Data Scientist - Technical Assessment

Executive Summary

This report presents a comprehensive analysis aimed at assisting the Supervision Manager in strategically allocating supervisory resources to insurance firms. The task involves evaluating firms based on three key characteristics: firm size, changing business profile, and outliers from the norm. The analysis utilises data from two datasets, combining them into a unified format for clarity and consistency.

Key Findings

1. **Firm Size:** Focusing on Gross Written Premium (GWP) as a metric for firm size, the analysis identifies a concentrated group of top firms dominating a significant portion of the market. A Pareto chart effectively highlights the most impactful firms, offering a practical strategy for resource allocation.
2. **Changing Business Profiles:** The analysis of the Net Combined Ratio (NCR) reveals insights into the evolving risk profiles of insurance firms. A focus on consecutive NCR year-on-year increases identifies 38 firms facing potential financial distress. Recommendations emphasise dynamic resource adjustments to address emerging challenges.
3. **Outliers from the Norm:** Examination of the relationship between SCR coverage ratio and gross claims incurred identifies high-risk firms requiring immediate attention. A scatter plot visually communicates the urgency of supervisory intervention for firms with low SCR coverage ratios and high claims.

Error Detection and Reporting

The report conscientiously identifies and addresses data quality issues, ensuring transparency and reliability in the analytical process. Recommendations for further analysis and considerations for future work contribute to a comprehensive view of the data.

Recommendations

The report concludes with actionable recommendations for the Supervision Manager:

- Periodic reviews of firm size allocation based on GWP updates.
- Dynamic adjustment of resources for firms with changing business profiles.
- Focused intervention for outliers identified in SCR coverage ratio and gross claims incurred.

Conclusion

In conclusion, the report provides a data-driven foundation for the Supervision Manager to allocate resources effectively, proactively address emerging risks, and safeguard the stability of the insurance market. The suggested strategies aim to optimise regulatory oversight in a dynamic and complex insurance environment.

Introduction

This report addresses the task given in the Technical Assessment brief. A Supervision Manager has asked for help identifying firms which their team should focus on. They have identified three key criteria for the allocation of supervisory resources and have provided data about different insurance firms. This report will use this data to suggest several groups of firms that the Manager should focus on

Basis of preparation

The data provided is split between two different tables. `Dataset 1 - General` contains key metrics from the financial statements of each firm (including the statements of profit and loss and of financial position) and `Dataset 2 - Underwriting` contains information about the insurance underwriting of the firms. Both files contain a number of different metrics for each firm from 2016 to 2020. Due to the similar structure of these two datasets, I decided to combine them into a single flat file containing four columns:

- `Firm` is the firm that the row relates to
- `Year` is the year that the row relates to
- `Metric` is the name of the metric that the row relates to
- `Value` is the numerical metric value that the relevant firm had at the given year-end

At this stage I note that there are 325 firms that exist in both datasets. `Dataset 2 - Underwriting` additionally contains data on 131 firms that are not featured in `Dataset 1 - General`.

Across the two datasets, there were no missing entries. However, I did note that 34% of records had a `Value` equal to `0`. This prompted an investigation into the possible reasons why. Because the majority of zero values occur consecutively at either the first or last years for most firms, I have concluded that most of the zero values are set accordingly because the firm was not trading during that particular year (e.g. it started trading after 2016 and/or ceased trading before 2020). The Supervision Manager may or may not want to allocate resources to these closed firms, so I have not filtered them out at this stage. For example, they may be required to monitor or report on the financial status of the closed firm. Additionally, the closed firms also provide useful benchmarking data.

There were some instances of firms that had certain metrics that had zero values nested between non-zero values. These are trickier to explain away, and are likely a result of missing data. Since the number of values affected by this is low (51), I decided to use a simple means of imputation for these metrics. I took the previous non-zero value and the next non-zero value for each record (taking into account the `Firm` and `Metric`) and replaced the zero value with the average of these two values. This means that we do not have to simply discard these firms.

In summary, the data preparation steps that I performed are:

1. Load the two datasets
2. Transform the datasets into the flat file format described above
3. Imputed values for rows which had a `Value` equal to `0` by taking the average of the previous and next relevant record

Allocation of supervisory resources

This section of the report will focus on how to allocate supervisory resources considering the three characteristics: firm size, changing business profile, and outliers from the norm. We will consider each characteristic separately but will discuss how a more holistic assessment can be made in the conclusion.

To ensure a comprehensive and well-rounded approach to resource allocation, the Supervision Manager should consider integrating the findings from each characteristic. A holistic assessment can be achieved by weighing the significance of each factor in the context of overall risk and regulatory objectives.

