Risk Classification and Prediction: A Logistic Regression Approach for Analyzing Property Risk Classes in Insurance Companies

Reem Adel Abdallah
Department of Industrial Engineering and Engineering Management
University of Sharjah, Sharjah 27272
United Arab Emirates
U20105511@sharjah.ac.ae

Doraid Dalalah, PhD
Associate Professor, Coordinator of the PhD Program
Department of Industrial Engineering and Engineering Management
University of Sharjah, Sharjah 27272
United Arab Emirates
ddalalah@sharjah.ac.ae

Abstract

Risk exists in every aspect of a business, as it cannot be eliminated but rather reduced to an attainable level through the utilization of effective risk assessment techniques. Risk impacts people differently, some can be seen as risk-seeking while the majority are risk-averse. For the insurance industry in particular, risk is traded and transferred to the insurance providers as insurance providers offer a shield from the exposure to risk consequences and the likelihood of loss, therefore, escalating the risk from the insured entity to the insurer for a given premium. In this research, a modern model to risk classification will be proposed for property lines insurance. The proposed model will be validated via data collected from case studies of an insurance company in the United Arab Emirates (UAE). The model is expected to serve as a tool that helps provide better estimates of risk for various properties.

Keywords
Risk, Insurance, Property, Classification and Regression

1. Introduction

Risks in the financial industry refers to the amount of variability in the anticipated outcome of a business activity and the associated chances of the payoffs resulting of each outcome. It describes the amount of diversion from the expected and planned situations; the greater the diversion the greater the risk that is to be faced or tackled. When companies face decisions that include risk, they may have the option of either taking the risk through further involvement or avoiding it through disengagement from any new actions. Avoiding the risk means preferring the status quo as compared to taking the risk in a project with probabilistic outcomes. The risk-aversion or risk-taking depends on the resources available, the variability in payoffs, possibilities of states of nature, gains, losses, and risk attitudes.

There are two different groupings of risk: systematic and unsystematic. Systematic risk refers to the external aspects that influence and cause a risk to a company's investment, while the unsystematic refers to the assets that may influence the capabilities of a company or investor as risk occurs. Risks come from various sources, such as risk of liquidity, sovereign, business or insurance. For the purpose of this research, risk in insurance will be particularly highlighted and the various types will be discussed for the purpose of classifying the risk in insurance companies.
Insurance industry helps safeguard companies and businesses from various risks that could occur every day. It gives a financial shield for the insured entity to redeem a loss when unpredicted events occur. Moreover, it aids people in managing and predicting risks in order to keep them at a minimum. Insurance providers are constantly facing

challenges in estimating the required amount of coverage with respect to non-life insurance policies, as there are many factors that contribute to the changing levels of risk. For instance, in the UAE, majority of insurance companies are continually seeking to have a robust technique to not only manage risks, but also to predict it beforehand. In this study, the problem of classifying the risk will be addressed, particularly, property line insurance. This research will provide many insurance companies with reliable and compelling model which can be utilized in early stages of insurance risk estimation, leading to efficient decisions that will enhance the financial performance of the company while reducing the risks carried.

1.1 Objectives

This research aims at constructing a risk classification model for various properties through the use of Binary Regression. Additionally, it focuses on describing the best practices employed in the UAE insurance industry, which will help improve the customer experience and enhance the company's profits. Finally, it aims at successfully minimizing potential losses by utilizing effective risk prediction models.

2. Literature Review

New business opportunities are being presented with the emergence of Internet of Things (IOT), enabling the market to better gather data which can be used to improve the process of risk prediction in insurance industry (Baecke and Bocca 2017). Data mining along with risk assessment methods were utilized in motor insurance, resulting in enhanced risk control and management for the company since they were able to modify the policy conditions based on the client's real needs, in other words, the best way to manage risk and get more efficient results is by customizing and integrating the insurance coverage along with the client's usage. Additionally, this implementation improved the speed of insurance quotation being proposed for the customer, since this method works efficiently without the use of large historical data to predict risks. Logistic regression is the simplest form of machine learning algorithm, having a binary dependent variable (0,1) and it is famously used for the purpose of classifying or categorizing into binary outputs (A/B, 0/1, P/F) with the condition of having at least 2 independent variables or predictors for the model. The predictors will be checked for how they predict while supervising their effect in the model as well.

