Exploring the application of information and data analytics in insurance: A Perspective of Institutional theory

Research Rationale

The insurance industry's risk analysis and risk management are growing increasingly significant and have become the core components of insurance technology management. New innovations in this field require more research and development. Emerging disciplinary technologies such as data mining have begun to evolve with the advent of the era of big data and machine learning as one of the most representative technologies. (Liu, 2019). On a daily basis, the insurance industry faces a number of fundamental issues, such as effective underwriting management. The majority of insurance executives, however, see changes in legislation and standards, shifts in company emphasis, the speed of technology development, changing client behaviors and expectations and competition from new market entrants as formidable difficulties (Cerchiello and Giudici, 2016). Most insurance businesses' survival strategies include implementing new core systems, automating all procedures and utilizing Artificial Intelligence (AI). AI is considered as the science of teaching a machine to execute a certain function that a human brain is incapable of performing when the data is large and complicated operations are required. All AI goals can be fulfilled using machine learning (ML) and deep learning (DL), which are the two pillars of AI (Dwivedi et al., 2021).

The continuity of insurance companies in performing their activities, ensuring their growth and development is linked to the quality of insurance services. These services offer to clients, their growing interest in financial technology in the insurance technology division. In fact, technological revolution that is sweeping the world rapidly by employing the ML technologies had an effect on insurance companies, where huge amounts of data could be reviewed in a short time, simplifying underwriting and settlement procedures (Klapkiv and Klapkiv, 2017). Thus, exploring the applications of ML worth studying in more detail, especially in the insurance industry, because their businesses depend mostly on trust. Furthermore, the most crucial objectives that insurance companies are pursuing to achieve are, focusing on reducing the cost of insurance services, increasing its market share, and expanding its customer bases.

This research aims to find an approach that increases the ability of insurance companies to study risks in more detail by analyzing large amounts of accurate data and as a result, individual risks are priced more accurately which will make premiums fairer as they will be more reflective of those risks by using the applications of ML and data analytics.

Literature Review

The change in technological advancement and frameworks is more rapid and need appropriate strategy to cope with the operations in the insurance industries. Therefore, this section will investigate and provide additional literature on the use of applications of (ML) in the insurance sector.

Pricing Optimization

In traditional actuarial science, insurers extract specified samples from long-term, multiple business practices to formulate a mathematical model used in pricing insurance products (Zheng and Guo, 2020). Consequently, the computed operational results become somehow less satisfactory. In addition, there are errors in pricing due to the limited nature of samples. Moreover, Zheng and Guo (2020) posit that this method of actuarial pricing relies on groups, not considering the individual, therefore, leading to

abstraction and generalization of the individual with the aim of grasping the overall trend. Therefore, big data analytics using the Internet of Things (IoT) and real-time risk assessment technology allow for large amounts of data to be collected to determine the laws behind the collected data, thus leading to accuracy in the statistical results (Nana, 2014). This enables product pricing to be more realistic. Additionally, big data significantly enriches factors in insurance risks, such as credit, leaving area, income, life work, browsing records, exercise frequency, rest, online time, hobbies, genes, risk preference and other personal information. According to Zheng and Gao (2020), the risk profile of an individual is enriched with multi-dimensional and comprehensive information by integrating individual data with a group's sample data, hence enhancing the insurance pricing ability, traditional actuarial theory and customer personalization and differentiation.

Customer Segmentation

Understanding a diversified consumer group is greatly aided by customer segmentation. Customers' socioeconomic, demographic, and frequency of various channels of communication were employed in customer segmentation in insurance using multinomial regression and count regression models (Dalla Pozza et al., 2018). A vehicle insurance firm was able to efficiently categorize their consumers into 'new,' 'risky,' 'uncertain,' and 'best' using a fuzzy analytic network process-based weighted recency, frequency, monetary value model, and K-means. Mau et al (2015) argued that understanding insurance usage across various communication channels was further aided by customer segmentation. Khalili-Damghani et al (2018) explained that when the strengths of clustering, classification, and rule mining methods are combined, hybrid models could be more efficient and accurate in consumer segmentation. Other algorithms have aided in the effective segmentation of clients.