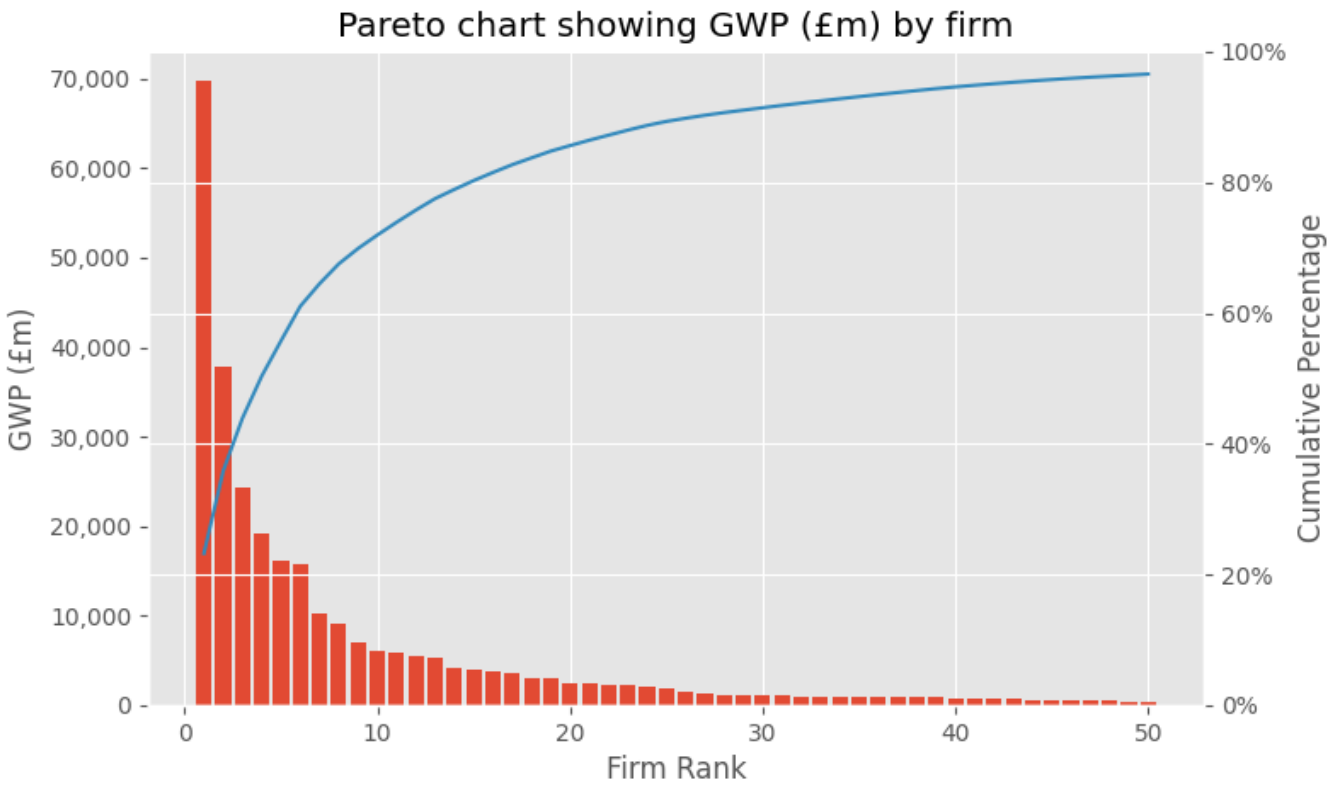
For each characteristic, I have chosen to focus on just one approach. Given the data available, there are many valid approaches but to address them all would result in a lengthy report. In all cases, I have justified why I settled on that particular approach.

Firm size

A goal of supervisory resource allocation is that bigger firms receive more attention. It therefore makes sense to focus on a small group of firms that constitute a large proportion of the insurance market. Although bigger firms will require more resources per firm, this is likely to be less than the number of resources required to supervise many tiny firms. I decided that I would choose a metric that describes the size of an insurance firm and then identify the largest firms in the market.

There are several metrics that I considered in order to best quantify firm size including gross written premium (GWP), total assets, total liabilities, and total equity. GWP provides a better focus on revenue generation but the other metrics better quantify financial structure. I decided to focus on GWP. GWP reflects the total amount of premiums collected from policyholders. A higher GWP should indicate a larger volume of business since a firm's ability to underwrite and manage a significant volume of policies is a key aspect of its size.

Further, I decided to focus on data from 2020 only (the most recent year). This will ensure that resources are allocated using the latest up-to-date information. If a firm is no longer trading then it will not have any GWP for this year. The Manager may wish to review these firms for regulatory purposes, but they will not be considered here and have been filtered out.



The above chart shows the top 50 firms ranked by GWP (there are 325 firms under consideration in total but are immaterial in comparison to the largest). The left-hand y-axis shows the GWP for each firm and the right-hand y-axis shows the cumulative percentage of total GWP represented. It is clear that a very small number of firms dominate a huge percentage of the market. A simple strategy for selecting firms to focus on would be to choose firms from the top, knowing that so long as you select at least 4 firms then you will be allocating supervisory resources to at least 50% of the market. The table below shows the data underlying this chart for the top 15 firms.

