14th Annual International Conference on Industrial Engineering and Operations Management
Dubai, United Arab Emirates (UAE), February 12-14, 2024
Publisher: IEOM Society International, USA DOI: 10.46254/AN14.20240282
Published: February 12, 2024

Risk Profile Classification and Premiums Estimation in Property Lines Insurance: A Multinomial Regression Approach

Reem Adel Abdallah
Department of Industrial Engineering and Engineering Management
University of Sharjah
Sharjah 27272, United Arab Emirates

Abstract

The notion of risk exists in every aspect of the business, as it cannot be eliminated but rather reduced to an attainable level through the utilization of effective risk management techniques. For the insurance industry in particular, risk is traded and transferred to the insurance providers as the company offers 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. This research is a development on a previously published paper by the author which had focused on the same issue but with the utilization of binary regression. The paper now proposes a modern model to risk classification which will be used for property lines insurance. The significance of the research lies in the fact that the process of risk prediction can be extremely complex since there are many parameters to keep in mind. In addition, accurate premium pricing can be difficult to estimate given the unpredictability of the risks occurring to the covered property. The data will be categorized into "A", "B", "C", and "D" using multinomial regression followed by a pricing model. The proposed model will be validated via data collected from surveyed properties of a UAE based insurance company. The model is expected to serve as a tool that helps provide better estimates of risk, premiums, and precise pricing.

Keywords
Insurance, Risk Classification, Machine Learning, Multinomial Regression

1. Introduction

The notion of risk differs from an industry to another, for this paper, the focal point will be risks in the financial industry as it describes the volatility and variances in the outcomes of business strategies and trades. According to Khandelwal et al. (2019), organizations are presented with a choice when faced with risky decisions: they can choose to take the risk by getting more involved or avoid it by doing nothing new. Risks can originate from a variety of sources, including sovereign, business, insurance, and liquidity risk. Because these risks can all affect the organization, it's critical to manage them well (Anton & Nucu, 2020). In order to categorize risks and determine insurance policy prices, this study will focus in particular on risk in insurance and explore the many kinds of risk. By giving the insured company a financial cushion to recover losses when unforeseen circumstances arise, the insurance sector assists in protecting organizations and enterprises from a variety of dangers that may arise on a daily basis. This is achieved by having the insured party pay a premium amount, where the amount is determined by various risk factors including the sum insured; the total amount of money in which the company is accountable to pay if a loss occurs to the covered property, the type of insurance coverage, the claims history along with the likelihood and size of potential losses. The load that is applied to the base premium varies depending on the insurance line; the higher the risk factors, the higher the load (Berger, 1988). Due to the numerous variables that affect the fluctuating levels of risk, insurance companies are continuously faced with difficulties in determining the necessary level of coverage and premiums for property

lines insurance policies (Boodhun & Jayabalan, 2018). For example, most insurance companies in the United Arab Emirates are always looking for a reliable method to forecast and manage risks. In insurance companies, risk management is divided into two categories: (1) risk classification, which necessitates the investigation of possible losses or damages through risk assessment in prior and comparable circumstances; and (2) premium estimation, which should ensure adequate profits for the insurance providers while remaining appealing to insurance policy carriers.

1.1 Objectives

This research will propose a new approach to risk prediction for insurance companies, as it will showcase novel techniques for risk classifications along with pricing, which are methods that have not been proposed jointly, specifically in the UAE. As a result, this will provide many insurance companies with reliable and compelling model which can be utilized in early stages of insurance pricing and risk estimation, leading to efficient decisions that will enhance the financial performance of the company while reducing the risks carried. This study builds on a previously published paper that employed binary regression to demonstrate the effectiveness of multinomial regression in more accurately modeling the data. Insurance of properties is an on-going major issue in UAE that is also faced by industry, since the classification and premium pricing methods are solely reliant on the expert judgement and background information on similar cases. The purpose of this research is to try to fill such a void by introducing a new approach for accurately classifying the risk and propose a pricing scheme, based on their probabilities and weight factor.

2. Literature Review

Where risk exists, there will always be unpredictability and ambiguity which often is accompanied by a loss. The insurance industry's sole purpose is to transfer the risk for some agreed premium that is to be paid by the client (Boodhun & Jayabalan, 2018). Insurance companies gain funds and finances by managing those risks that are to be faced. Risk management aims at acknowledging and addressing risks to understand and be able to better predict it in the future (Khandelwal et al., 2019). It is important that companies and organizations address those risks, not to just avoid reducing it, but rather to grasp potential market opportunity. Brand new market opportunities are presented as the Internet of Things (IoT) emerged in the business industry, allowing them to gather great number of data which can enhance risk prediction procedures in insurance industry (Mavrogiorgou et al., 2017). Risk assessment methods regarding various data mining methods were explored in motor insurance, resulting in better risk management and control, since the company had the ability to modify their coverage to match the client's needs (Baecke and Bocca, 2017). Severino and Peng, (2021) put forward the idea of predicting fraud in property lines through the utilization of machine learning techniques, the suggested model was examined repeatedly to compare the mean outcomes for any inaccuracies or inconsistencies. This revealed that by using ensemble techniques along with deep neural networks, results were given with highest accuracy and greater performance as opposed to the traditional models. On the other hand, Patil et al., (2018) suggested a model to detect fraud in credit cards through the inspection of real-time card transactions, revealing that using random forest decision tree gave the most accurate results.

Managing risks is an important role in every organization, as they face a variety of challenges with the increasing customer demand and technological advancements. Hence, implementing automated, continues risk assessments, and classification techniques can improve and support the process of effective risk management to deal with unexpected risks that may occur (Paltrinieri et al.,2019). Research suggests the use of machine learning to predict risks and determine premium prices for properties in insurance companies comes with higher confidence (Boodhun & Jayabalan, 2018).

