**Improving Diversified Pricing Strategy with Machine Learning**

INTEGRATED PROJECT | SUMMER 2024

Team Arocha

Luyang Feng | Orange | Ruizhe

August 9, 2024

~~Cover~~

Table of Contents

~~Executive Summary~~

~~Business Background~~

~~Goals of the Project~~

Analysis of Result

Recommendations & Lessons

Future Considerations

Appendices

Reference

Table of Contents

Executive Summary………………………………………….…[2](#_9ap8iagbpr76)

Business Background…………………………………………...[4](#_b19b4vc0a7o5)

Goals of the Project……………………………………………..[6](#_emug0jdayvug)

Details of Work performed………………………………….…[11](#_wc2ll5fo0r06)

Analysis of Result………………………………………………6

Recommendations & Lessons………………..……………..…..7

Future Considerations…………………………………………..8

Appendices…………………………………………………..….9

Reference………………………………………………….…..10

# Executive Summary

By using a combination of machine learning and exposure rating, we are able to help Columbia Insurance Company(CIC) classify the risk levels in their product lines and create an improved pricing algorithm for its products. The business issue that CIC currently faces, the specifics of our solution and analysis, and recommendations and future considerations are provided in this report.

In the past, CIC has mostly employed aggregate loss modeling and monte carlo simulation for its pricing strategy, with satisfactory results. However, the same pricing strategy has seen a decline in performance metrics with the introduction of the new product lines in commercial properties, auto, marine cargo, and aviation hull since 2018. This decline can be attributed to the insufficient claim experience in the new product lines, as experience rating typically requires three years of loss data.

In cases where there is a want of claim experience to perform experience rating, exposure rating becomes an invaluable tool for us to look at the claim experience of another portfolio of the same kind. In using exposure rating, we can derive the missing claim experience from a reference portfolio whose claim experience is sufficiently supported statistically, and we can use the reference portfolio to derive a better loss distribution for the insurance products. In our analysis, the reference portfolio we employed comes from the Maxwell-Boltzmann, Bose-Einstein, and Fermi-Dirac distribution (MBBEFD). The MBBEFD is a well known distribution in statistical mechanics, and it’s well adapted at modeling loss on the interval [0,1] and on the interval [0,∞]. When the parameters of 1.5, 2, 3, 4, respectively, are put into the MBBEFD distribution, the outputs are the exposure curves that approximate the four SwissRe curves used in practice by professionals to model loss. The four SwissRe curves will be referred to as y1, y2, y3, and y4, with y1 being the least risky and y4 being the most risky, and it is these four curves that will be used as the exposure curves to model the loss distribution of CIC’s products.

In the industry, the standard practice is to look at the probable maximum loss (PML), namely, the maximum loss an insurer would expect to incur on a policy, and assign the exposure curves based on arbitrarily decided ranges (e.g. 0-4 million in y1; 4-10 million in y2, etc). It is our opinion that this method falls short of the standards of work we aim to achieve. As such, we have opted to include in the extra step of creating a clustering model to first classify the risk classes for each policy, and then use the aforementioned exposure curves for pricing. Our clustering model uses features such as property value, Probable Maximum Loss (PML), deductible, and the Coefficient of Variation (CV) of the claim amount of each property type. More details on the model are included in the details of work performed section.

Through our analysis, we have found that four risk classifications work best in accurately representing the risk levels of CIC’s product lines. We have also developed a pricing model in python, wherein CIC can input new polices’ property type, property value, PML, deductible, and receive an output of price as a percentage of risk premium.

To improve the quality of our risk classification model, we recommend CIC collect and provide more comprehensive claims data, including geographic location (high risk, hurricane, tornado, proximity to high risk area), age of property, in the case of cars and aviation, age of drivers, and in the case of house, construction type, in the case of buildings, number of floors, property age, etc. In general, the more granular the data provided, the better the clustering model and the pricing model downstream perform. In addition, we also recommend that CIC continuously monitor claim experience and update the model every 2 years using new claim data to keep up with any changes in the market.

# 

# Business Background

Columbia Insurance Company, which started its operation in 2004, has been a small insurance company that primarily offers homeowner’s insurance. In the past, CIC has used aggregate loss modeling and monte carlo simulation as part of its pricing strategy. The company has seen significant growth in its homeowner’s product. As such, CIC has decided to open up new product lines in commercial property, auto, aviation hull, and marine cargo in 2018. Since the introduction of the new product lines, however, the experience rating pricing strategy that CIC traditionally used has been performing suboptimally, and the company has seen a decrease in profitability.

