claims. Mostly, the amount of the claim is higher than the amount of the premium paid by the insurer. Therefore, identifying the features that lead to a claim is highly desirable to the insurance company.

In this paper, we analyze the features that influence the insurer's claim. The data are taken from Nasrudien (2019), comprising 11 attributes (columns) and 63326 entries. The target attribute is the claim status column, which contains "Yes" or "No" as values. 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" or "Fail", "Win" or "Lose", "Yes" or "No", "Healthy" or "Sick", etc. Logistic regression was applied in various studies. In Weissert et al. (1990), logistic regression was used to compare the mobile nursing system established between 1977 and 1985 in the USA. Binary logistic regression was used to calculate people's retirement age based on age, sex, economics, and social status (Williamson & McNamara, 2001). In Barniv et al. (2002), logistic regression was used to classify companies' bankruptcy in America. Classification of car accidents in America was analyzed using logistic regression based on several criteria (Sullivan, 2003).

RESEARCH METHOD

For an explanation of logistic regression, we cite Berenson et al. (2012). Logistic regression is a model used to predict the probability of a particular categorical response for a given set of independent variables. This model uses the odds ratio, representing the probability of an event of interest compared with not having an event of interest. The formula to determine the odds ratio is:

Corresponding author
dadang.hamzah@president.ac.id; averia.kalambe@student.president.ac.id
https://doi.org/10.31098/quant.613

$$\text{odds ratio} = \frac{\text{Probability of an event of interest}}{1-\text{Probability of an event of interest}} \qquad (2.1)$$

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:

$$\ln(\text{odds ration}) = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_k X_{ki} + \varepsilon_i \qquad (2.2)$$

where $\beta_n, n = 1,2,\dots,k$ are the regression coefficients and ε_i is the random error in observation i. In application, finding the actual natural logarithm of the 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 follows:

$$\ln(\text{Estimated odds ratio}) \ = b_0 + b_1 X_{1i} + b_2 X_{2i} \ + \cdots + b_k X_{ki} \qquad (2.3)$$

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

$$\text{Estimated odds ratio} = \exp(b_0 + b_1 X_{1i} + b_2 X_{2i} + \cdots + b_k X_{ki}) \qquad (2.4)$$

Finally, we use (2.4) to estimate the probability of an event of interest as:

$$\text{Estimated probability of an event of interest} = \frac{\text{Estimated odds ratio}}{1+\text{Estimated odds ratio}} \qquad (2.5)$$

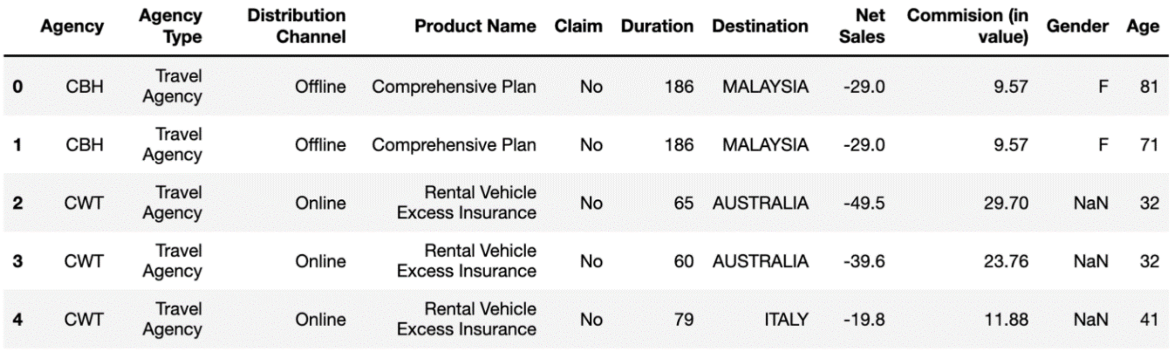
The value determined from (2.5) is an estimation for the value in (2.1). In practice, to find (2.3), all observations in each independent variable X_{ki}must be numerical. Therefore, if possible, the observation that is not numerical must be converted into numerical or removed from the analysis. The binary classification will be determined based on the value of (2.5), 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 data from Kaggle named Travel Insurance (Nasrudien, 2019). The model construction will be divided into three processes: data observation, data preprocessing, and model building and evaluation. Data observation is intended to determine the independent and dependent variables and observe the content of the data, known as the data type. Data preprocessing is intended to address missing data, inconsistencies, non-numeric observations, and outliers. In this process, we ensure that the data are ready for analysis by eliminating the aforementioned situations. Once the data is cleaned, we divide the data into training and testing. The final step is model building and evaluation. The resulting model is applied to the training data and returns the estimated claim data. The resulting data are then compared with the actual data to examine the model.