De Menezes et al. (2017) suggested the necessity of integrating logistic regression model with boosting to enrich the accuracy of the model, since the logistic regression doesn't factor in any noise in data. Therefore, expert systems are most suitable for such complex scenarios, where boosting is utilized progressively by enforcing a classification model to the re-weighted category of the data which is specified for training purposes. The analysis was made to differentiate the traditional approach with the maximum likelihood model, along with the logistic regression estimated through the boosting method for the binary classification. This was implemented on the Coronary Heart Disease (CHD) as a function of multiple biological parameters gathered from patients, with the intention of deciding the presence or absence of CHD. In conclusion, the results concluded that the model revealed more strength than the traditional approach. Moreover, it performed better in terms of area under the curve, responsiveness, precision and reduced false alarm rates. Furthermore, the application of logistic linear regression extend to building prediction models for COVID-19 patients, as it was used to determine mortality rates in China based on the age and time taken for triage (Josephus et al., 2021). The model demonstrated around 90% accuracy in predicting the probability of mortality in the patients, revealing that age is the highest contributing factor for patients.

In insurance industries, the company offers indemnity if the event occurs to the insured, for a given premium to be paid. In terms of life and health insurance, the type of risk being faced is the amount of money to be paid as a result of an unfortunate event such as injury or death. In 2018, a study by Grant et al., (2018) proposed the use of predictive models for enhancing risk prediction in the healthcare industry. The study implemented logistic regression model to examine the cardiothoracic surgeries carried out, demonstrating an outline of probable risk that could materialize and allowing for better prediction prior to its occurrence. Moreover, as the healthcare system is profoundly dependent on the amount of money estimated and paid by insurers, it is critical to present accurate determination for such amounts as it directly impacts the healthcare system's performance. A recent study captured the need for a developed model to better predict risk by putting forward a model which can detect the essential factors in determining life insurance prices, by utilizing Neuro Fuzzy Inference System (ANFIS) to capture the non-linear relationship between the data (Mladenovic et al., 2020). Consequently, the outcomes of the study showed that the most influencing factor in life insurance pricing was smoking. Wang et al., (2021) used a combination of classification along with logistic regression model in order to look into the introductory risk element for e-bike

users. This study collected relevant risk data for a period of three years, such as user's behavior, unacceptable conduct of users and environmental factors. The users were divided into five categories on the ranking tree analysis, with the category of non-professional users above the age of 55 in the suburban areas are correlated with the highest probability of damage when compared with other categories. Nonetheless, the logistic regression revealed that the categories showed that risk elements such as highways have repeatedly impacted the severity of damage for different categories. When comparing the crashes on highways against lower speed routes, the probability of damage for the highway has increased from approximately 9% to being around 42% for the e-bike users. Dong and Chan, (2013) explored the dynamic modeling of the long tail loss reserving data, which illustrates the state space mean model along with the Beta distribution to the CTP loss reserving data under the effect of legislation shifts. Nonetheless, the paper discussed the Beta distribution with the integration framework of individual loss data, revealing heterogenous traits and allowing for changing parameters with the groups. On the contrary, a model for comparing confident bands and managing the credit risk was proposed by Kiatsupaibul et al., (2017) and it deals with incorporating the interferences of the parameters alpha and beta in order to calculate the upper confidence level.

In summary, the application of logistic regression throughout the years extend to many applications and purposes. This review has revealed the power of logistic regression in classifying and categorizing risks in various industries, including the health sector. However, it was observed that there is a gap in applying this methodology in property lines of insurance companies in the UAE, when in fact the insurance industry is in need for such implementation to enhance its own performance along with the financial health of the country. Having said that, this paper will focus on constructing and implanting regression models for the purpose of predicting risks in insurance companies, property lines in particular.

3. Methods

In this paper, a UAE based insurance company's data will be assessed with the aim of accurately and effectively classifying risk categories (high risk, low risk) of various properties, in order to improve the financial performance with respect to how companies assess and predict risks. Moreover, the proposed model will be compared against the traditional approach currently implemented in order to validate the outcomes and demonstrate the effectiveness of implementing machine learning in risk prediction and classification models. A binary logistic regression model will be enforced to predict how risky a property is, given the input parameters which will be provided by the insured to the insurer. The logistic regression will predict the correct category to place the property into, meaning it will estimate the probability of it falling within one of the two categories (A being less risk and B being more risk). Furthermore, the risk categories can then be used to identify the risk profiles and the mean Damage Ratio (MDR) which will aid in forecasting the expected damage under risk neutrality assumption.