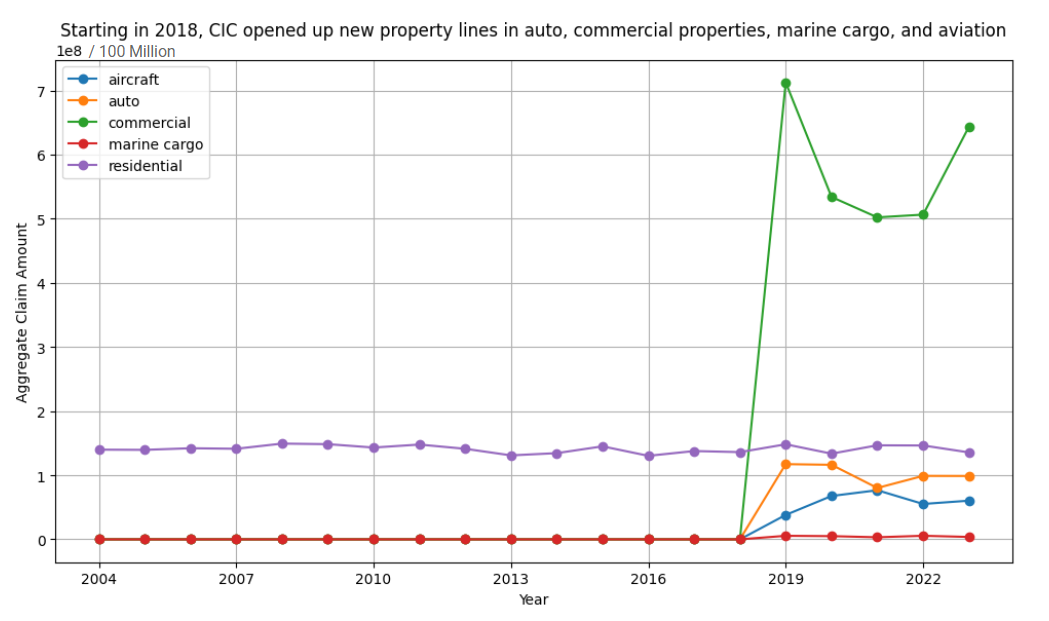
Figure 1: Aggregate claim amount for each product line by year

Figure 1 shows the aggregate claim amount for each product line from 2004 to 2023. Note that only residential (i.e. homeowner’s insurance), as indicated by the purple line, has value above zero prior to 2018, as the other product lines were first introduced in 2018. We can see that the residential property line has been mostly steady and showing little growth over the past 20 years. There is no evidence to suggest that the new product line has grown much either.

This stagnant growth trajectory for these product lines are particularly problematic when compared to the market landscape. In contrast to CIC’s performance, the forecasted growth rates for various insurance markets are looking notably optimistic. The global marine cargo insurance market is projected to grow at a compound annual growth rate (CAGR) of 5.86%, reaching US$34.0 billion by 2032. Similarly, the aircraft insurance market is expected to expand at a CAGR of 2.98%, increasing from USD 14,505 million in 2022 to USD 18,893 million by 2031. Meanwhile, the auto insurance market, valued at USD 652.5 billion in 2021, is anticipated to grow at a robust CAGR of 8.7%, reaching USD 1,383 billion by 2030. Furthermore, the commercial property insurance sector is projected to achieve a CAGR of 11.3%, with its market size soaring from US$254.9 billion in 2022 to US$724 billion by 2032. These figures underscore the positive outlook for these sectors, highlighting the problematic nature of CIC's stagnant growth compared to their dynamic expansion.

In light of the challenging landscape faced by CIC, our team was hired to develop an alternative pricing strategy aimed at revitalizing CIC's profitability. The insurance market as a whole is experiencing significant growth, with sectors like marine cargo, aviation hull, auto, and commercial property insurance all forecasting impressive compound annual growth rates. These figures underscore a positive outlook for these sectors, highlighting the problematic nature of CIC's stagnant growth compared to their dynamic expansion.

Our analysis indicates that CIC's current pricing strategies are not aligned with market trends, resulting in a stagnation that is becoming increasingly problematic as competitors capitalize on expanding opportunities. To address this issue, our team has conducted an in-depth market analysis and identified several strategic pricing adjustments that could help CIC regain its competitive edge and return to a path of profitability.