Pal (2012) examined three feature selection models' performance in terms of Multi-class Linear Regression (MLR) using two hyperspectral sets of data. The results showed that the MLR selected 10-15 traits as put forward by Cawley and Talbot (2006) and gave noteworthy improvements in classifying with high precision. The use of multi-class classification and machine learning techniques are widely used in cancer diagnosis. A study conducted in 2018 proposed a technique Adaptive Multinomial Regression with a Sparse Overlapping Group Lasso penalty (AMRSOGL) with regards to lung cancer in patients (Li et al., 2018). The results displayed that it could detect the disease-causing genes of every category and notably surpassing the other methods of binary classifications. Moreover, in the application of categorical analysis, a study conducted by El-Habil, 2012 considered various health factors in children who have been physically abused, to classify the relationship between the variables and their correlation. Furthermore, emotion recognition in artificial intelligence has been studied through multinomial models for the purpose of identifying depression, happiness, and pain level (Najar & Bouguila, 2022). A recent study in Korea

focused on enhancing the weather forecasting by implementing multinomial regression along with correlation-based feature selection (Moon & Kim, 2020) for various geographical areas for midrange forecasts. The results revealed higher accuracy when using multinomial regression in comparison to traditional methods. Moreover, multinomial regression was used for predicting the patient behavior towards treatment pursuance (Brainard, Weston, Leach & Hunter, 2020), revealing that people who had a positive attitude towards the benefits of pursuing medical treatment were more likely to visit a medical professional. Furthermore, a study compared both binary and multinomial regression for users' prediction in terms of movement choice of bicyclist and correlation with the number of choices to be made at an intersection (Twaddle & Busch, 2019).

In conclusion, this review has demonstrated the effectiveness of multinomial regression in risk classification and categorization across a range of industries. It was noted, nevertheless, that there is a lack of application of this methodology in the property lines of insurance companies in the United Arab Emirates, despite the fact that the industry itself requires its implementation to improve both its own performance and the nation's financial stability. Having said that, the main objective of this paper is to build and implement regression models for the purpose of estimating premiums for property lines in particular and predicting risks in insurance companies.

3. Methods

In order to improve the financial performance with regard to how companies assess and predict risks, the data of an insurance company based in the UAE will be evaluated in this paper with the goal of accurately and effectively classifying risk categories (A, B, C, and D) of various properties. In addition, the suggested model will be contrasted with the conventional method now in use in order to verify the results and show how well machine learning can be incorporated into risk prediction and classification models. Given the input parameters that the insured will supply to the insurer, a multinomial regression model will be applied to forecast the riskiness of a property. Additionally, premium pricing estimates for various risk levels of potential insured entities can be made using the risk categories.

The first stage of the research process involves observing the financial performance and the manner in which companies estimate risks and premiums. This is known as the problem environment. After that, a multinomial regression-based model will be created to more accurately depict and explain the problem statement. After the model is constructed, the accuracy and reasonableness of the conducted model's solution will be evaluated by contrasting it with the actual company results.

The following kinds of information could be gathered from the company's risk survey:

- Construction type (high rise, medium rise, low rise)
- Material of the property
- Occupation type (office, residential, mall, factory, warehouse)
- Protection system (full protection, medium protection, low protection)
- Exposure type (sea/river, desert, near airport, hailstorm/rain/snow)
- Neighboring buildings
- Territory and safety (high safety, medium safety, low safety)
- Age of property
- Type of required insurance
- Estimated Maximum Loss (EML)

The data in this research has been refined in two rounds, the first is when the model was applied to all the data columns (input variables). This process was done by applying the Wald test, which has been re-tested on the new dataset after exclusion of insignificant variables. It was found that there was no correlation between the different input variables in this model, otherwise the correlated columns would have been eliminated during the process. Having no correlation indicates that one variable can't predict the other, which is true in this case. In addition, there were no missing values in the models, hence no missing value analysis was required for this.

3.1 Regression Model Construction

The proposed model will consist of two main parts: Classification of risk and estimating the premiums (through damage distributions). For the classification model, a regression model will be utilized to predict the risk categories based on the input parameters provided by the user. The risk categories in the proposed model will help identify the

damage ratio distribution, and the Mean Damage Ratio (MDR) will assist in decision making by predicting the expected damage or loss for a property, under the risk neutrality assumption. As a result, the expected damage ratio will help find the breakeven premium for the insurance provider. Figure 1 illustrates a brief structure and methodology of this research as it begins with selecting the input parameters for the properties, which will be used in the regression model to get the predicted classification.

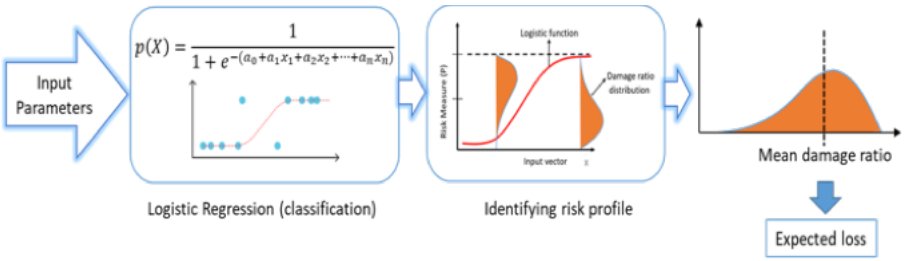

Figure 1. Sketch of the Model

The descriptive statistics of the Beta distribution can be expressed as:

$$\text{Probability Density Function} = \frac{x^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha,\beta)} \tag{1}$$

where: $B(\alpha,\beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$ and Γ is the Gamma function

$$\text{Mean} = \frac{\alpha}{\alpha+\beta} \tag{2}$$

The mean value given in equation (2) of the damage ratio will be used for the pricing and estimation of the premiums as shown in equation (5). The variance of the Beta distribution is given by equation (3) where the variance will decrease by increasing the values of Alpha and Beta (Thomas Bayes, 1763). Similarly, the mode will tend to be in the center of the range of the Beta distribution ranging from 0 to 1. As a remark for the Mode, when the values of Alpha and Beta approach each other, the Beta distribution will appear almost the same as anormal distribution in shape.