Customer Churn Prediction

ML algorithms have been employed in a range of businesses including insurance to forecast whether or not a customer will churn. Several classification techniques including logistic regression, decision trees, Subversion (SVN), artificial neural network (ANN) and others have proven to be effective in forecasting customer turnover (Toreini et al., 2020). Given the complexity of customers' data, hybrid models like the Logit Leaf model may be preferable to standard methods (De Caigny et al., 2018). Adding word frequency and inverse document frequency to base classifiers like Support Vector Machines (SVM) and Nave Bayes has greatly improved customer churn prediction accuracy.

Prediction of Customer Lifetime Value

Customer's lifetime value (CLV) is critical in understanding a company's financial value to its customers from the perspective of the company. Customers' purchased products or services, socioeconomic and demographic information are used by various ML algorithms such as classification and regression trees (CART), SVM, additive regression, K-Star Method, multilayer perceptron, wavelet neural network, regression, and Nave to predict the CLV, which helps understand the probability of customers' attitude and behavior toward successfully maintaining their policies (Rathi and Ravi, 2017). Other advanced methods for CLV prediction, such as synthetic minority oversampling paired with deep neural networks, have yielded encouraging results (Sifa et al., 2018).

Application of Machine Learning in Insurance Risk Management

New cognitive technologies such as big data, machine learning and natural language processing are displacing traditional methods of analysis to quantitatively analyze and process the ever-growing data sets generated in the insurance market. This allows for the identification of various risk indicators and

thus more effective risk management. Massive data about human consumption, entertainment, credit, and other behaviors have arisen as a result of the increase in Internet Technology. Traditional software solutions are now incapable of acquiring, managing, and analyzing large data sets in a timely manner. New processing concepts and technical tools are now required. In order to ensure insurance operations or profit from insurance operations, insurers have improved their risk assessments and developed and implemented a variety of innovative tools for new disaster preventive and depletion technologies (Liu, 2019).

The objective of an insurance company's risk control strategy is to modify the company's risk status, assist the firm in avoiding risk and preventing losses and aim to lessen the risk's negative impact on the loss in the event of a loss. Machine learning is a modeling and analysis tool based on huge data. It extracts useful information from a large volume of information. It adjusts training samples on a regular basis by continuously selecting data, constructing model data, checking data and re-adjusting models. The data itself searches for inherent patterns and rules that, in the end, give the best results. The majority of risk financing decisions made by insurance companies are a combination of retention and transfer risk to varied degrees (Lin et al., 2017). Quantifiable indicators should be used as much as is feasible in the design process, and certain qualitative indicators should be used to more consistently indicate financial concerns that are not captured by quantitative indicators. Large companies use big data mining to create their own credit rating systems, whereas small and medium-sized businesses use information sharing and third-party services to receive risk assessment services. Risks will be transferred by insurance firms by making self-retaining decisions which necessitates a simple tradeoff between two factors. There are many learning approaches for instance, neural network, support vector machine, random forest and clustering analysis techniques that correspond to distinct algorithms. Thus, contrasting prediction results for various application modes, data sets and prediction objectives will be given by different algorithms (Liu, 2019).

There are some practices that insurance companies should adopt in order to mitigate the consequences when disputes happen. Pham et al (2019) stated that internal programs and compliance functions could improve cybersecurity processes, increase claim transparency and facilitate communication among parties. Agyei et al (2020) explained that in the insurance industry, customers buy promises; therefore, fair insurance premiums greatly affect customers' purchase decisions. Meyers and Van Hoyweghen (2018) argued that since the quality of the product that customers buy is invisible, fair premium plays a primary role in the insurance industry.

Research Questions

This research will be guided by the following research questions:

1. What is the relationship between competitive advantage of insurance companies and data analytics applications in Saudi?
2. How does machine learning affect the success of insurance firms?
3. Which elements of the machine learning approach are desirable to achieve positive outcomes in Saudi companies?
4. What levels of efficacy do these elements achieve in enhancing the work of insurance companies?
5. How important is big data in achievement work by Saudi companies?
6. What are the challenges in the application of data analytics?

Research Methodology

Samples

The sample here will be non-probability, and since the researcher will study a particular group within companies, the purposive homogeneous sample method will be used. We will use large data set in insurance companies such as Altawuniya, Alsagr, Malath and Allianz, along with implementing advanced machine learning methods.