Data Observation

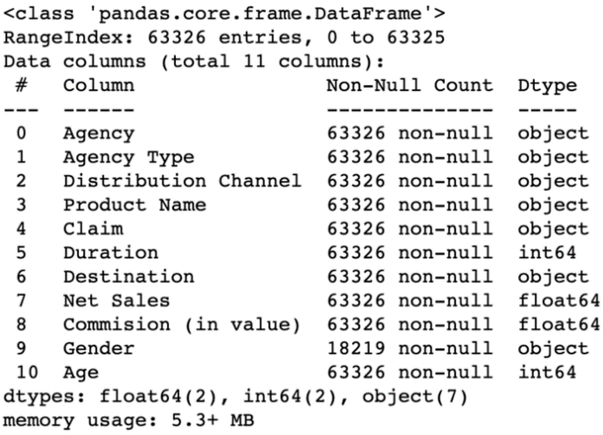
We observe the data using the Python data analysis library or pandas (McKinney, 2011). The data consist of 11 attributes (columns) with 63326 entries. The attributes are Claim Status (Claim) which will be our target, Name of agency (Agency), Type of travel insurance agency (Agency Type), Distribution channel of travel insurance agencies (Distribution Channel), Name of the travel insurance products (Product Name), Claim Status (Claim), Destination of travel (Destination), amount of sales of travel insurance policies (Net Sales), Commission received for travel insurance agency (Commission), and Gender of insured (Gender). The overview of the first five-rows of the data can be seen in Figure 1.