$$\text{Variance} = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)} \tag{3}$$

$$\text{Mode} = \frac{\alpha-1}{\alpha+\beta-2} \tag{4}$$

$$\text{Insurance Premium of client } i = \frac{\alpha_i}{\alpha_i+\beta_i} \times M_i \tag{5}$$

Where α_i,β_i, M_i are the distribution parameters and the maximum claimed amount of client i.

3.2 Risk Profile

Risk insinuates the degree of unpredictability of the future and its outcomes. It can be found in many forms such as investor, liquidity, natural disasters, operational, and insurance risks (Sendova, 2007). Implementing risk management plans can help businesses consider the wide range of possible risks that may occur and how to mitigate risks without compromising the strategic goals of the business. Nowadays, risks that organizations face are becoming more complex and highly unpredictable with the use of modern-technology and rapid climate change. A risk management plan is a document containing written information regarding the process to be taken to minimize the impact of risks on an organization, and aids in identifying, assessing, and controlling them (Ramos et al., 2020). Table 1 represents the risk management plan with the severity and probability of each risk. The green areas indicate that the risk can be easily dealt with by the organization. The orange areas indicate that the risk could impact the organization if there are no sufficient resources available to mitigate it. While the red areas indicate that the risk could lead to significant financial damage to the organization.

Risk profile on an individual level describes the inclination of a person towards taking risks, while on an organizational level it describes the external exposure that is being faced as threat. There is always a trade-off between risk and reward. In economics, safe outcomes are characterized by narrow range of payoffs, on contrast, risky outcomes are those which exhibit higher range, hence, higher standard deviation. This statement is valid for both discrete and continuous distributions. The range of possible damage in the event of losses and the range of possible payoffs in the event of gains determine the amount of risk (Roeser, 2013). Higher range means higher risk even in the event of gains. The same concept also applies for losses, the perception towards risk varies depending on the individual and how willing they are to accept the risk and understand it.

Table 1. Risk Management Plan

	Rare	Unlikely	Possible	Likely	Almost Certain
Catastrophic					
Major					
Moderate					
Minor					
Insignificant					

The general logistic regression formula is given by:

$$p(X) = \frac{1}{1 + e^{-(a_0 + a_1 x_1 + a_2 x_2 + \ldots + a_n x_n)}} \tag{6}$$

Where, $p(X)$ is the probability of the classification of safe relative to risky insurance transaction, the input vector X is given by $[x_1, x_2, \ldots, x_n]$ and the set of coefficients by $a_1x_1 + a_2x_2 + \ldots + a_nx_n$, which should be optimized using a statistical software such as Minitab or SPSS.

The logistic regression begins with initial values of the coefficients a_i, $i=1, 2, \ldots, m$. The sum of the coefficients is multiplied by the input parameters which is then evaluated through a sigmoid function that will help in the classification process Berkson (1944). The resulting classification is then compared with the actual classification, moreover, the error percentage is also optimized by repeatedly changing the values of the coefficients until there are no further changes in the percentage to be observed. The final resulting coefficients are used in the finalized model to predict the classification and help in establishing the Beta distribution and ultimately finding the risk premiums as it can be seen in Figure 2.

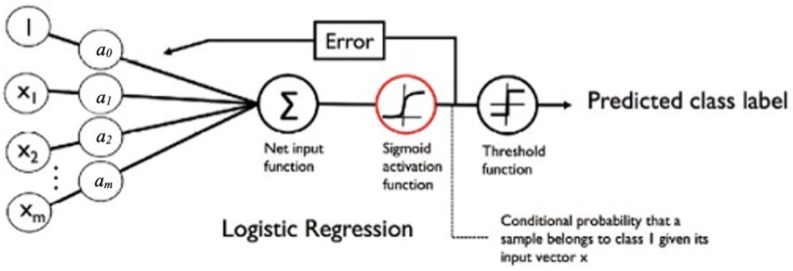

Figure 2. Logistic Regression

3.3 Mean Damage Ratio (MDR)

The notion of Damage Ratio refers to the ratio of the amount of money required to renovate a property versus the amount required to rebuild it. The size of a damage on a property during a catastrophe or claim will vary depending on various factors, however, some attributes give good indication on the vulnerability of the property, in other words, its most probable damage ratio. This factors in the economic loss, casualties, and downtime of the model (Fu, Gao & Li, 2021). The average of the statistical distribution shown in Figure 3 below illustrates the concept of MDR and the intensity function along with the plot can be described as vulnerability function, the figure also shows how the MDR increases with the intensity. In order to calculate the loss distribution in the financial model, the ratio of the damage

distribution is multiplied by the property replacement value, such calculations help insurance companies in predicting the policyholder's loss more precisely.