# Goals of the Project

Project Goal 1: Transition from Experience Rating to Exposure Rating

To help CIC find a solution to its business issue, it’s essential to first identify the root causes of these challenges. Our analysis indicates that CIC’s reliance on experience rating is a significant factor contributing to its current difficulties. Although experience rating has historically performed well for CIC’s homeowner insurance products, it requires at least 3 to 5 years of loss data to be effective. This reliance on experience rating has become problematic for CIC's new property lines, where sufficient historical loss data is unavailable, leading to inaccurate pricing models and financial performance issues.

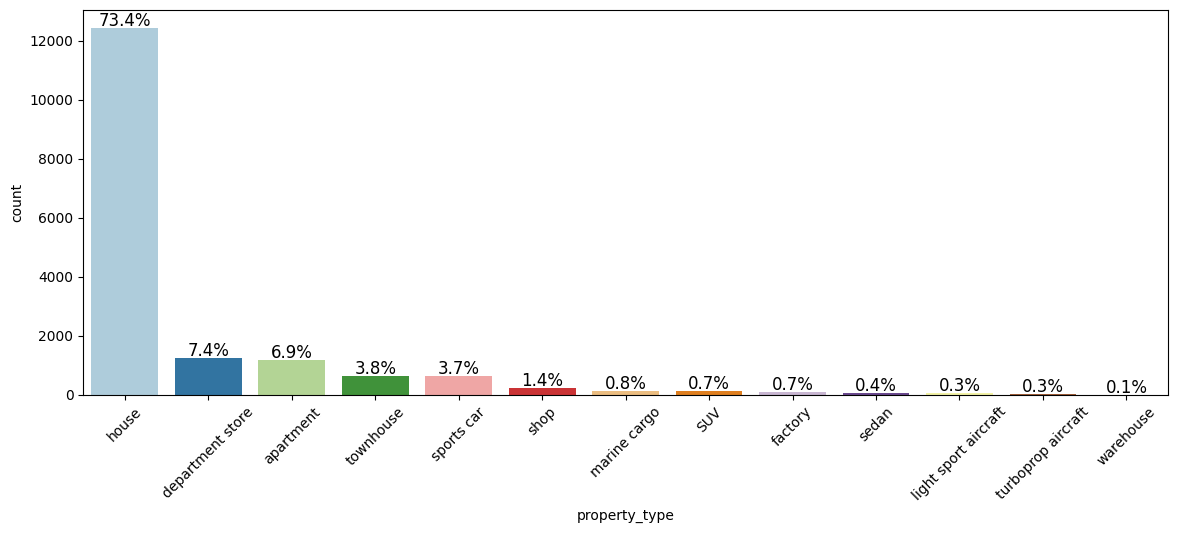


Figure 2: CIC’s Proportion of Products

Figure 2 above shows that the majority of CIC’s products are in homeowner’s insurance. To address this lack of claim experience data in CIC’s new products, our first project objective is to develop a method for utilizing exposure rating. Exposure rating serves as an excellent supplement to experience rating and becomes vital as an alternative in situations where there is a deficiency of sufficient claim data, as is the case with CIC's newer offerings. By using exposure rating, we aim to derive the missing claim experience from a reference portfolio that is statistically robust, thereby allowing CIC to make more informed pricing decisions.

For this purpose, we will employ the Swiss Re Exposure Curves as our reference portfolio. These curves are derived from the Maxwell-Boltzmann, Bose-Einstein, and Fermi-Dirac distribution, collectively known as the MBBEFD distribution. Widely recognized in physics, the MBBEFD distribution is adept at modeling loss and provides a strong foundation for insurance pricing models.

By inputting specific parameters into the MBBEFD distribution, we can approximate the Swiss Re exposure curves, labeled as y1, y2, y3, and y4. These curves represent different levels of risk, with y4 modeling the least risky policies and y1 modeling the most risky ones. Developed by Swiss Re, one of the world’s largest reinsurance companies, the Swiss Re Exposure Curves have been empirically validated to accurately model loss based on real-world data. These four curves will serve as the cornerstone for modeling our loss distribution, ensuring that CIC's pricing strategy is both precise and competitive.

In summary, by transitioning to exposure rating and leveraging the Swiss Re Exposure Curves, CIC can develop a more reliable pricing model that reflects true market conditions and diverse risk profiles. This approach will enable CIC to improve pricing accuracy, enhance competitiveness, and restore profitability across its new property lines.