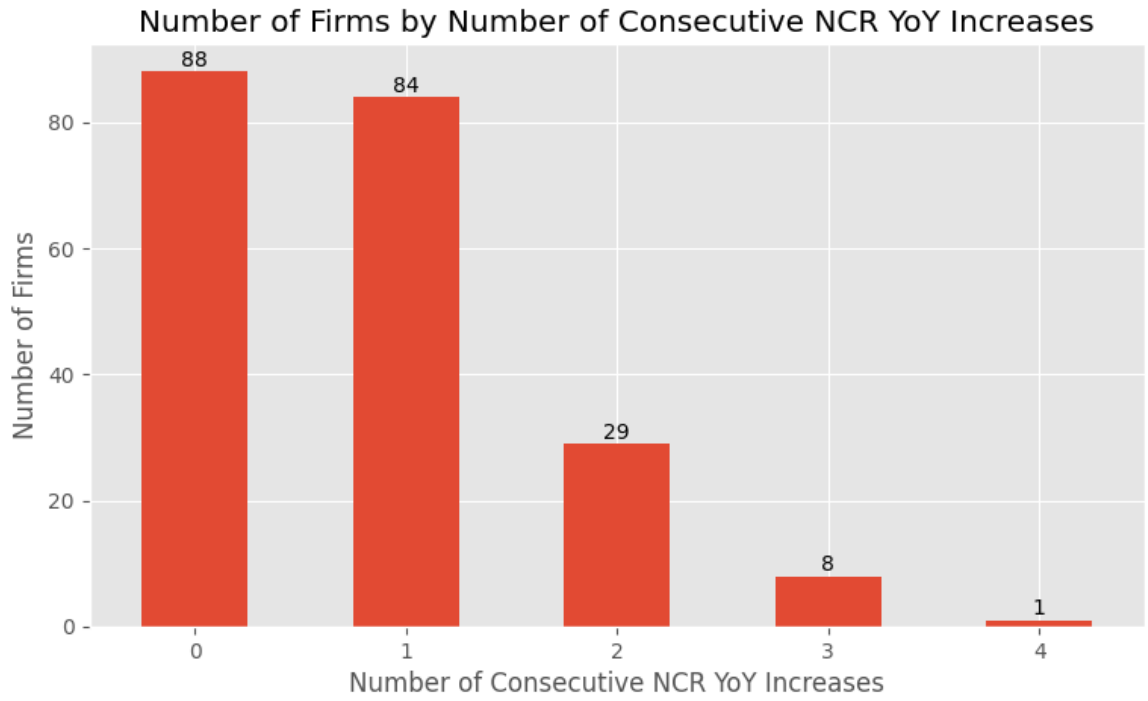
Rank	Firm	GWP 2020 (£m)	Cumulative %
1	Firm 210	69,698	23.3%
2	Firm 4	37,762	35.9%
3	Firm 311	24,251	43.9%
4	Firm 34	19,275	50.4%
5	Firm 7	16,184	55.8%
6	Firm 151	15,825	61.1%
7	Firm 26	10,352	64.5%
8	Firm 247	9,115	67.6%
9	Firm 25	7,032	69.9%
10	Firm 73	6,174	72.0%
11	Firm 105	5,857	73.9%
12	Firm 276	5,587	75.8%
13	Firm 104	5,257	77.5%
14	Firm 6	4,209	78.9%
15	Firm 234	4,052	80.3%

Changing business profile

The risk profile of a business can change over time. The Supervision Manager will want to ensure that firms that are facing increasing risk over time receive more attention so that this risk can be better understood and managed. Risk may be quantified with several metrics such as net written premium (NWP) or the net combined ratio (NCR).

I have chosen to analyse the NCR in more detail. NCR is a metric, expressed as a percentage, that measures the profitability of an insurance firm's underwriting operations. If the NCR is below 100% then this indicates that a firm is making a profit. This is essential for the long term sustainability and solvency of the business. A consistently increasing NCR may signal potential financial distress for an insurance firm. The Supervision Manager can therefore use this metric as an early warning system to identify firms that might be facing challenges in their underwriting operations. However, a high NCR will not always be a problem. Indeed, the business may have made a strategic choice to operate at a lower profitability level, such as to promote growth.

If a firm's NCR has been increasing year-on-year for a number of consecutive years, then the Supervision Manager may be concerned about this firm's performance and want to allocate additional resources. I calculated a metric equal to the number of consecutive years of NCR increases (prior to 2020—the most recent date in the dataset), referred to as Number of Consecutive NCR YoY (year-on-year) Increases. The chart below shows the number of firms for each by their number of consecutive NCR YoY increases (excluding 217 firms that have no NCR data in 2020).



88 firms did not have an increase in NCR between 2019 and 2020, and the Supervision Manager may have fewer concerns about these firms. Many firms, 84, did have an increase in NCR between 2019 and 2020, but they did not experience an increase the year prior. There could be many benign reasons for this, including minor variation in the NCR between years. Again, the Supervision Manager is unlikely to be concerned about these firms.

Some firms have 2 or more consecutive NCR YoY increases. This is indicative of an insurance firm that is struggling to make a sustainable profit on its underwriting operations. These organisations could be in financial distress and therefore the Supervision Manager should pay close attention to the 38 firms that fall into this category. In particular, they should direct resources to the 8 firms that have experienced 3 years of consecutive increases to the NCR and the 1 firm that has increased every year between 2016 and 2020. The most concerning 15 firms are included in the table below, selected on the basis of having the highest number of consecutive NCR YoY increases and the highest NCR in 2020. The Supervision Manager should consider diverting more resources to these firms. Note that the columns `NCR 20XX` correspond to the value of the NCR in the given year form that firm.