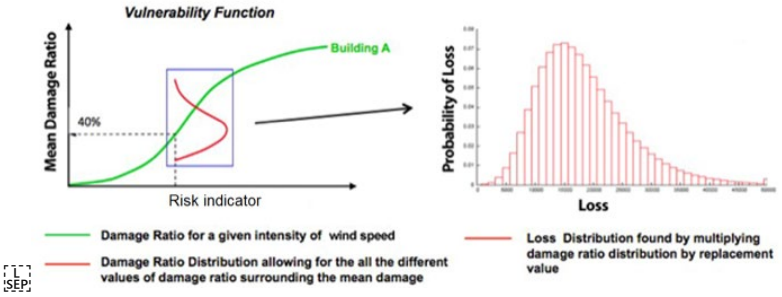

Figure 3: V*ulnerability model (left) and Financial Model (right). Modified from:* *http://understandinguncertainty.org/node/622*

3.4 Multi Category Risk Classification

The multinomial regression, known as Multi-class Linear Regression, is used in classification when there are three or more categories for the outcomes. It can predict an outcome that can be classified into multiple categories, and it can account for the interconnectivity between independent variables to anticipate the dependent variables. The assumptions of the multinomial regression are:

1- The dependent variables can't be anticipated from the independent variables perfectly.
2- The independent variables don't have to be statistically independent.
3- There should be no multicollinearity (having two or more independent parameters with high correlation amongst each other).

As for sample size, the model requires a minimum of 10 samples for every independent variable (Schwab, 2002). In this research, and to stay in line with practice in the selected case study, the risk is classified in to four groups given the same input variable. Since in multinomial classification there is no odds ratio used, the log of probabilities is computed according to a reference class (group) as follows:

$$Ln\left(\frac{\text{prob(being in group } m)}{\text{prob(being in reference group)}}\right) = \alpha_m + \beta_{m1}x_1 + \beta_{m2}x_2 + \ldots + \beta_{MK}x_K \tag{7}$$

Rearrange the above to get the probability of being in group m, we get:

$$\frac{\text{prob(being in group } m)}{\text{prob(being in reference group)}} = e^{\alpha_m + \beta_{m1}x_1 + \beta_{m2}x_2 + \ldots + \beta_{mk}x_k} \tag{8}$$

and the probability of being in the reference group is:

$$Prob \text{ (being in the reference group)} = \frac{1}{1 + e^{\alpha_m + \beta_{m1}x_1 + \beta_{m2}x_2 + \ldots + \beta_{MK}x_K}} \tag{9}$$

Where m=1, …M, k=1, …, K; M and K are the number of groups and independent variables respectively. For this study, M=4 and K= 16, meaning that four equations will result, each of which consist of 16 independent variables. In multinomial regression, the prediction is conducted as follows:

$$\text{Predicted group} = Prob \text{ (being in } q) | \ Prob \text{ (being in } q) = \max_{m=1,..,M} (prob\ of\ being\ in\ m) \tag{10}$$

Of note, each client will be characterized by specific probabilities and therefore, the probability indices will also be referred to by the client index as in equation (11). As for the pricing, the expected loss is computed by utilizing the

weight factors f_m which for this study is given by {0.025, 0.05, 0.075, 0.1} which is developed in this research. Hence, the expect loss for client i is estimated by:

$$\text{Expected loss of client } i = EML_i \times \sum_1^M f_m p_{im} \tag{11}$$

As will be illustrated in the discussion part, while the above expression can represent the expected loss of a client, for an insurance firm to make profit, the premiums should be higher than the above quantity.

4. Model Analysis

A multinomial regression model will be developed to classify risks into various categories while considering the interconnectivity amongst the independent variables to predict the dependent variable, which is the risk classification. The results of this section will illustrate the prediction alongside the estimation of the premium amount for various properties that are being insured. The classification model will aim at predicting the levels of risk into four classes, while the estimation model will utilize the probabilities given to each of the four categories for the purpose of estimating the premiums depending on the level of risk associated with the property. This model will result in classification of risks into four different categories "A", "B", "C", and "D" as shown in the following Table 2:

Table 2. Risk Classification and Description

Risk Category	Description
A	Safe
B	Neutral
C	Risky
D	Very Risky

The four classifications are advantageous in terms of prediction accuracy, as it gives a more detailed classification for the risks associated with a given probability of the property to fall within each of the four categories. The way that the software computes this is by providing a probability for every case falling within one of the four categories, and the category with the highest probability is the one to be predicted as it has a higher likelihood of being true. By plugging the data into the multinomial logistic regression using SPSS, the model results in the classification Table 3. The table illustrates the classification for the multinomial model where the diagonal represents the correct predictions and the off-diagonal represents the incorrect predictions. The model has correctly predicted 86% of the cases. Out of the 14% incorrect cases, none of them demonstrated significant errors, for instance, hardly one can find an "A" class predicted as a "D" class and vice versa.

Table 3. Classification Table for Multinomial Model

Classification

	Predicted				
Observed	A	B	C	D	Percent Correct
A	7	4	0	0	63.6%
B	2	55	0	1	94.8%
C	0	3	17	0	85.0%
D	0	3	1	7	63.6%
Overall Percentage	9.0%	65.0%	18.0%	8.0%	86.0%

The parameters for each of the risk categories "A", "B" and "C" are shown in below Tables (4, 5 and 6) respectively. However, it can be noted that category "D" is considered as a reference category, hence, there is no coefficient table for category "D" as it is calculated by adding the probability of three categories and subtracting them from 1. In terms of parameter estimates for category "A", it can be observed that the significant variables highly influencing the classification are occupancy, rise, exposure to flood, and maintenance level. Furthermore, the significant variables for category "B" are exposure from neighboring buildings, maintenance level, and fire protection level. In terms of category "C", the significant variable is only the age of the property. Given the significant variables for every category being different, it can be concluded that every risk category has different parameter estimates that highly contributes or influences the output of the categorical prediction.