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

Figure 1. Travel insurance data observation

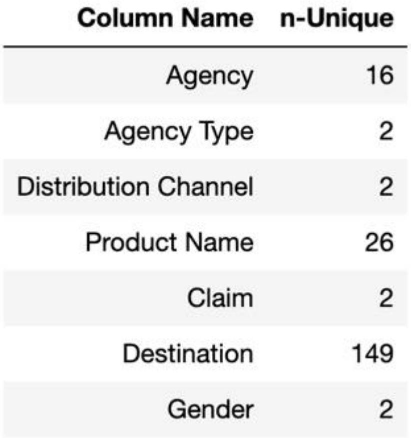
From Figure 1, we can see that not all the types of the data are numerical. The data type in each column can be seen in the figure 2. From 11 column, there are 7 columns which is not numerical. These columns are Agency, Agency Type, Distribution Channel, Product Name, Claim, and Destination.



```
<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 2. Travel insurance data type

Non-numeric data types will be converted to numeric types. Every response in each column will be converted into a value based on the number of unique responses in the corresponding column. The number of unique responses in each column is presented in Figure 3.



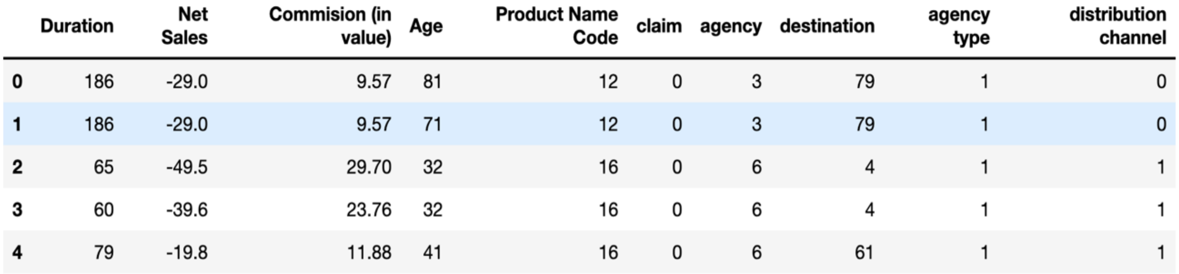
Column Name	n-Unique
Agency	16
Agency Type	2
Distribution Channel	2
Product Name	26
Claim	2
Destination	149
Gender	2

Figure 3. Number of individual responses in the object data type column.

The only column with missing data is Gender, which contains 45107 entries. This value takes about 71.23% of the total entries. Therefore, we remove this column from our analysis.

Data Preprocessing

In this step, we first convert the column with an object data type to a numeric data type. The resulting table is shown in Figure 4.



	Duration	Net Sales	Commision (in value)	Age	Product Name Code	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

Figure 4. Converted travel insurance data

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 with the target variable. We select all columns that have a positive correlation with the target variable. The cross-correlation coefficients for each pair of variables in the data are shown in Figure 5. From Figure 5, we can see that the variables highly positively correlate with the claim column are 'Duration', 'Net Sales', 'Commission (in value)', and 'destination'. We do not include the 'Product Name Code' column because the correlation value is too small.

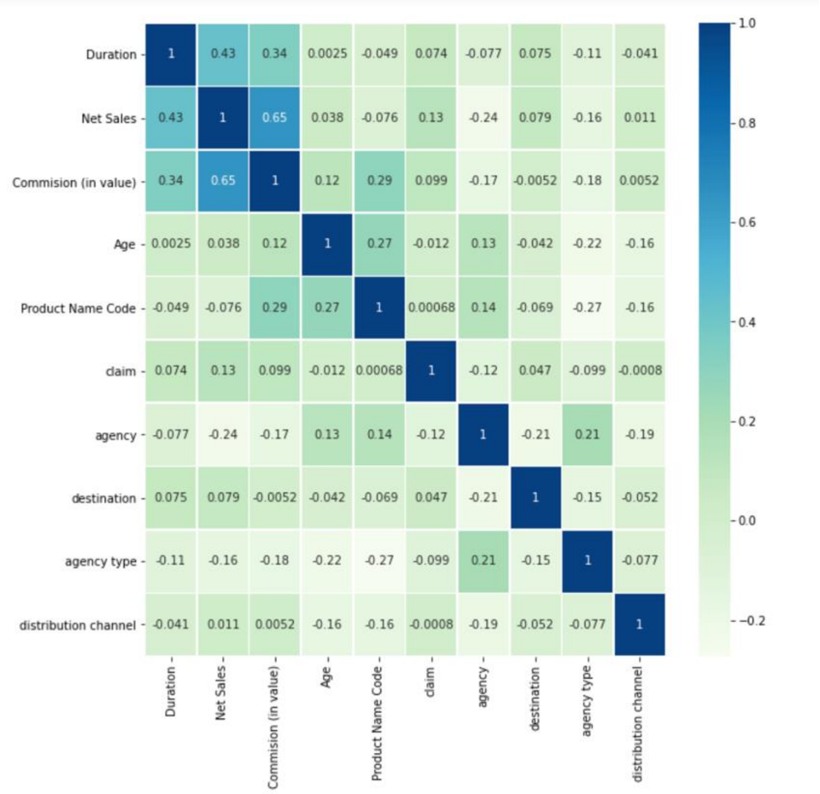


Figure 5. Cross Correlation Value

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% is "No". This situation will affect the model's precision in predicting the "Yes" answer, as 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%.

Model Building and Evaluation

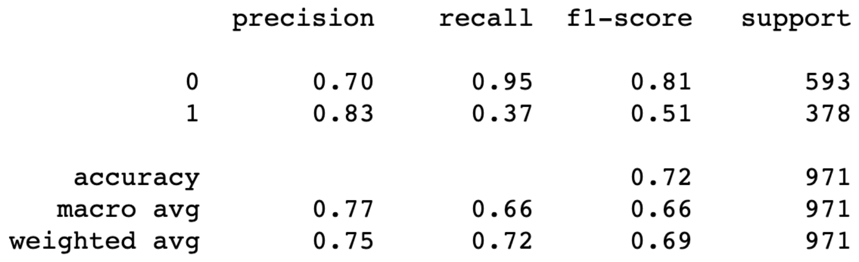
The logistic regression model is created using the Python package named scikit learn. The result is described in true positive, true negative, false positive, and false negative values. True positive (TP) is the number of occurrences when the actual is "Yes" agreed with predicted as "Yes".

False-positive (FP) is the number of occurrences when the actual is "No" but predicted as "Yes". False-negative (FN) is the number of occurrences when the actual answer is "Yes" but predicted as "No". True negative (TN) is the number of occurrences when the actual is "No" agree with predicted as "No". In our case, the "Yes" answer is denoted by "1" and "No" answer is denoted by "0". The table that describes all those values is called a confusion matrix. The confusion matrix for our observations is shown in Table 1.