Firm	Number of Consecutive NCR YoY Increases	NCR 2016	NCR 2017	NCR 2018	NCR 2019	NCR 2020
Firm 19	4	30.8%	144.8%	166.3%	173.3%	185.2%

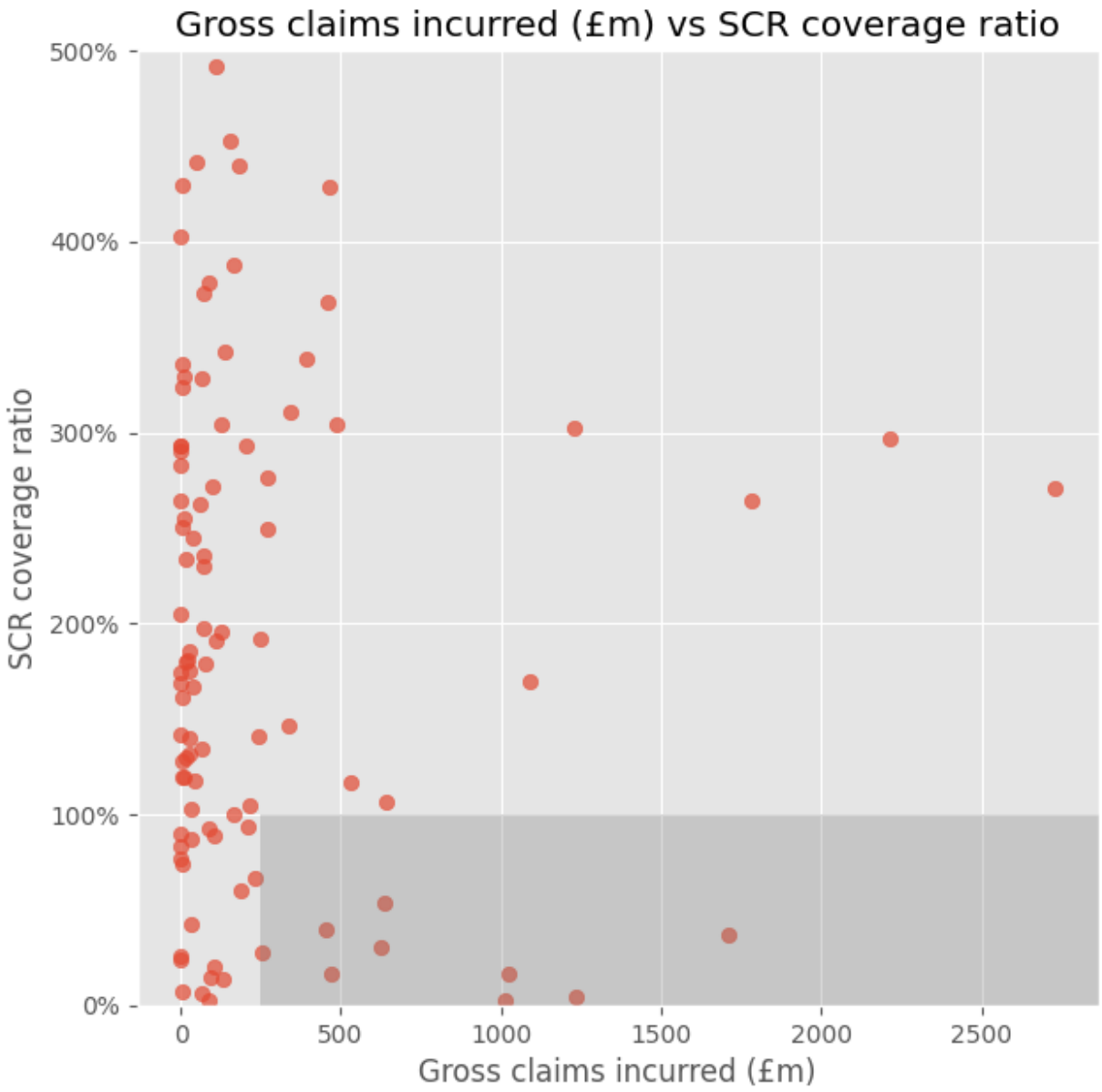
Firm	Number of Consecutive NCR YoY Increases	NCR 2016	NCR 2017	NCR 2018	NCR 2019	NCR 2020
Firm 394	3	0.0%	0.0%	56.8%	121.3%	275.1%
Firm 311	3	0.0%	0.0%	23.0%	170.5%	189.9%
Firm 22	3	177.6%	78.5%	92.8%	162.4%	178.0%
Firm 102	3	83.3%	66.0%	129.1%	137.9%	167.8%
Firm 344	3	0.0%	0.0%	1.7%	118.4%	145.5%
Firm 349	3	52.0%	10.3%	31.3%	136.9%	144.7%
Firm 230	3	130.5%	91.2%	119.7%	119.8%	131.4%
Firm 198	3	66.6%	50.7%	95.9%	101.0%	118.5%
Firm 203	2	108.4%	135.3%	63.8%	123.8%	365.4%
Firm 89	2	0.0%	239.6%	19.9%	152.5%	239.3%
Firm 300	2	184.2%	187.9%	114.6%	149.7%	232.5%
Firm 91	2	0.0%	178.3%	124.9%	179.1%	215.5%
Firm 92	2	0.0%	91.6%	17.7%	36.2%	200.4%
Firm 287	2	77.0%	150.1%	0.5%	158.9%	191.3%

Outliers from the norm

The Supervision Manager will be interested in identifying firms that deviate significantly from the norm. When looking at what value is 'expected' for any given metric it is important to consider it in the context of other data points. For example, a high value in one column may be normal if there is a high value in another column but unusual if that other column has a low value.