Table 4. Parameter Estimates for Category "A"

Parameter Estimates

Risk Classification of Property[a]		B	Std. Error	Wald	df	Sig.	Exp(B)	95% Confidence Interval for Exp (B) Lower Bound	Upper Bound
A	Intercept	22.142	2205.668	.000	1	.992			
	Class Activity of Property	-.572	.774	.547	1	.460	.564	.124	2.571
	Occupancy of Property	-7.278	3.383	4.630	1	.031	.001	9.115E-7	.523
	Rise or Height of Property	-6.848	2.556	7.178	1	.007	.001	7.085E-6	.159
	Age of Property	-.251	.219	1.308	1	.253	.778	.506	1.196
	Is the property's material highly combustible?	-1.103	2.611	.178	1	.673	.332	.002	55.407
	Rate of fire exposure from neighbouring building	-.501	1.183	.179	1	.672	.606	.060	6.160
	Rate of damage exposure from aircraft crossing over property	-9.483	1562.310	.000	1	.995	7.612E-5	.000	.[b]
	Rate of exposure to natural disaster of Storm	-.924	3.330	.077	1	.781	.397	.001	270.951
	Rate of exposure to natural disaster of Earthquake	-4.869	2.950	2.725	1	.099	.008	2.370E-5	2.489
	Rate of exposure to natural disaster of Tsunami	-6.847	1556.939	.000	1	.996	.001	.000	.[b]
	Rate of exposure to natural disaster of Flood	8.539	3.247	6.917	1	.009	5109.522	8.805	2964994.146
	Is the property enforcing strict non-smoking policies?	6.156	3.779	2.653	1	.103	471.494	.286	777269.508
	Rate of Maintenance level	7.661	2.474	9.589	1	.002	2123.865	16.644	271019.619
	Rate of Housekeeping level	1.275	2.086	.374	1	.541	3.579	.060	213.371
	Rate of Fire Protection level	1.040	2.074	.251	1	.616	2.829	.049	164.798
	Estimated Maximum Loss of Property	-.056	.046	1.494	1	.222	.945	.864	1.034

Table 5. *Parameter Estimates for Category "B"*

B	Intercept	-8.505	977.527	.000	1	.993			
	Class Activity of Property	-.651	.620	1.103	1	.294	.521	.155	1.758
	Occupancy of Property	-2.102	1.772	1.407	1	.236	.122	.004	3.940
	Rise or Height of Property	-1.437	1.051	1.872	1	.171	.238	.030	1.862
	Age of Property	-.054	.174	.096	1	.757	.948	.673	1.333
	Is the property's material highly combustible?	-3.172	1.922	2.723	1	.099	.042	.001	1.813
	Rate of fire exposure from neighbouring building	-1.951	.828	5.556	1	.018	.142	.028	.720
	Rate of damage exposure from aircraft crossing over property	2.758	1.816	2.305	1	.129	15.762	.448	554.357
	Rate of exposure to natural disaster of Storm	-20.419	977.492	.000	1	.983	1.355E-9	.000	.[b]
	Rate of exposure to natural disaster of Earthquake	-2.425	2.620	.856	1	.355	.089	.001	15.035
	Rate of exposure to natural disaster of Tsunami	26.966	.000	.	1	.	5.143E+11	5.143E+11	5.143E+11
	Rate of exposure to natural disaster of Flood	3.018	2.048	2.173	1	.140	20.459	.370	1131.709
	Is the property enforcing strict non-smoking policies?	.400	1.311	.093	1	.761	1.491	.114	19.475
	Rate of Maintenance level	3.343	1.555	4.622	1	.032	28.302	1.344	596.191
	Rate of Housekeeping level	1.633	1.459	1.253	1	.263	5.118	.293	89.294
	Rate of Fire Protection level	2.931	1.258	5.428	1	.020	18.741	1.592	220.578
	Estimated Maximum Loss of Property	.001	.028	.000	1	.984	1.001	.947	1.058

Table 6. Parameter Estimates for Category "C"

C	Intercept	5.736	977.528	.000	1	.995			
	Class Activity of Property	.213	.686	.096	1	.757	1.237	.322	4.748
	Occupancy of Property	-3.179	1.791	3.151	1	.076	.042	.001	1.393
	Rise or Height of Property	-.915	1.089	.705	1	.401	.401	.047	3.387
	Age of Property	.396	.163	5.913	1	.015	1.486	1.080	2.045
	Is the property's material highly combustible?	-1.443	1.885	.586	1	.444	.236	.006	9.500
	Rate of fire exposure from neighbouring building	-.478	.815	.344	1	.558	.620	.126	3.063
	Rate of damage exposure from aircraft crossing over property	.296	1.889	.025	1	.875	1.345	.033	54.558
	Rate of exposure to natural disaster of Storm	2.408	3.205	.565	1	.452	11.116	.021	5946.045
	Rate of exposure to natural disaster of Earthquake	-1.778	2.527	.495	1	.482	.169	.001	23.904
	Rate of exposure to natural disaster of Tsunami	7.122	977.495	.000	1	.994	1238.685	.000	.[b]
	Rate of exposure to natural disaster of Flood	-1.312	2.222	.349	1	.555	.269	.003	20.959
	Is the property enforcing strict non-smoking policies?	-.689	1.451	.226	1	.635	.502	.029	8.625
	Rate of Maintenance level	-.573	1.337	.184	1	.668	.564	.041	7.753
	Rate of Housekeeping level	-2.644	1.804	2.149	1	.143	.071	.002	2.438
	Rate of Fire Protection level	1.472	1.235	1.421	1	.233	4.359	.387	49.052
	Estimated Maximum Loss of Property	.002	.028	.006	1	.937	1.002	.949	1.058

a. The reference category is: D.

The "likelihood ratio test" for the multinomial model is represented in Table 7 showing the significant variables for the overall model such as occupancy, rise, age, exposure from neighboring buildings, exposure from aircraft crossing over, exposure to storm, exposure to tsunami and flood, maintenance level, housekeeping level, and rate of fire protection. These variables are highly influential when it comes to classifying risks.