Project Goal 2: Implementing Advanced Risk Classification through Machine Learning

To effectively utilize exposure rating, it is crucial to first undertake a comprehensive risk classification process. The core concept of risk classes is that policies with similar exposure to a particular risk form a risk class. Each risk class is then modeled by a corresponding exposure curve, enabling more accurate risk assessment and pricing strategies. In the insurance industry, the standard practice for risk classification involves examining the Probable Maximum Loss (PML) and assigning risk classes based on that analysis. For instance, in a product line where policy PML ranges from 0 to 10 million dollars, policies with a PML from 0 to 2 million might be assigned to risk class one and modeled using the Swiss Re y1 curve. Similarly, policies with a PML between 2 to 5 million could be assigned to risk class two and modeled using the Swiss Re y2 curve. This method relies heavily on extensive datasets to accurately determine risk thresholds, which can present challenges when historical data is scarce, as is the case with CIC.

At Team Arocha, we acknowledge this limitation and are committed to devising a solution that extends beyond conventional methods. The second part of our project objective focuses on implementing unsupervised machine learning in our risk classification process to overcome the constraints posed by limited data. Specifically, we will utilize the k-means clustering model, incorporating features such as property value, PML, deductible, and the coefficient of variation of claim amounts for each property type. Figure 3,4,5, and 6 shows our exploratory data analysis on our features.