Data Collection Methods

This study will adopt the philosophy of interpretation, which indicates that earthly phenomena are very complex to be regulated by particular rules and laws (Wegner, 2008). Interpretivism permits to represent the complicated issues from the researcher's perspective. In this work, answering the research question tends to be difficult and multi-disciplinary. In other words, there is an overlap among data science and enterprise risk management. Interpretivism tends to determine the contributions of big data analytics in the insurance sector and how information and machine learning methods are applied in the insurance industry.

Regarding secondary data, academic and empirical articles, books and official websites will be used. Secondary data is more likely to help the researcher in building a baseline regarding the theoretical development of different types of data analytics detailed in the literature review above. Moreover, it will be helpful for identifying the limitations and values of enterprise risk management. The primary data will be collected through an empirical investigation into the efficacy of data analytics and how it impacts on the insurance industry. Semi-structured interviews will be the main data collection tool after contact with the insurance company's stakeholders.

Data Analysis

Machine Learning Approach to Predict Fair Premium

The purpose of predicting the fair premium is to estimate the risks. Let Y = 0,1 be the possible result, with 0,1 being categorical values indicating 'high risk' or 'not high risk,' respectively.

This machine learning model's goal is to estimate the probability of fair premiums. This algorithm is based on data that represents a consumer that is either has a high risk or low risk. Therefore, the problem can be classified as a binary classification, with (a) indicating a high risk and (b) indicating not high risk. There are a variety of classification algorithms that can be utilized. Given the current status of the data, some of them perform better, while others perform worse.

$Pr(Y = 0|X = xi)$ (1)

$Pr(Y = 1|X = xi)$ (2)

where X is a collection of instances xi that represent all of the i-th policyholder's known information (Hanafy and Ming, 2021).

Regression Analysis

The linear relationship between a response variable (target) and a set of predictor variables is estimated by using linear regression. When the target variable is binary, however, linear regression is not appropriate (Sabbeh, 2018). Logistic Regression (LR) is an appropriate model for evaluating regression for binary-dependent variables. LR is a statistical method for describing how a binary target variable is linked to a set of independent variables. It has a lot in common with linear regression.

For dichotomous results, LR is a multivariable learning method. It is the best classification method for models with two outputs, such as yes/no decision making (Musa, 2013). Therefore, it can be used to anticipate a reasonable insurance premium using only two factors (high risk or not high risk). The functioning of LR is comparable to that of linear regression. In contrast to the categorical output, we want in the binary target variable, linear regression produces a continuous result. LR uses a single output variable, yi, with I = 1,..., n with each yi holding one of two values: 0 or 1. (but not both). As seen in the following equation, this follows the Bernoulli probability density function:

$$p(yi) = (\pi i)^{yi} \ (1 - \pi i)^{1-yi} \qquad (3)$$

When the probability is (πi) this takes the value 1 and takes the value 0 when the probability is (1 – πi); the interest in this is when yi = 1 with an interesting probability (πi). The projected label will then be output by the classifier; (πi) is equal to 1 if it is larger than or equal to its threshold (by default, 0.5) (Hanafy and Ming, 2021).

Expected contribution of research

This research will benefit several financial institutions in the insurance sector. The information in the research will provide frameworks for more precise decisions and techniques for avoiding risks that may have been poorly calculated. As a result, the firms will have improved service delivery when they have a competitive advantage. Therefore, insurance firms and other financial institutions would maintain their relevance. In addition, the research will pave way for more research on big data analytics, not only in the insurance industry but also in other industries in the economy.

Timeline

Research Activity	Month	-3	-2	-1	6	6	5	7	12
Problem identification									
Proposal development									
Project Approval									
Site preparation									
Recruitment of participants									
Data collection									
Data analysis									
Report writing									

The duration for the research process will be 36 months. The negative months represents time taken to carry out ethics approval, problem identification and proposal approval. Next will be site preparation which will take 6 months. The recruitment of participants will also take 6 months, collection of data will be 5 months, while data analysis, 7 months. Report writing will be the last step and will take 12 months. However, in the event of an unfinished step, the schedule will be adjusted and updated to accommodate the new changes.