3.1 Data Collection

A sample of 100 data were collected from COPE (Construction, Occupancy, Protection and Environment) survey of the company regarding the construction height, material of property, business activity, fire protection systems, susceptibility to natural disasters, age of building, territory and estimated maximum loss. Additionally, the risk classification for the data collected were made available for the purpose of comparing the results of the model against the actual classification made by the company's experts. The samples will be analyzed by using SPSS binary logistic regression and assuming a cut probability of 0.7, which is the most commonly used in literature. In this proposed model, A is set to be a less risky and B being more risky property.

4. Model Analysis and Discussion

4.1 Model Description

The collected data was coded in SPSS, meaning that the age of the building is a continuous variable therefore it will have one coefficient. However, for the other variables such as rise of building or natural disasters, they will have categorical variables and levels (low, medium, high). The coding was made for the purpose of making the data a good fit for a binary logistic regression. The case summary of the cases along with the dependent variable coding are presented in tables 1 and 2, respectively. The coding for the dependent variable indicates that the cases labeled as A (lower risk), their internal value of the system will be 0, however, for cases which are labeled as B (higher risk), their internal value will be set to 1.

Table 1. Case processing summary

Unweighted Cases[a]		N	Percent
Selected Cases	Included in Analysis	100	100.0
	Missing Cases	0	.0
	Total	100	100.0
Unselected Cases		0	.0
Total		100	100.0

Table 2. Dependent variable coding

Original Value	Internal Value
A	0
B	1

Table 3. Chi-square test results

Step	Chi-square	df	Sig.
1	5.641	8	.687

4.2 Model Fit

In order to determine whether the model constructed is a good fit for the data, several tests have been made to verify this. For example, the Hosmer and Lemeshow test as shown in table 3, is similar to a chi-square test but is interpreted differently as statistical significance would mean that the model does not fit the data. Statistical significance would reveal that the model is a not good fit for the data. The model is observed to have a P-value of 0.687 which is greater than alpha 0.05 making the null hypothesis true ($P \leq 0.05$) and concluding that the model does not have a statistical significance and therefore it is fitting the data. Data regarding the types of class activity for the properties included in this analysis is illustrated in figure 1 and table 4. The highest frequency of class activity observed is both Residential Building and Commercial Property making 22% of the total classes each, while Tower was the least with only 9%. Figure 1 shows that Towers type buildings is least frequent followed by warehouses. Low frequency of class activity may result on higher emphasis on classes of higher frequency. Therefore, we may expect more risk to be involved in commercial buildings as compared to Towers.

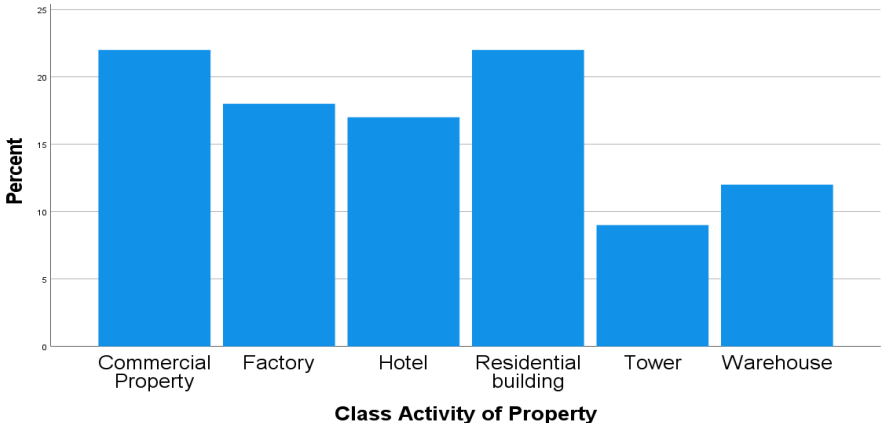

Figure 1. Percentage of each class activity.

