Summer Internship Report

Development of a Predictive Customer Segmentation Solution For An Insurance Agency

Submitted by:

Kalthoum Dridi

For the Company:

LLOYD ASSURANCES

4th Year DS8

Academic Year 2023-2024

Acknowledgments

I would like to express my heartfelt gratitude to all those who contributed to the completion of this project.

First, I sincerely thank the stakeholders from the Lloyd Assurances's agency for their support and trust in our work. Your collaboration has been invaluable in guiding this project.

I also extend my appreciation to my professors and academic teachers for their support and guidance, especially in helping me understand machine learning technologies and techniques througout this great year.

Lastly, I will be forever grateful to everyone who offered encouragement and assistance along the way. Your support has been greatly appreciated.

Contents

List of Figures

List of Tables

CHAPTER 1

General Introduction

In an era marked by the constant evolution of data-driven decision-making, businesses across various sectors must explore innovative solutions to unlock the potential of data, foster innovation, and secure a competitive edge. This internship report describes a collaborative project with Lloyd Assurances, a leading insurance company in Tunisia, as we aim to revolutionize customer segmentation and profiling.

At the core of this project is the careful design and implementation of a cutting-edge solution for customer segmentation within one of Lloyd Assurances' agencies. This initiative addresses the challenges of traditional customer segmentation methods and responds to the growing demand for personalized and efficient insurance services. The primary goal is to use data science to create precise customer segments, establishing a streamlined and data-driven approach that mitigates challenges and propels Lloyd Assurances to the forefront of insurance innovation.

Throughout this report, you will follow our journey as we tackle industry challenges, embrace digital transformation, and contribute to Lloyd Assurances' commitment to operational excellence. This narrative spans the convergence of data science and insurance, highlighting our dedication to reshaping the future of customer management in the insurance sector.

As we navigate the complexities of this project, our aim is to provide not just a solution but a comprehensive blueprint for industry stakeholders. We will explore the synergies between data science and insurance expertise, showcasing our commitment to shaping a future where innovation and efficiency seamlessly converge in the financial landscape.

CHAPTER 2

Business Understanding

2.1 Introduction

In this chapter, we delve into Business Understanding, laying the foundation for our collaboration with Lloyd Assurances. Our focus is on understanding the intricacies of the insurance landscape, deciphering challenges, and aligning our efforts with Lloyd Assurances' objectives. This chapter sets the stage for subsequent phases, ensuring a nuanced approach as we strive to deliver a solution that surpasses expectations in the dynamic world of customer segmentation and profiling

2.2 Study of the Existing

In understanding the current landscape, a thorough study of the existing approach employed by one of Lloyd Assurances' agencies lays the groundwork for our transformative journey. This examination delves into the intricacies of current customer segmentation practices, challenges posed by data diversity, and the impact of manual processes.

Some of these aspects are highlighted through the following key elements:

2.2.1 Fragmented Data Sources

The agency's customer data is scattered across various sources such as contracts, bordereaux, insurance letters, claims, and other documents. This fragmentation makes it challenging to obtain a comprehensive and accurate view of each customer, impeding effective segmentation and personalized service delivery.

2.2.2 Outdated Customer Segmentation Practices

The agency currently employs manual methods for customer segmentation. These practices can be inefficient, prone to errors, and may not fully utilize available data for comprehensive customer understanding.

2.2.3 Data Diversity Challenges

The agency faces challenges in handling diverse data sources, including customer demographics, policy details, and claims histories. Variability in data formats and quality presents obstacles for accurate and unified customer profiling.

2.2.4 Limitations of Manual Processes

Manual processes in customer segmentation lead to time-consuming operations and inhibit scalability. This approach lacks the agility needed to respond quickly to market changes and customer preferences.

2.2.5 Inconsistencies in Data Interpretation

The absence of standardized methodologies results in inconsistencies in data interpretation and classification. This inconsistency can impact decision-making processes and hinder personalized service delivery.

2.2.6 Dependency on Manual Analysis

Current practices heavily rely on manual analysis, which introduces subjectivity and potential biases in customer profiling. Human interpretation limitations may lead to

varying insights and recommendations.