I have decided to consider two columns: the SCR coverage ratio and the gross claims incurred. SCR coverage ratio gives an indication of whether a firm is meeting its prudential capital requirements whereas gross claims

incurred represents a significant cost to an insurer. By analysing the relationship between these two variables we can comment on each firm's financial stability, capital adequacy, and risk management.



In the above chart, each dot represents a firm. The shaded region in the bottom right represents the most concerning firms that have a low SCR coverage ratio but high gross claims incurred. These firms may have underestimated risk or do not have adequate capital reserves to cover potential losses. This could place the firm under strain if it faces a surge in claims. The Supervision Manager should closely monitor such firms and allocate the maximum resources to them to minimise risk. The 9 firms with an SCR coverage ratio below 100% and gross claims incurred of over £250m are listed in the table below.

Firm	Gross claims incurred (£m)	SCR coverage ratio
Firm 286	1,711	37.3%
Firm 234	1,232	4.7%

Firm	Gross claims incurred (£m)	SCR coverage ratio
Firm 210	1,024	16.4%
Firm 297	1,014	2.3%
Firm 165	634	53.6%
Firm 26	625	31.0%
Firm 118	470	16.8%
Firm 230	452	39.7%
Firm 227	253	27.7%

A low SCR coverage ratio combined with low gross claims incurred also raises significant concerns about an insurance firm's financial stability and may require prompt supervisory intervention to mitigate potential risks and protect policyholders. The manager may consider allocating resources to these firms if they believe that the low level of claims may be masking the financial vulnerability resulting from the inadequate capital position. However, since these firms will tend to be smaller than the previous group they may consider them a lower priority.

Firms with low gross claims incurred and a high SCR coverage ratio may also be worthy of attention. Although this would generally indicate a favourable position since the firm has a surplus of capital beyond regulatory requirements and demonstrates effective risk management, they still represent unusual deviations from the norm. The manager may have an opportunity to acknowledge the firm's financial strength and encourage the continuation of sound risk management strategies.

Error Detection and Reporting

The datasets used to create this report were generally clean and suitable for purpose. However, there were some minor data quality issues. Some of these issues faced have already been discussed, but a full list of data quality issues noted can be found below. Please note that this list only mentions issues identified in the analysis to generate this report—there may be additional issues impacting other columns.

- There are 325 firms that exist in both datasets. `Dataset 2 - Underwriting` additionally contains data on 131 additional firms that are not featured in `Dataset 1 - General`.
- 34% of records have a `Value` equal to `0`. Most of these values occur consecutively at either the first or last years for most firms and are reasonably assumed to represent years when the firm was not trading.
- There are 51 instances of firms that have certain metrics that had zero values nested between non-zero values. These are assumed to be the result of missing data. They have been imputed by taking the previous non-zero value and the next non-zero value for each record (taking into account the `Firm` and `Metric`) and replacing the zero value with the average of these two values.
- The `NWP (£m)` metric is erroneously named in the raw data. It has a trailing space after the text.
- There are 4 firms with negative GWP in 2020, which do not make logical sense. However, these were immaterial to the analysis conducted (the most negative value was approx. -£8.5m) on GWP so were left untreated.

- Some values for NCR are impossibly large. 11 firms that were affected by this have been removed from the analysis.
- Some values for NCR are less than 0. 23 firms that were affected by this have been removed from the analysis.
- Some values for the SCR coverage ratio in 2020 are very high. In particular, Firm 127 has an SCR coverage ratio of over 8,000,000%. This number is clearly incorrect and this firm was ignored in the analysis.

Recommendations and Conclusion

The analysis conducted on the insurance firms' data highlights specific areas that warrant the Supervision Manager's attention for effective resource allocation. The three key characteristics—firm size, changing business profile, and outliers from the norm—provide valuable insights into potential areas of concern and opportunities for proactive supervision.

To ensure a comprehensive and well-rounded approach to resource allocation, the Supervision Manager should consider integrating the findings from each characteristic. If more time was available, it would be interesting to investigate how a holistic assessment can be achieved by weighing the significance of each factor in the context of overall risk and regulatory objectives.

Key Recommendations

1. Periodic Review of Firm Size Allocation

Regularly reassess the allocation of supervisory resources based on firm size, considering updates to GWP and other relevant metrics. This ensures that the oversight strategy remains aligned with the evolving landscape of the insurance market.

2. Dynamic Adjustment of Resources for Changing Business Profiles

Implement a dynamic resource allocation strategy for firms with evolving risk profiles. Conduct periodic assessments of NCR trends and adjust resource allocation to address emerging challenges in underwriting operations effectively.

3. Focused Intervention for Outliers

Develop a targeted intervention plan for outliers identified in the SCR coverage ratio and gross claims incurred analysis. Tailor resources based on the severity of financial instability, giving priority to firms with a higher likelihood of facing challenges in meeting their capital requirements.

Conclusion

In a dynamic and complex insurance environment, a one-size-fits-all approach to supervisory resource allocation may prove insufficient. By holistically considering firm size, changing business profiles, and outliers, the Supervision Manager can adopt a nuanced and adaptive supervisory strategy. This report has demonstrated how the Manager is able to use a data-driven approach to ensure the effective use of resources and also position regulatory oversight to proactively address emerging risks and safeguard the stability of the insurance market.