Table 4. Class activity of property

	Frequency	Percent	Valid Percent	Cumulative Percent
Commercial Property	22	22.0	22.0	22.0
Factory	18	18.0	18.0	40.0
Hotel	17	17.0	17.0	57.0
Residential building	22	22.0	22.0	79.0
Tower	9	9.0	9.0	88.0
Warehouse	12	12.0	12.0	100.0
Total	100	100.0	100.0	

4.2 Model Probability Analysis

A regression model's log-likelihood value is a technique of determining the model's quality of fit. The greater the log-likelihood number, the better the model matches the dataset. For a particular model, the log-likelihood value might vary from negative infinity to positive infinity. The likelihood summary of the model is demonstrated in table 5, showing the relationship between the predictors and the outcome. For the Nagelkerke R Square which ranges from 0 to 1, the model has an R-square value of 0.697 meaning that it is a good fit.

Table 5. Likelihood model summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	55.533[a]	.495	.697

The general logistic formula is described as follow:

$$p(X) = \frac{1}{1+e^{-(a_0+a_1x_1+a_2x_2+\cdots+a_nx_n)}} \quad (1)$$

Given that p(X) refers to the probability of the classification of safe relative versus risky insurance scenario, the input vector X is given by $[X_1, X_2, \ldots, X_n]$, the set of coefficients are $[a_1X_1, a_2X_2, \ldots, a_nX_n]$ which will be optimized by utilizing the statistical software. The equation of the line for the model is described in table 6, the table shows the significant variables in the model which are age of property, rate of fire exposure from neighboring buildings, rate of damage from aircrafts crossing over property, rate of maintenance level, rate of housekeeping level, rate of fire protection level, rate of exposure to storms and floods. The variables in the equation show the regression coefficients, the predicted change in log odds for every unit positive change of predictor. A positive value of B represents an increasing value on the predictor and a positive association, while a negative value would mean that there's a decreasing likelihood or a negative association. For instance, for Age it can be seen that there is a positive association and it is statistically significant (sig <0.001) meaning that it has great influence on the outcome of the risk classification when compared to other variables that aren't significant, such as Estimated Maximum Loss of Property with significance of 0.882 which is greater than alpha (0.005) therefore having the lowest impact on the decision of risk classification.

Table 6. Variables in the equation

		B	S.E.	Wald	df	Sig.	Exp(B)	95% C.I.for EXP(B) Lower	Upper
Step 1[a]	Class Activity of Property	.475	.364	1.698	1	.192	1.608	.787	3.283
	Occupancy of Property	.640	.887	.521	1	.471	1.896	.333	10.779
	Rise or Height of Property	1.005	.684	2.160	1	.142	2.733	.715	10.448
	Age of Property	.336	.100	11.304	1	<.001	1.399	1.150	1.702
	Is the property's material highly combustible?	1.398	.934	2.239	1	.135	4.046	.649	25.240
	Rate of fire exposure from neighbouring building	1.282	.587	4.775	1	.029	3.604	1.141	11.381
	Rate of damage exposure from aircraft crossing over property	-2.307	1.055	4.787	1	.029	.100	.013	.786
	Rate of exposure to natural disaster of Storm	4.234	1.448	8.554	1	.003	68.999	4.042	1177.958
	Rate of exposure to natural disaster of Earthquake	.904	1.267	.509	1	.476	2.469	.206	29.585
	Rate of exposure to natural disaster of Tsunami	-1.619	1.607	1.015	1	.314	.198	.008	4.624
	Rate of exposure to natural disaster of Flood	-3.889	1.311	8.795	1	.003	.020	.002	.267
	Is the property enforcing strict non-smoking policies?	-.888	.907	.960	1	.327	.411	.070	2.432
	Rate of Maintenance level	-3.536	1.239	8.141	1	.004	.029	.003	.331
	Rate of Housekeeping level	-3.007	1.009	8.886	1	.003	.049	.007	.357
	Rate of Fire Protection level	-1.672	.707	5.590	1	.018	.188	.047	.751
	Estimated Maximum Loss of Property	.003	.017	.022	1	.882	1.003	.970	1.037
	Constant	9.919	4.797	4.275	1	.039	20320.626		