2.2.7 Adaptability to Market Dynamics

The existing approach may struggle to adapt swiftly to evolving market dynamics and regulatory changes, affecting the relevance and accuracy of customer segmentation strategies.

2.2.8 Data Security and Privacy Concerns

Manual handling of sensitive customer data raises concerns about data security and privacy compliance. Ensuring robust security measures becomes crucial to protect customer information.

This study of the existing approach reveals significant challenges in fragmented data sources, outdated customer segmentation practices, data diversity, and the need for enhanced automation and standardized methodologies in insurance customer profiling.

2.3 Business Problem

One of the agencies within Lloyd Assurances, specifically Agency 127 of the Lloyd Agencies Network, faces several pressing issues related to customer data management. These challenges hinder the agency's ability to effectively segment and profile customers. The key business problems include the following:

2.3.1 Fragmented Data Sources

The most significant challenge is the lack of consolidated customer data. Customer information is scattered across various sources such as contracts, bordereaux, insurance letters, claims, and other documents. This fragmentation makes it difficult to obtain a complete and accurate view of each customer, impeding effective segmentation and personalized service delivery.

2.3.2 Data Complexity

The agency encounters significant challenges due to the diversity of customer data sources and the formats in which this data is presented. Customer information can be complex, with various formats, structures, and presentations. This complexity makes it difficult for the agency to efficiently process and analyze these data sets, especially when information is dispersed or poorly organized.

2.3.3 Manual Data Handling

The absence of an automated system for customer segmentation requires extensive manual data handling within the agency. Manual processes increase the risk of errors, duplications, and inconsistencies. These errors can significantly impact the accuracy of customer profiles and compromise the integrity of subsequent analyses and decisions. Manual data handling can also lead to operational inefficiencies and additional costs associated with error correction and data re-entry.

2.3.4 Need for Standardization

Customer data needs to be structured and standardized to facilitate analysis and ensure an efficient approach to segmentation. Data entry errors, variations in data formats, and inconsistent data practices can lead to significant discrepancies.

We can conclude that this agency within Lloyd Assurances faces significant challenges in processing and analyzing customer data. The fragmentation of data sources, complexity of data, manual handling errors, and lack of standardization compromise the reliability and efficiency of customer segmentation processes. To overcome these challenges, it is essential to invest in automated systems and standardization practices to ensure data integrity and enhance decision-making in customer management.

2.4 Business Objectives

With a focus on optimizing customer data management processes, the agency aims to enhance efficiency by streamlining data entry and analysis. The project's core objectives include the following goals:

2.4.1 Consolidation of Customer Data

The project aims to centralize customer data from various sources and insurance documents into a unified system. This will provide a comprehensive view of each customer, improving the accuracy and completeness of customer profiles.

2.4.2 Consistency and Standardization

The objective is to establish standardized formats and structures for customer data, ensuring consistency and facilitating data analysis and comparison. Standardization will help in creating reliable customer segments.

2.4.3 Enhancement of Decision Making

The project seeks to provide timely and accurate customer data that supports decision-making processes, enabling better-informed business strategies and personalized service delivery. Enhanced decision-making will contribute to improved customer satisfaction and retention.

2.4.4 Adaptability to Market Changes

The objective is to create a flexible and scalable system that can quickly adapt to evolving market dynamics and regulatory changes. This will ensure that the agency remains competitive and responsive to customer needs.

By achieving these objectives, the agency within Lloyd Assurances will be better positioned to effectively segment and profile customers, ultimately enhancing customer satisfaction and driving business growth.

2.5 Data Science Objectives for Insurance Customer Segmentation

In alignment with the business objectives of optimizing customer segmentation in the insurance sector, our data science objectives focus on utilizing machine learning techniques to improve the precision and effectiveness of segmentation. These objectives aim to enhance customer understanding and drive business growth:

2.5.1 Integration and Consolidation of Customer Data

Utilize Python libraries such as Pandas and NumPy to integrate and consolidate customer data from multiple sources, including policyholder records, claims data, and demographic information. This process ensures the creation of a unified customer profile for more accurate segmentation.