Annex - Task II - Machine Learning

In order to complete Task II, I first considered several alternative opportunities to use machine learning. Then, I selected one of these and built a proof-of-concept solution in Python.

Possible Approaches

I considered several possible approaches using machine learning techniques that might draw additional insights for the Supervision Manager:

- **Regression Analysis for SCR Coverage Ratio:** Use regression analysis to understand the relationship between various metrics (e.g., GWP, NWP, Gross claims incurred) and the SCR coverage ratio. This can help in predicting whether a firm is likely to meet its prudential capital requirements and allow the manager to identify firms that might be at risk of not meeting them in the future.
- **Cluster Analysis on Profitability:** Apply clustering algorithms (e.g. K-means) to form groups based on their profitability to identify clusters of firms with similar profitability profiles. The Supervision Manager can pinpoint clusters of firms that are consistently profitable or face challenges in maintaining profitability and tailor strategies accordingly.
- **Time Series Analysis for Gross Claims Incurred:** Utilise time series analysis to identify patterns and trends in gross claims incurred over time. This can help anticipate financial stress and allocate resources to firms with increasing claims over time.
- **Ensemble Learning for Overall Ranking:** Implement ensemble learning techniques (e.g. Random Forest) to combine predictions from different metrics and create an overall ranking model that can take into account multiple metrics simultaneously. This can guide the Supervision Manager in prioritising attention based on a holistic view of each firm's performance.
- **Feature Importance Analysis:** After training a machine learning model, analyse feature importance to identify which metrics contribute the most to predicting certain outcomes (e.g. profitability, meeting capital requirements). This information can help guide the Supervision Manager in focusing on the most critical aspects during their supervision activities and more efficiently allocate resources.
- **Anomaly Detection for Outliers:** Implement anomaly detection algorithms (e.g. Isolation Forest) to identify firms that deviate significantly from the norm in a single reporting period. The Supervision Manager can quickly identify and prioritise resources to unusual behaviours.

Proof of Concept

Building a robust machine learning solution takes time to engineer descriptive features and fine-tune the model. I decided to implement a proof-of-concept for one of these approaches. I decided to focus on anomaly detection for outliers, because it is a relatively straightforward method that directly address the aspect of identifying outliers, which is a critical concern for supervision. My full proof of concept can be found in the file `Task II - Machine Learning.ipynb`. This report will summarise my methodology, considerations made, and discuss how the output could be used.

The goal of this proof of concept will be to create an anomaly detection model that the Supervision Manager can use to identify outlying firms and consider whether they should allocate resources to them.

Data Load

I loaded the data in the same way that I loaded it in for Task I. I then filtered out the metrics from `Dataset 2 - Underwriting`. Focusing on a subset of features has two advantages:

1. There will be fewer records with missing values. I will have to deal with any missing values, and this will help simplify my methodology.
2. There are only a few hundred firms in this dataset at most. Therefore, if I was to introduce too many features into the machine learning solution I would be at risk of the 'curse of dimensionality'. Anomaly detection relies on identifying firms that are spatial 'outliers' in the data and increasing the number of features increases the dimensionality of the dataset, which increases the space between firms and makes it harder to identify anomalies.

Because I had already spent time analysing the data in Task I, I decided not to perform additional exploratory data analysis. If I had more time, this is something that I may have considered in order to understand the dataset even better.

Handle Missing Values and Erroneous Data

For this proof of concept, I wanted to maximise the amount of data that I had available whilst using minimal effort, in the interest of time. There are many firms that are no longer trading by 2020, so I decided to take the most recent non-zero value for each firm and metric. This would ensure that my dataset included as many firms as possible (this will introduce some bias, since many firms that have stopped trading would have stopped trading for a reason such as becoming insolvent). Then, I removed any firms that were missing values in any metric.

In Task I, it was identified that some firms have an erroneously high SCR coverage ratio value. I filtered out 5 firms that had an SCR coverage ratio of more than 30 (although this will still leave some incorrect values in).

Feature Engineering

When performing machine learning, it is important to avoid having features that are highly correlated with each other. This helps to eliminate redundant information, reduces multicollinearity, and improves the model's interpretability and performance by avoiding undue influence of correlated predictors on the model training process. Therefore, I selected a few features to bring through to my model as they are in the original data (GWP, total assets, and SCR coverage ratio). I then engineered several new features that express some of the other original features in relative terms:

- **Reinsurance Retention Ratio:** NWP divided by GWP. The proportion of insurance risk retained by the company rather than transferred to a reinsurer.
- **Debt-to-Assets Ratio:** Liabilities divided by assets. A financial metric that measures the proportion of a company's assets financed through debt, indicating the level of financial leverage.
- **Debt-to-Equity Ratio:** Liabilities divided by equity. A financial ratio that compares a company's total debt to its total equity, providing insight into the relative contribution of debt and equity to the company's capital structure.

I performed correlation analysis on these features and found no strong correlations between the features. These features should provide a good overview of the size of a firm, how it is financed, the risk that it takes on, and the adequacy of its capital reserves. They are also simple and interpretable, which will help make the model more explainable. I am therefore happy to proceed with these features for outlier detection.