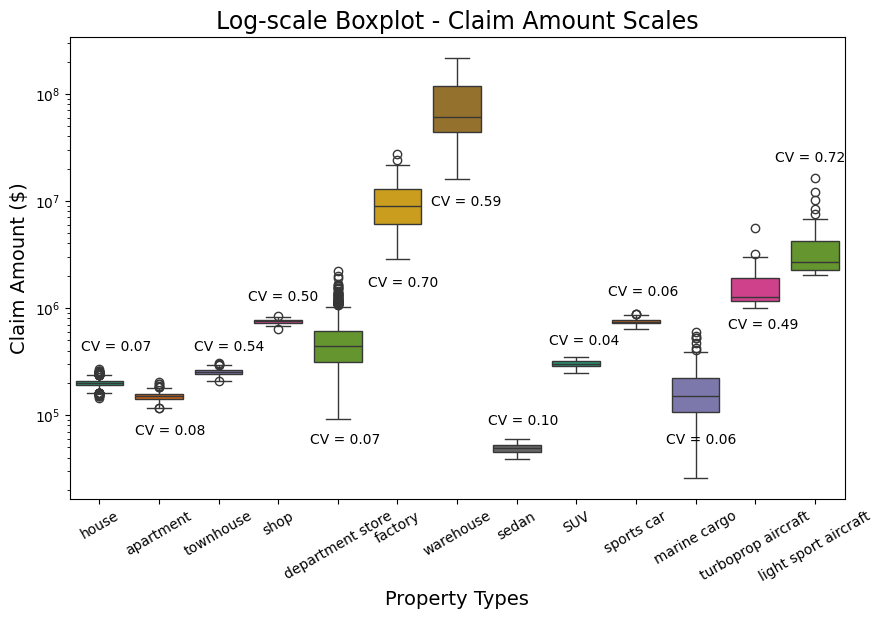


Figure 3: Coefficient of Variation (CV) of Claim Amount for Property Type

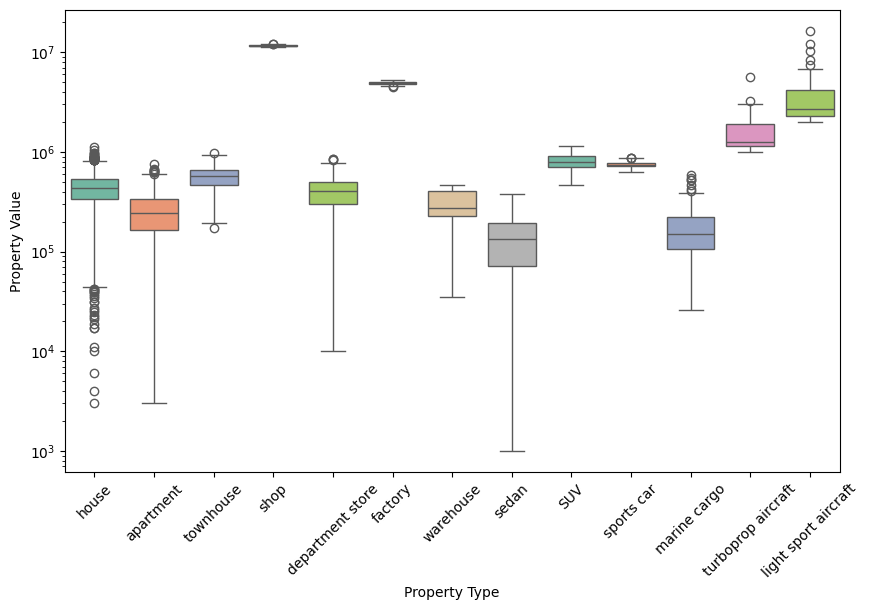


Figure 4: Boxplot of Property Value for each Property Type

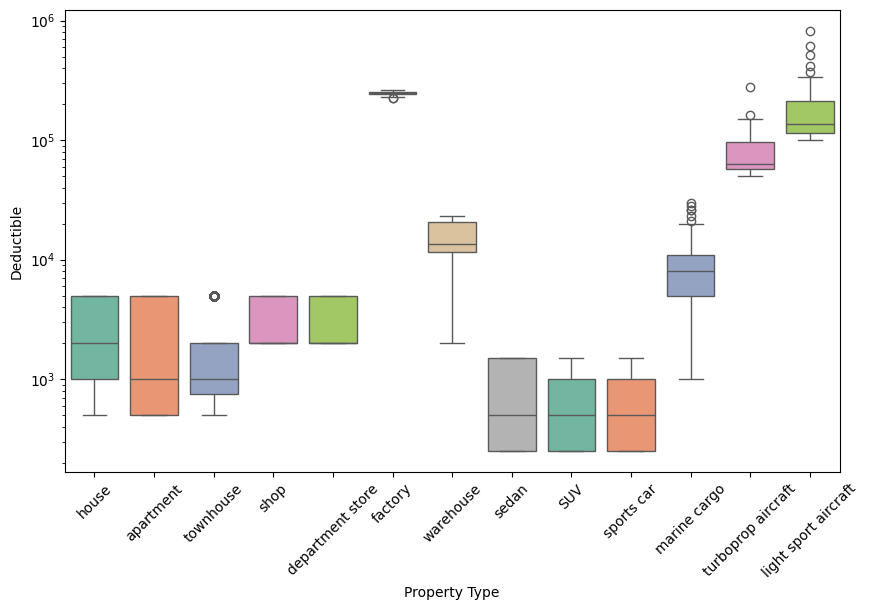


Figure 5: Boxplot of Deductible for each Property Type

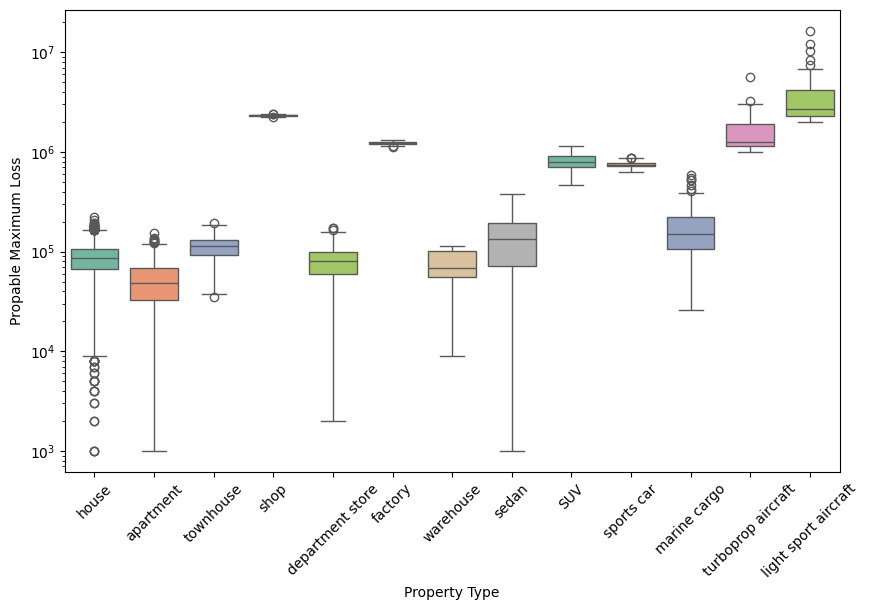


Figure 6: Boxplot of Probable Maximum Loss for each Property Type

This innovative approach allows us to leverage both the policy profile and claims profile comprehensively, circumventing CIC's data limitations and enabling more precise risk classification. By integrating these machine learning techniques, we aim to enhance the accuracy and effectiveness of CIC's exposure rating strategy, providing a robust framework for pricing that aligns with modern industry standards.