Table 7. Classification table

			Predicted		
			Risk Classification of Property		Percentage Correct
	Observed		A	B	
Step 1	Risk Classification of Property	A	68	1	98.6
		B	10	21	67.7
	Overall Percentage				89.0

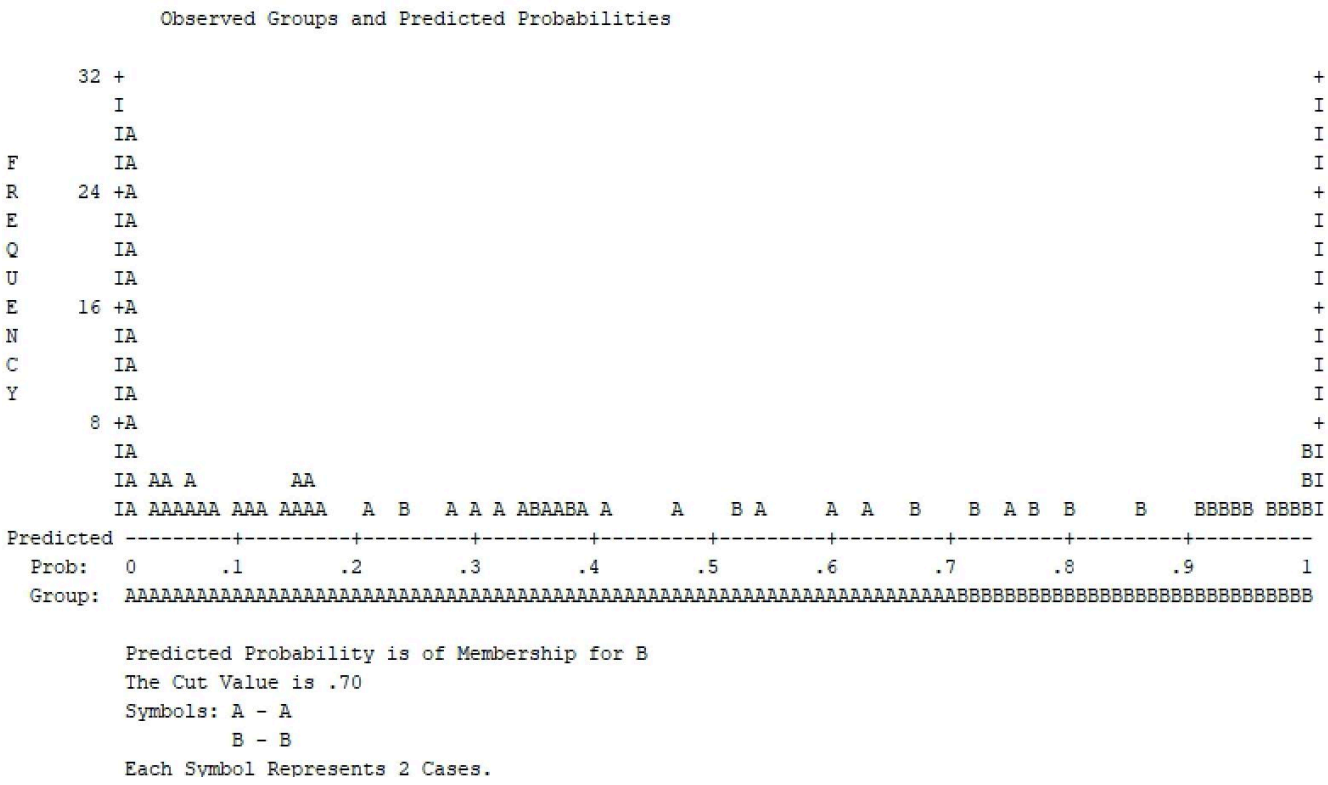

Figure 2. Classification plot

4.2 Model Classification and Discussion

The classification table lists the observed versus the predicted risk classification. In terms of prediction, the diagonal values represent the hits where the model predicted the risk classification accurately, while the off diagonals were the misses. Essentially, out of the 100 data points, the model predicted 89% of the cases and missed only 11% of them as seen in Table 7. The model has accurately predicted majority of the cases correctly. Hence, revealing the power of using this test for prediction. The cut-probability in this test was assumed to be 0.7 as it was the optimal value that gave the highest prediction percentage. Lastly, the classification plot is illustrated in figure 2 where each symbol representing two cases. This plot gives a graphical representation of the risk classes taken from the classification table, showing the cut observed group versus the predicted group and their frequencies alongside their probabilities. As it can be seen, any point falling beyond the cut-probability 0.7 is to be considered as class B (risky), while the others being less than 0.7 are to be considered as class A (safe). The model has predicted 89% of the points accurately, while only 11% of them incorrectly and this can be seen from the plot. The points falling at the extreme left have the lowest probability and are to be considered as class A (safe), while the ones falling at the extreme right carry the highest probabilities and are to be classified as class B (risky).

5. Conclusion and Future Research

In summary, the paper evaluated the binary logistic regression model for predicting the risks in property lines in an insurance company. It displayed a high performance and accuracy in predicting risk for a given set of input variables regarding the properties. The model was capable of detecting majority of the points (89%) and cases accurately when comparing the predicted cases versus the observed cases in terms of risk classification. As a result, it can be concluded that this case study underlined the benefits of utilizing regression models in risk prediction, assisting the companies in making better decisions on whether they should insure a property given the level or classification of risk it falls under. Consequently, this will enhance their financial performance along with the client experience, since the insurer will be able to give the best cover option for the insured based on the outputs of the classification model. Future research can be further made on the basis of significant variables only, to see how the model will respond in terms of correct predictions.