References

AGYEI, J., SUN, S., ABROKWAH, E., PENNEY, E. K. & OFORI-BOAFO, R. 2020. Influence of trust on customer engagement: Empirical evidence from the insurance industry in Ghana. *SAGE Open,* 10**,** 2158244019899104

.

CERCHIELLO, P. & GIUDICI, P. 2016. Big data analysis for financial risk management. *Journal of Big Data,* 3**,** 1-12

DALLA POZZA, I., BROCHADO, A., TEXIER, L. & NAJAR, D. 2018. Multichannel segmentation in the after-sales stage in the insurance industry. *International Journal of Bank Marketing*

COUSSEMENT, K. & DE BOCK, K. W. 2018. A new hybrid classification ,.DE CAIGNY, A algorithm for customer churn prediction based on logistic regression and decision trees. *European Journal of Operational Research,* 269**,** 760-772

AARTS, G., COOMBS, C., CRICK, T., DUAN, Y., ,.DWIVEDI, Y. K., HUGHES, L., ISMAGILOVA, E DWIVEDI, R., EDWARDS, J. & EIRUG, A. 2021. Artificial Intelligence (AI): Multidisciplinary perspectives on emerging challenges, opportunities, and agenda for research, practice and policy. *International Journal of Information Management,* 57**,** 101994

HANAFY, M. & MING, R. 2021. Machine learning approaches for auto insurance big data. *Risks,* 9**,** 42.

KHALILI-DAMGHANI, K., ABDI, F. & ABOLMAKAREM, S. 2018. Hybrid soft computing approach based on clustering, rule mining, and decision tree analysis for customer segmentation problem: Real case of customer-centric industries. *Applied Soft Computing,* 73**,** 816-828

KLAPKIV, L. & KLAPKIV, J. 2017. Technological innovations in the insurance industry.

TIAN, R. & YU, J. 2017. Pension risk management in the enterprise ,.LIN, Y., MACMINN, R. D risk management framework. *Journal of Risk and Insurance,* 84**,** 345-365.

LIU, Q. Research on risk management of big data and machine learning insurance based on internet finance. Journal of Physics: Conference Series, 2019. IOP Publishing, 052076.

MAU, S., CVIJIKJ, I. P. & WAGNER, J. 2015. Understanding the differences in customer portfolio characteristics and insurance consumption across distribution channels. *Working Paper presented at the WRIEC in Munich*.

MEYERS, G. & VAN HOYWEGHEN, I. 2018. Enacting actuarial fairness in insurance: From fair discrimination to behaviour-based fairness. *Science as Culture,* 27**,** 413-438.

MUSA, A. B. 2013. Comparative study on classification performance between support vector machine and logistic regression. *International Journal of Machine Learning and Cybernetics,* 4**,** 13-24.

NANA, P. N. 2014. Legal rights, information sharing, and private credit: new cross-country evidence. *The Quarterly Review of Economics and Finance,* 54**,** 315-323.
PHAM, H. C., BRENNAN, L., PARKER, L., PHAN-LE, N. T., ULHAQ, I., NKHOMA, M. Z. & NGUYEN, M. N. 2019. Enhancing cyber security behavior: an internal social marketing approach. *Information & Computer Security*

Customer lifetime value measurement using machine learning .2017 .RATHI, T. & RAVI, V techniques. *Artificial Intelligence: Concepts, Methodologies, Tools, and Applications.* IGI Global.

SABBEH, S. F. 2018. Machine-learning techniques for customer retention: A comparative study. *International Journal of Advanced Computer Science and Applications.*

SIFA, R., RUNGE, J., BAUCKHAGE, C. & KLAPPER, D. Customer lifetime value prediction in non-contractual freemium settings: Chasing high-value users using deep neural networks and SMOTE. Proceedings of the 51st Hawaii International Conference on System Sciences, 2018.

TOREINI, E., AITKEN, M., COOPAMOOTOO, K., ELLIOTT, K., ZELAYA, C. G. & VAN MOORSEL, A. The relationship between trust in AI and trustworthy machine learning technologies. Proceedings of the 2020 conference on fairness, accountability, and transparency, 2020. 272-283.

ZHENG, L. & GUO, L. Application of Big Data Technology in Insurance Innovation. International Conference on Education, Economics and Information Management (ICEEIM 2019), 2020. Atlantis Press, 285-294