# 

# 

# Details of Work performed

The advantage of the clustering

pricing model is an improvement

one goal in insurance is to have homogeneous group

Why do we do classification(Doesn’t answer why we don’t do as the industry standard of looking at pml, but should explain the basic idea of risk classification, u can probably use it as an intro)

The basic idea of insurance is that policies that have a similar exposure to a particular risk, form a risk class. Many factors contribute to pricing. To arrive at homogeneous risk classes, the CIC portfolio must be divided into smaller classes. But how many? For example, if three features were used and each of the features had 10 possible values, this would lead to 1,000 risk classes. Many of these classes would contain very few risks, with little or no statistical support. If, on the other hand, the portfolio is not divided, the assumption of homogeneity will not hold. In essence, one underlying claim distribution would be used for policies of very different nature. According to Bühlmann and Gisler [1], “there are no homogeneous classes in insurance”, because each policy is unique, and is a function of a large number of features. But many features are

1. Not quantifiable (e.g., temperament of a driver, safety procedures at a factory, etc.)
2. Difficult to verify (e.g., number of miles driven by a driver)
3. Politically sensitive (e.g., race, gender, etc.)
4. Administratively difficult to keep track of

Our work is an attempt to find a relatively small number of risk classes for pricing purposes, through an unsupervised Machine Learning algorithm, using a limited number of features. The main idea is to form risk categories using “clusters.”

[1] Bühlmann, H and Gisler, *A Course in Credibility Theory and Its Applications*, pages 1—2, Springer Berlin Heidelberg, 2005

# Recommendations & Lessons

Recommendations

To enhance CIC's business and pricing strategies, we recommend that CIC collect and provide more comprehensive data. This data should include the driver’s age for auto and aviation policies, the number of floors and property ages for real estate policies, and geographical information for most policies. With these detailed features, we can perform a thorough rating factor analysis, such as using regression models to identify significant factors for various property types. This will enable us to offer deeper insights that can significantly improve CIC's operations. Additionally, the new data and the results from our rating factor analysis can be utilized to refine and enhance our risk clustering model, leading to more accurate risk assessments and better-informed pricing decisions.

Additionally, we also recommend that CIC continuously monitor claim experience and update the model every two years using new claim data to keep up with any changes in the market. Regularly updating the model will ensure that it remains accurate and reflective of current trends, allowing CIC to adapt to shifts in the industry and emerging risks. By incorporating the latest claim data, the model can provide more precise risk assessments and pricing strategies, enhancing CIC's ability to manage risk and maintain competitive advantage. This proactive approach to monitoring and updating will also help identify new patterns and anomalies early, enabling timely adjustments to policies and practices. Ultimately, a biennial update cycle will support CIC in maintaining a robust, data-driven strategy that is responsive to the evolving insurance landscape.

Lessons

To help CIC further improve their business

Reference

Exposure rating

<https://www.swissre.com/dam/jcr:7137dac0-83a6-4cfa-80a4-93d33c35562f/exposure-rating-brochure.pdf>

Number of years of claim experience needed

<https://www.ncci.com/Articles/Documents/UW_ABC_Exp_Rating.pdf>

MBBEFD

<https://www.casact.org/sites/default/files/2021-03/8_Bernegger.pdf>

Loss sensitive features

<https://www.actuaries.org.uk/system/files/documents/pdf/mata_0.pdf>

Marine Cargo Growth <https://finance.yahoo.com/news/marine-cargo-insurance-market-set-133000167.html>

Commercial Property Growth <https://www.insurancebusinessmag.com/us/news/breaking-news/commercial-property-insurance-market-to-hit-us724bn-by-2032-458055.aspx#:~:text=The%20commercial%20property%20insurance%20industry,report%20from%20Allied%20Market%20Research>.

Auto Growth <https://straitsresearch.com/report/auto-insurance-market#:~:text=The%20global%20auto%20insurance%20market,of%20an%20accident%20or%20theft>.

Aviation Growth <https://finance.yahoo.com/news/aircraft-insurance-market-size-expected-143500313.html>