Table 1. Confusion Matrix

	Actual	
Predicted	Yes	No
Yes	138	240
No	28	565

From the confusion matrix in Table 1, we can evaluate the classification results shown in Figure 6. The classification report in Figure 6 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 it is simply a ratio of correctly predicted observations to the total observations. Precision is the ratio of correctly predicted positive observations to the total predicted positive observations. Recall is the ratio of correctly predicted positive observations to all observations in the actual class - yes. F1 Score is the weighted average of Precision and Recall. This score takes both false positives and false negatives into account. 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 are very different, it's 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 6. Classification report

CONCLUSION

We analyzed the attributes that influence the travel insurance claim, based on the correlation coefficients with the target attribute. These attributes are 'Duration', 'Net Sales', 'Commission (in value)', and 'destination'. Imbalanced response from the target attribute is resolved by reconstructing the data such that the percentage of "No" is 61.80% and "Yes" is 38.20%. From this arrangement, we can construct a logistic regression model that yields good results, as shown in Figure 6. We observe that, besides the rearrangement of the data, the process that influences the result in Figure 6 is the random process that occurs when we split the data into training and test data. In sci-kit-learn, this random process is denoted by a random state. The random state is an integer value used to control the shuffle of "yes" and "no" answers in both the training and test data. The train and test data composition 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 and 40%

for the test data. Different treatments in handling the imbalance and missing data could lead to a different result. Also, applying different methods and compare their result would be an interesting future work.

REFERENCES

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

Berenson, M., Levine, D., Szabat, K. A., & Krehbiel, T. C. (2012). *Basic Business Statistics: Concepts and Applications*. Pearson Higher Education AU.

Kagan, J. (2021, May 28). *Comprehensive Guide to Travel Insurance: What It Covers and Why You Need It*. Investopedia. https://www.investopedia.com/terms/t/travel-insurance.asp

McKinney, W. (2011). Pandas: A Foundational Python Library for Data Analysis and Statistics. *Python for High Performance and Scientific Computing*, *14*(9), 1-9.

Nasrudien, Z. (2019, January 28). *Travel Insurance*. Kaggle. https://www.kaggle.com/mhdzahier/travel-insurance

Sullivan, K. (2003). *Transportation & Work: Exploring Car Usage and Employment Outcomes in the LSAL Data*. National Center for the Study of Adult Learning and Literacy, Harvard Graduate School of Education.

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).

Williamson, J. B. and McNamara, T. K. (2001, November). *Why Some Workers Remain in the Labor Force Beyond the Typical Age of Retirement*. Boston College CRR Working Paper No. 2001-09, http://dx.doi.org/10.2139/ssrn.290095