Table 7. Likelihood Ratio Tests

Likelihood Ratio Tests

Effect	Model Fitting Criteria: -2 Log Likelihood of Reduced Model	Likelihood Ratio Tests: Chi-Square	df	Sig.
Intercept	109.133	11.126	3	.011
Class Activity of Property	101.532	3.525	3	.318
Occupancy of Property	107.150	9.144	3	.027
Rise or Height of Property	112.268	14.262	3	.003
Age of Property	130.604	32.597	3	<.001
Is the property's material highly combustible?	103.192	5.185	3	.159
Rate of fire exposure from neighbouring building	108.188	10.181	3	.017
Rate of damage exposure from aircraft crossing over property	107.510	9.503	3	.023
Rate of exposure to natural disaster of Storm	114.541	16.534	3	<.001
Rate of exposure to natural disaster of Earthquake	102.232	4.225	3	.238
Rate of exposure to natural disaster of Tsunami	117.676	19.669	3	<.001
Rate of exposure to natural disaster of Flood	125.937	27.930	3	<.001
Is the property enforcing strict non-smoking policies?	103.461	5.454	3	.141
Rate of Maintenance level	120.732	22.725	3	<.001
Rate of Housekeeping level	111.833	13.826	3	.003
Rate of Fire Protection level	107.825	9.818	3	.020
Estimated Maximum Loss of Property	101.039	3.032	3	.387

The chi-square statistic is the difference in -2 log-likelihoods between the final model and a reduced model. The reduced model is formed by omitting an effect from the final model. The null hypothesis is that all parameters of that effect are 0.

The "goodness of fit" table represents how fit the model is to the data given, the below Table 8 contains both the Deviance and Pearson Chi-Square tests, where non-significance gives an indication that the model is a good fit for the presented data. However, both tests don't usually provide the same results, as for the Deviance test the p-value is less than 0.05 therefore indicating it is of significance and that the model is not a good fit for the data. While on the other hand, the Pearson test reveals that the p-value is greater than 0.05 meaning that it is not significant and that the model is a good fit for the data.

Table 8. Goodness of Fit Test

Goodness-of-Fit

	Chi-Square	df	Sig.
Pearson	626.076	246	<.001
Deviance	98.007	246	1.000

4.1 Pricing in Multinomial Regression Model

Once the risk classification is established, the estimation model can be constructed from the outputs. For each of the four risk categories, a probabilistic value is estimated and the category with the higher probability will be assigned as the predicted category. The probabilities of each category are estimated using the equations presented previously in the paper. The four probabilities will be used to calculate the expected weight factor of each case, i.e., client. Table 9

shows the used factors for each risk category, where higher risks will have higher risk factor, however, each risk categories comes with a specific probability in the multinomial model. The expected weight can be found by multiplying each probability with the factors of each class, which is developed in this research. That is:

$$\text{Expected weight factor of each case} = \sum_{i \in \{A,B,C,D\}} f_i \times p_i \quad (17)$$

Table 9: Damage Factor for the Classes

Risk Class	A	B	C	D
Damage factor for each class	0.025	0.05	0.075	0.1

The estimated loss for this model is computed by multiplying the maximum estimated loss by the expected weight, where higher risks result in higher estimated loss, which is given by:

$$\text{Estimated loss} = \text{EML} * \text{Expected weight factor.} \quad (18)$$

Equations (17 and 18), Table 9 shows the damage factor for each class and estimation for the multinomial regression model, respectively. These equations are developed in this research. The risk classes have been assigned weights to accommodate for the different risk levels, hence the weighted maximum estimated loss is calculated by 0.01 which is the commonly used weight by the company. For instance, as seen in table 10, the highlighted case is predicted to be of class "B" since the highest probability is 42% (which was assigned to class "B"). This case has an expected weight factor of 0.07 and a given EML of AED 8,000,000 hence, the estimated loss for this case is set to be the product of multiplying EML with expected weight factor, which yields AED 560,000 in terms of estimated loss. However, some weights need to be applied to this in order to accommodate for the different risk levels of the four different risk classes, therefore the weighted maximum estimated loss is calculated by 0.01 (a value commonly used by the company), then the weighted estimated loss can be computed by multiplying the weighted maximum estimated loss by expected weight factor, yielding AED 5,600 for the breakeven value of this property.

Table 10: Sample of Multinomial Regression Estimation

Category	PCP_1	ACP_1	Expected Weight Factor	EML	Estimated Loss	Weighted EML	Weighted Estimated Loss
A	0.95	0.95	0.02625	986,096,567	25,885,035	9,860,966	258,850
B	0.65	0.27	0.04575	919,496,567	42,066,968	9,194,966	420,670
B	0.87	0.87	0.04675	598,252,689	27,968,313	5,982,527	279,683
B	0.93	0.93	0.05375	210,472,986	11,312,923	2,104,730	113,129
C	0.78	0.78	0.0695	165,937,500	11,532,656	1,659,375	115,327
B	0.87	0.87	0.05625	787,500,000	44,296,875	7,875,000	442,969
C	0.96	0.96	0.0755	382,276,398	28,861,868	3,822,764	288,619
D	0.49	0.39	0.0775	250,000,000	19,375,000	2,500,000	193,750
B	0.91	0.91	0.05425	340,000,000	18,445,000	3,400,000	184,450
B	0.93	0.93	0.05375	278,000,000	14,942,500	2,780,000	149,425
B	0.98	0.98	0.04975	643,500,000	32,014,125	6,435,000	320,141
B	0.97	0.97	0.05075	328,000,000	16,646,000	3,280,000	166,460
B	0.69	0.69	0.05775	856,342,705	49,453,791	8,563,427	494,538
B	0.97	0.97	0.0515	653,100,000	33,634,650	6,531,000	336,347
B	0.99	0.99	0.05025	210,472,000	10,576,218	2,104,720	105,762
B	0.42	0.42	0.07	8,000,000	560,000	80,000	5,600
B	0.55	0.55	0.0695	28,276,600	1,965,224	282,766	19,652
B	0.56	0.56	0.06275	252,000,000	15,813,000	2,520,000	158,130

5. Results and Discussion

The classification model revealed high accuracy and prediction power, as it accommodates the risk category and the highly associated variables to predict the correct category to place the property into. Furthermore, it provides the user with a probability for each of the four categories to illustrate the likelihood of the property being in one of the four categories, where the category with the highest percentage reveals that it is the best fit for the property. This is advantageous as it allows the company to judge the results based on some qualitative measures that further influence the decision made regarding the categorization.