References

Baecke, P. and Bocca, L., The value of vehicle telematics data in insurance risk selection processes. *Decision Support Systems*, 98, pp.69-79, 2017.

De Menezes, F., Liska, G., Cirillo, M., and Vivanco, M., Data classification with binary response through the Boosting algorithm and logistic regression. *Expert Systems With Applications*, 69, 62-73, 2017.

Dong, A., and Chan, J., Bayesian analysis of loss reserving using dynamic models with generalized beta distribution. *Insurance: Mathematics And Economics*, 53(2), 355-365, 2013.

Grant, S., Collins, G. and Nashef, S., Statistical Primer: developing and validating a risk prediction model. *European Journal of Cardio-Thoracic Surgery*, vol. 54, no. 2, pp.203-208, 2018.

Josephus, B., Nawir, A., Wijaya, E., Moniaga, J., and Ohyver, M., Predict Mortality in Patients Infected with COVID-19 Virus Based on Observed Characteristics of the Patient using Logistic Regression. *Procedia Computer Science*, vol. 179, no. 871-877, 2021.

Kiatsupaibul, S., Hayter, A., and Somsong, S., Confidence sets and confidence bands for a beta distribution with applications to credit risk management. *Insurance: Mathematics And Economics*, vol. 75, pp. 98-104, 2017.

Mladenovic, S., Milovancevic, M., Mladenovic, I., Petrovic, J., Milovanovic, D., Petković, B., Resic, S. and Barjaktarović, M. Identification of the important variables for prediction of individual medical costs billed by health insurance. *Technology in Society*,vol. 62, p.101307, 2020.

Wang, J., Song, H., Fu, T., Behan, M., Jie, L., He, Y., and Shangguan, Q. Crash prediction for freeway work zones in real time: A comparison between Convolutional Neural Network and Binary Logistic Regression model. *International Journal Of Transportation Science And Technology*, 2021.

Wang, Z., Huang, S., Wang, J., Sulaj, D., Hao, W., & Kuang, A. Risk factors affecting crash injury severity for different groups of e-bike riders: A classification tree-based logistic regression model. *Journal Of Safety Research*, vol. 76, pp. 176-183, 2021.

Biographies

Reem Adel Abdallah is a graduate student in the Department of Industrial Engineering and Engineering Management at the University of Sharjah, UAE. She obtained her B.Sc. degree in Industrial Engineering and Engineering Management from the University of Sharjah.

Doraid M. Dalalah received his BSc in mechanical engineering from Jordan University of Science and Technology. He received his master's degree in Industrial Engineering from Jordan University in 1999. He worked as a maintenance engineer at Jordan Cement Factories till 2000. Doraid finished his Ph.D. degree from Lehigh University-USA in industrial and system engineering, 2005. Currently, Dr. Dalalah is an associate professor in Industrial engineering and Engineering Management at University of Sharjah.