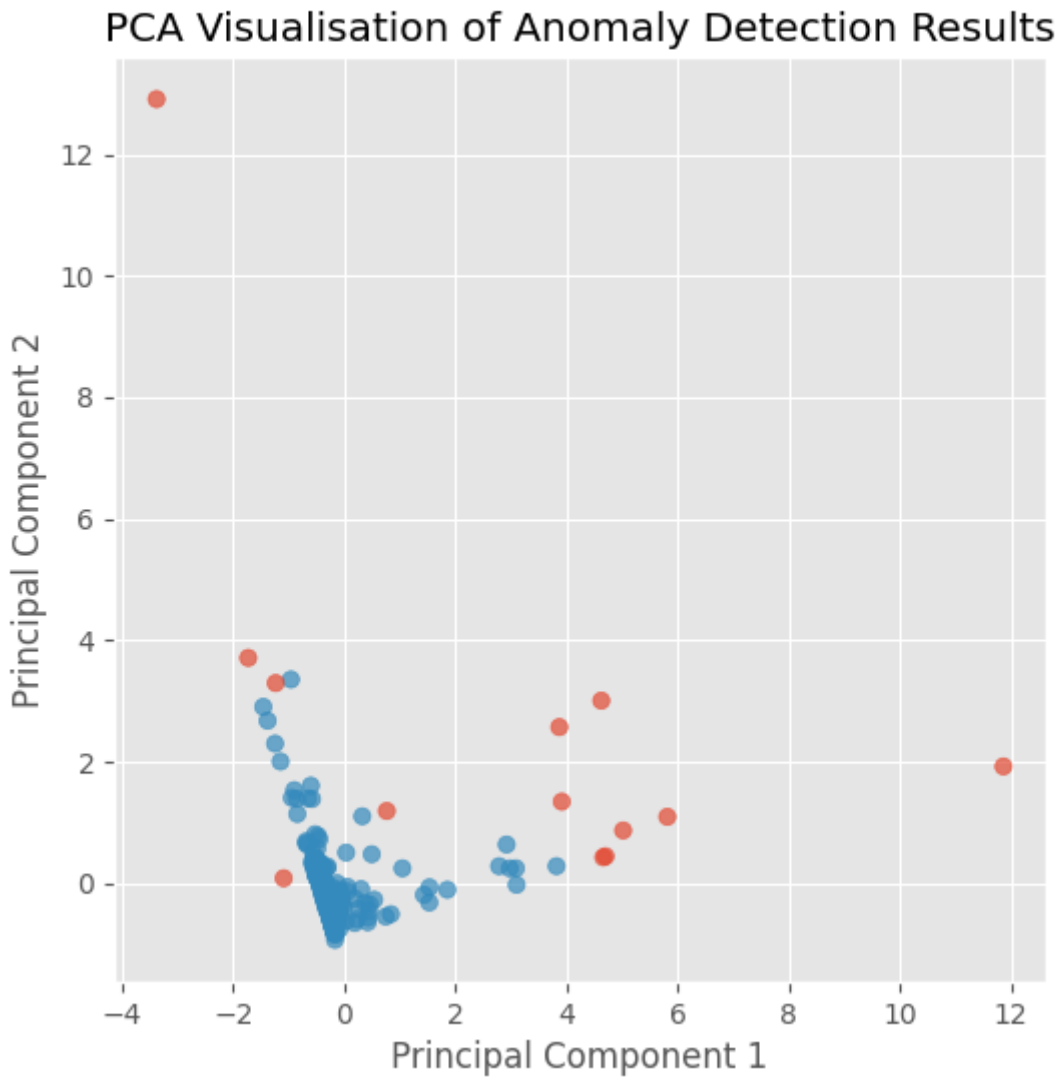
Normalisation

I normalised the features by scaling each feature to have a mean of 0 and a standard deviation of 1. Normalisation is important to ensure that all features contribute equally to the analysis, preventing features with larger scales from dominating the anomaly detection process and allowing the algorithm to focus on patterns and variations rather than absolute magnitudes.

Outlier Detection

I used an isolation forest, implemented in `scikit-learn` in order to detect outliers in the dataset. Isolation forests are well-suited for situations where anomalies are expected to be rare and different from the norm. Their underlying mechanism of randomly partitioning the data into isolation trees and isolating anomalies in shorter paths makes it robust to the presence of outliers. Additionally, isolation forests do not assume any underlying data distribution, making them versatile and applicable to various types of financial datasets. The algorithm's ease of implementation and interpretability align well with the requirements of this proof of concept, making it a pragmatic choice for identifying anomalies for the Supervision Manager to focus on.

After fitting the model, it identified 13 outliers in the dataset. I performed principal component analysis in order to visualise these outliers in two dimensions and you can see that the algorithm did a good job of identifying anomalous firms (highlighted in red). I have also included a table of the 13 firms identified as outliers.