As for the estimation model, the model works as a great tool for the underwriters of the insurance company, as they can easily identify the right amount of premium to be given to a client to maximize the profits for providing insurance coverage. The output given works as a tool in decision making, while keeping in mind that they are the breakeven values, if the company is to insure the full property and EML amount. Moreover, the EML provided during the survey covers the total loss of a property and every aspect of the risk. However, in practice this is not always the case as the company usually covers losses from specific risks, for instance any loss occurring due to fire with exclusions, such as plate glass. Furthermore, the insurance providers usually set a limit to the losses to be recovered fully when occurring, for instance the limits for each loss occurring would be set to AED 1,000,000 despite the loss amount being more than that. This will reduce the premium to be paid by the insured since the insurance provider will only cover losses up to that specified limit. On the other hand, the company could collaborate with other insurance providers to reinsure this client, where the risk is split between them if the company fails to cover the EML value solely. The term reinsurance refers to the method used by insurance companies to transfer a percentage of the risk to another insurance provider, with an agreement to collaborate in paying off the amount to the client in case a loss occurs to the insured's property. By doing this, insurance providers diversify their portfolio, reduce their liabilities, and the chances of paying a large amount of money (Park & Xie, 2014). To further illustrate the significance of this research, Table 11 showcases the advantages of using multinomial regression in contrast to the binary regression. This research reflects the power of utilizing multinomial regression, which has enhanced the power of the previous research (Abdallah, R.A & Dalalah, D., 2022) by adding more risk factors which replicate the actual model used in UAE based insurance companies. Additionally, by allocating damage factors for each category, EML has been estimated for premium pricing purposes. This continuation is aiming to ease the classification and pricing procedure for commercial insurers, by reducing time and improving accuracy while automizing the process.

Table 11: Binary Logistic Regression Versus Multinomial Regression

Criteria	Binary Logistic Regression	Multinomial Regression
Number of risk categories*	Limited to two	Four categories
Model prediction	Predicted 89% of the cases (A,B)	Predicted 86% of the cases (A,B,C,D)
Probability	One cut-probability used throughout	Provided a probability for each of the four categories
Pricing	Due to limitations in actual risk categories, this model did not consider premium pricing	Premium pricing and EML included for all four categories, which is based on the damage factor

**The company's model includes four risk categories.*

6. Conclusion

In conclusion, risk is a factor that will continue to exist in every aspect of life as it cannot be excluded completely, especially in the business industry. However, with application of the right risk assessment techniques, it can be reduced to an acceptable level and have a minimal impact on the financial status of the business. The accuracy of risk predicted can in fact be enhanced by implementing machine learning and big data analysis (Jung et al., 2022). For the insurance industry, risk is often transferred from the insured client to the insurance provider for a given amount of premium that is specified for the respective duration of cover. Insurance companies provide a financial shield to the insured and transfer the liability to the company, therefore protecting the client from the likelihood and magnitude of a loss. This research is a continuation of a previously published paper (Abdallah, R.A & Dalalah, D., 2022) which focused on binary logistic regression, hence this particular paper has developed to include multinomial regression to predict and classify risks. Multinomial regression model has been developed for the purpose of predicting and classifying risks into various categories. Moreover, a new model for pricing has been suggested in this thesis for the purpose of

estimating the right amount of premium for each client. This will aid insurance companies in enhancing their risk prediction, attract new clients while matching their future needs, remain competitive in the market, and estimate claim amounts with higher accuracy.

The main conclusions can be summarized in the following bullet points, comparing the performance of the multinomial regression In comparison to the previously used method (Binary regression):

- The multinomial regression model demonstrated a more detailed classification by providing a probability for each of the four categories.
- Multinominal regression was precise enough to avoid severe misclassification of risk categories and showed no significant errors.
- The pricing model for the multinomial regression revealed high accuracy, as it allocated an expected weight factor for each of the four risk categories for calculating the expected loss.

For future work, the model may be further extended to predict more than four risk categories, should the data be available. Nonetheless, more input parameters can be included to enhance the prediction accuracy such as credit score, reputation, number of commercial policies issued under the same account, and claims history.

Acknowledgements

Special thanks and sincere appreciation to Professor Dorid Dalalah (who was a part of the previous paper) for serving as an exceptional mentor throughout this research endeavor. His invaluable guidance, unwavering support, and profound insights have been instrumental in shaping my academic journey and professional growth. His mentorship has not only contributed to the successful completion of this research but has also played a pivotal role in shaping my career aspirations. I am deeply grateful for his dedication, patience, and encouragement that have propelled me to reach this milestone in my academic pathway.

References

Abdallah, R.A. and Dalalah, D., 'Risk classification and prediction: A logistic regression approach for analyzing property risk classes in insurance companies', *Proceedings of the International Conference on Industrial Engineering and Operations Management* [Preprint], 2022. doi:10.46254/af03.20220019.

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

Bargmann, W., & Schadé, J., Progress in brain research. Amsterdam: Elsevier, 1963.

Blavatskyy, P. R., "Stronger utility", *Theory and Decision*, 76, 265-286, 2014.

Bell, D., Raiffa, H., & Tversky, A., Decision making. Cambridge: Cambridge University Press, 1988.

Chan, H., Chang, C., Chen, P., & Lee, J., Using multinomial logistic regression for prediction of soil depth in an area of complex topography in Taiwan. *CATENA*, *176*, 419-429, 2019. doi: 10.1016/j.catena.2019.01.030

Collins Terry R., Manuel D. Rossetti, Heather L. Nachtmann, James R. Oldham, "The use of multi-attribute utility theory to determine the overall best-in-class performer in a benchmarking study", *Benchmarking: An International Journal*, Volume: 13 Issue: 4, 2006.

De Menezes, F., Liska, G., Cirillo, M., & Vivanco, M., Data classification with binary response through the Boosting algorithm and logistic regression. *Expert Systems With Applications*, *69*, 62-73, 2017. doi: 10.1016/j.eswa.2016.08.014

Dionne, G., & Rothschild, C., Risk Classification and Health Insurance. SSRN Electronic Journal, 2012. doi: 10.2139/ssrn.2134190

Dong, A., & Chan, J., Bayesian analysis of loss reserving using dynamic models with generalized beta distribution. *Insurance: Mathematics And Economics*, *53*(2), 355-365, 2013. doi: 10.1016/j.insmatheco.2013.07.001

Douglass W. Shaw, Richard T. Woodward, "Why environmental and resource economists should care about non-expected utility models. *Resource and Energy Economics*, Volume 30, Issue 1, January, Pages 66–89, 2008.

Duan, Z., Chang, Y., Wang, Q., Chen, T., & Zhao, Q., A Logistic Regression Based Auto Insurance Rate-Making Model Designed for the Insurance Rate Reform. *International Journal Of Financial Studies*, *6*(1), 18, 2018. doi: 10.3390/ijfs6010018

Grant, S., Collins, G. and Nashef, S., Statistical Primer: developing and validating a risk prediction model. *European Journal of Cardio-Thoracic Surgery*, 54(2), pp.203-208, 2018.

Grover, G., Goyal, D., & Magan, R., Estimation of Seasonal Quality-Adjusted Life-Year Using Seemingly Unrelated Regression Equation Models With an Application to Orthopedic Data. *Value In Health Regional Issues*, *29*, 86-92, 2022. doi: 10.1016/j.vhri.2021.09.003

Hu, L., Hu, X., Wan, J., Lin, M., & Huang, J., The injury epidemiology of adult riders in vehicle-two-wheeler crashes in China, Ningbo, 2011–2015. *Journal Of Safety Research*, *72*, 21-28, 2020. doi: 10.1016/j.jsr.2019.12.011

Hu, L., Hu, X., Wang, J., Kuang, A., Hao, W., & Lin, M., Casualty risk of e-bike rider struck by passenger vehicle using China in-depth accident data. *Traffic Injury Prevention*, *21*(4), 283-287, 2020. doi: 10.1080/15389588.2020.1747614

Hu, X., Luo, H., Guo, M., & Wang, J., Ecological technology evaluation model and its application based on Logistic Regression. *Ecological Indicators*, *136*, 108641, 2022. doi: 10.1016/j.ecolind.2022.108641

Huang, Y., & Meng, S., Automobile insurance classification ratemaking based on telematics driving data. *Decision Support Systems*, *127*, 113156, 2019. doi: 10.1016/j.dss.2019.113156

Josephus, B., Nawir, A., Wijaya, E., Moniaga, J., & Ohyver, M., Predict Mortality in Patients Infected with COVID-19 Virus Based on Observed Characteristics of the Patient using Logistic Regression. *Procedia Computer Science*, *179*, 871-877, 2021. doi: 10.1016/j.procs.2021.01.076

Kaplinski, O., Risk Management of Construction Works by Means of the Utility Theory: A Case Study. *Procedia Engineering*, *57*, 533-539, 2013. doi: 10.1016/j.proeng.2013.04.068

Kiatsupaibul, S., Hayter, A., & Somsong, S., Confidence sets and confidence bands for a beta distribution with applications to credit risk management. *Insurance: Mathematics And Economics*, *75*, 98-104, 2017. doi: 10.1016/j.insmatheco.2017.05.006

Li, J., Wang, Y., Song, X., & Xiao, H., Adaptive multinomial regression with overlapping groups for multi-class classification of lung cancer. *Computers In Biology And Medicine*, *100*, 1-9, 2018. doi: 10.1016/j.compbiomed.2018.06.014

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*, 62, p.101307, 2020.

Pal, M., Multinomial logistic regression-based feature selection for hyperspectral data. *International Journal Of Applied Earth Observation And Geoinformation*, *14*(1), 214-220, 2012. doi: 10.1016/j.jag.2011.09.014

Patil, S., Nemade, V. and Soni, P., Predictive Modelling For Credit Card Fraud Detection Using Data Analytics. *Procedia Computer Science*, 132, pp.385-395, 2018.

Sendova, K., Operational Risk: Modeling Analytics. Technometrics, 49(4), 492-492, 2007. doi: 10.1198/tech.2007.s685

Severino, M. and Peng, Y., Machine learning algorithms for fraud prediction in property insurance: Empirical evidence using real-world microdata. *Machine Learning with Applications*, 5, p.100074, 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*, *76*, 176-183, 2021. doi: 10.1016/j.jsr.2020.12.009

Xu, X., Ge, Y., Wang, W., Lei, X., Kan, H., & Cai, J., Application of land use regression to map environmental noise in Shanghai, China. *Environment International*, *161*, 107111, 2022. doi: 10.1016/j.envint.2022.107111

Biography

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 Sc degree in Engineering Management from the University of Sharjah.