GWP	Total Assets	SCR Coverage Ratio	Reinsurance Retention Ratio	Debt-to-Assets Ratio	Debt-to-Equity Ratio
882.48	195835.58	0.6	0.25	1.3	28.06
15824.69	124326.02	0.58	0.35	0.76	129.51
0.44	0.05	23.51	1.26	87.65	2.83
-0.0	33.81	4.72	4410.46	0.17	0.4
69697.93	160518.7	0.16	0.87	0.08	10.97
9115.19	100085.1	7.97	0.02	0.8	5159.3
7031.52	3614.91	3.16	1.39	8.42	1124.7
-0.0	9.78	24.75	-5.24	0.01	14.01
24251.48	61836.61	1.8	0.6	8.0	61.89
19274.96	160124.46	2.55	0.32	1.14	34.05

GWP	Total Assets	SCR Coverage Ratio	Reinsurance Retention Ratio	Debt-to-Assets Ratio	Debt-to-Equity Ratio
37761.88	37423.54	1.2	1.3	2.47	3.82
2524.62	32271.72	2.43	1.5	0.64	8913.19
219.32	11715.2	21.78	-0.01	0.58	34.17

Although we have here performed a visual inspection of the data, it would be prudent to further evaluate the proof of concept. One approach to this could be getting the Supervision Manager's feedback on these firms in order to evaluate how effective the model was. If there are any issues, then we could fine-tune the parameters of the isolation forest, perform additional feature engineering, or collect data on more firms.

Conclusion

In this exploration of machine learning applications for resource allocation in supervision, various approaches were considered to draw additional insights from the provided metrics. A proof of concept focused on anomaly detection was implemented using an isolation forest algorithm. The primary goal was to identify outliers, allowing the Supervision Manager to pinpoint firms deviating significantly from the norm and allocate resources accordingly. This proof of concept provides a foundation for leveraging machine learning in supervision activities. By iteratively refining the model based on feedback and exploring additional techniques, the Supervision Manager can enhance the effectiveness of resource allocation and decision-making in a dynamic business environment.

Annex - Task III - Cloud Technologies

The task involves analysing insurance firm data and creating an end-to-end data processing and analytics pipeline for daily batch processing so that a Supervision Manager can allocate resources efficiently. This necessitates a robust, scalable, and secure solution to effectively handle the diverse needs of the insurance domain. Before designing an architecture to carry out this report using cloud technologies in Microsoft Azure, there were several considerations that I made.

- **Scalability:** The chosen services should be able to scale based on the data processing needs. Consideration was given to how the architecture handles an ever-growing volume of data and potential increases in the number of firms or metrics.
- **Cost Management:** Azure services have varying costs that often scale with computation requirements. Costs should be monitored and optimised. For example, serverless solutions may be preferable if the pipeline only has to run once a day for a short amount of time.
- **Data Security:** The architecture involves storing and processing confidential firm information, emphasising the need for robust security measures. Azure Entra ID for authentication and role-based access control for authorisation should be considered and integrated.
- **Resilience:** The impact of potential system inaccessibility was considered. Azure services with built-in redundancy and backup options were chosen to ensure data availability.
- **Integration:** Services were selected to ensure seamless integration within the data pipeline, fostering a cohesive and efficient workflow.
- **Compliance:** The Bank may have internal and external regulations governing data handling tasks that need to be taken into account during the design process.

With the above considerations, I propose the following architecture in Microsoft Azure:

1. Data Ingestion - Azure Data Factory (ADF): ADF allows users to create, schedule, and manage data pipelines that move data from various source systems to designated destinations, facilitating the efficient extraction and loading of data. In this context, ADF can be configured to ingest daily batch data from a diverse range of sources. It has a visual interface that simplifies the creation of ELT workflows, making it accessible to users with varying levels of technical expertise. ADF's integration capabilities with Azure Data Lake Storage ensures seamless storage of raw data for further processing and analysis. Its scalability and scheduling features align well with the daily batch processing requirement.

2. Data Storage - Azure Data Lake Storage (ADLS): ADLS is an optimal choice for data storage in this task due to its scalability, high performance, and compatibility with other Azure services. ADLS supports the storage of large volumes of data in an efficient hierarchical file system structure. ADLS integrates seamlessly with other Azure services like Azure Databricks, enabling smooth data processing workflows. Additionally, ADLS provides fine-grained access control and security features, aligning with the need for handling sensitive insurance data securely. As a result, ADLS serves as a reliable and scalable storage solution, ensuring that the pipeline can efficiently store and retrieve data as part of the end-to-end data processing and analytics workflow in Microsoft Azure.

3. Data Processing - Azure Databricks: Azure Databricks is an ideal solution for data processing in the specified task due to its powerful Apache Spark-based analytics platform and seamless integration with Azure services. With Databricks, large-scale data processing tasks can be efficiently executed. Its collaborative environment allows data engineers, data scientists, and analysts across the Bank to work together in a unified workspace. The built-in optimizations for Spark jobs in Databricks enhance performance, making it well-suited for handling the expanding data volumes. The integration with Azure services like Azure Data Lake Storage ensures easy data access and storage. Overall, Azure Databricks provides a scalable, collaborative, and high-performance environment for data processing.

4. Data Visualisation - Power BI: Power BI is a highly effective data visualisation tool, offering intuitive and interactive reporting capabilities. With Power BI, users can connect directly to the stored data in ADLS, and create compelling visualisations. Power BI can handle batch data by setting up refresh triggers that seamlessly integrate with other Azure services. Power BI's sharing and collaboration features also make it easy for stakeholders to access and interpret the insights, making it a well-suited choice for creating informative and accessible reports for the Supervision Manager.

In conclusion, the proposed architecture leverages key Azure services to address the specific needs of insurance metric analysis. It ensures scalability, cost-effectiveness, security, resilience, and seamless integration, aligning well with the task requirements for an end-to-end data processing and analytics pipeline on Microsoft Azure.