2.5.2 Data Quality, Consistency, and Feature Engineering

Leverage Python-based tools for data cleaning, standardization, and transformation to ensure high data quality. Feature engineering techniques will be applied to extract relevant features such as policy duration, claim frequency, and premium history, optimizing the input for machine learning models.

2.5.3 Enhanced Customer Segmentation

Develop and implement machine learning algorithms, including k-means clustering, hierarchical clustering, and decision trees, using Python libraries like Scikit-learn. These models will be used to create precise customer segments based on factors like demographics, policy details, and claim patterns.

2.5.4 Improvement of Decision Making

Integrate model outputs with visualization tools such as Matplotlib and Seaborn to present insights for decision-making. Future potential developments could include the

use of real-time processing frameworks like Apache Kafka and Apache Flink to further enhance decision-making capabilities.

By achieving these data science objectives, the insurance project will leverage machine learning tools to effectively segment and profile customers, driving better business outcomes.

2.6 Relative Metrics

The following metrics are essential for evaluating the effectiveness of our machine learning approach in customer segmentation within the insurance industry. These metrics emphasize the accuracy, efficiency, and scalability of unsupervised learning methods while considering the specific requirements of the project.

2.6.1 Silhouette Score

The silhouette score measures how similar a data point is to its own cluster compared to other clusters. It ranges from -1 to 1, with higher values indicating better-defined clusters. This metric helps assess the quality of customer segmentation, ensuring that the model forms well-separated and distinct groups of customers.

2.6.2 Inertia (Within-cluster Sum of Squares)

Inertia measures the sum of squared distances between each customer and the centroid of the cluster to which it is assigned. Lower inertia values indicate more compact clusters. This metric is useful for evaluating how well the clustering algorithm fits the data and captures the natural structure of the customer base.

2.6.3 Davies-Bouldin Index

The Davies-Bouldin Index evaluates the average similarity ratio of each cluster with its most similar cluster, considering the distance between clusters and the compactness of the clusters themselves. A lower Davies-Bouldin Index indicates better clustering, as

it reflects well-separated and compact clusters. This is crucial for ensuring that customer segments are distinct and meaningful.

2.6.4 Processing Speed

Processing speed measures how efficiently the system can process and segment the customer data. This metric is important for evaluating the performance of the machine learning algorithm, ensuring that segmentation can be performed quickly enough to provide timely insights, particularly as new customer data is incorporated.

2.6.5 Scalability

Scalability measures the ability of the clustering algorithm to handle increasing volumes of customer data without a significant drop in performance. High scalability ensures that the model remains effective as the customer base grows, which is essential for the long-term success of customer segmentation in the insurance sector.

2.7 Work Methodology

The Team Data Science Process (TDSP) is an agile, iterative data science methodology to deliver predictive analytics solutions and intelligent applications efficiently. TDSP helps improve team collaboration and learning by suggesting how team roles work best together.

TDSP includes best practices and structures from Microsoft and other industry leaders to help toward successful implementation of data science initiatives. The goal is to help companies fully realize the benefits of their analytics program following its four important and crucial steps:

2.7.1 Business Understanding

We start by deeply understanding the challenges related to customer segmentation and profiling within the agency. This includes identifying issues with data availability,

manual processes, and inconsistencies. We set clear project objectives in collaboration with stakeholders, ensuring our solutions align with the agency's strategic goals and enhance customer understanding and service delivery.

2.7.2 Data Acquisition and Understanding

We acquire diverse customer data from various sources such as contracts, bordereaux, insurance letters, and claims. A thorough analysis of these data sources helps us understand their structure, format, and the challenges they present. This analysis informs our data preprocessing and exploratory data analysis efforts, enabling us to make informed decisions in the modeling phase.

2.7.3 Modeling

Using insights from data exploration, we develop robust predictive models for customer segmentation and profiling. We experiment with various machine learning algorithms and statistical techniques to ensure high accuracy and efficiency in real-world scenarios. The models are designed to create precise customer segments that can be used to tailor marketing strategies and improve service delivery.

2.7.4 Deployment

In this phase, we seamlessly integrate our automated customer segmentation solution into the agency's existing workflows.

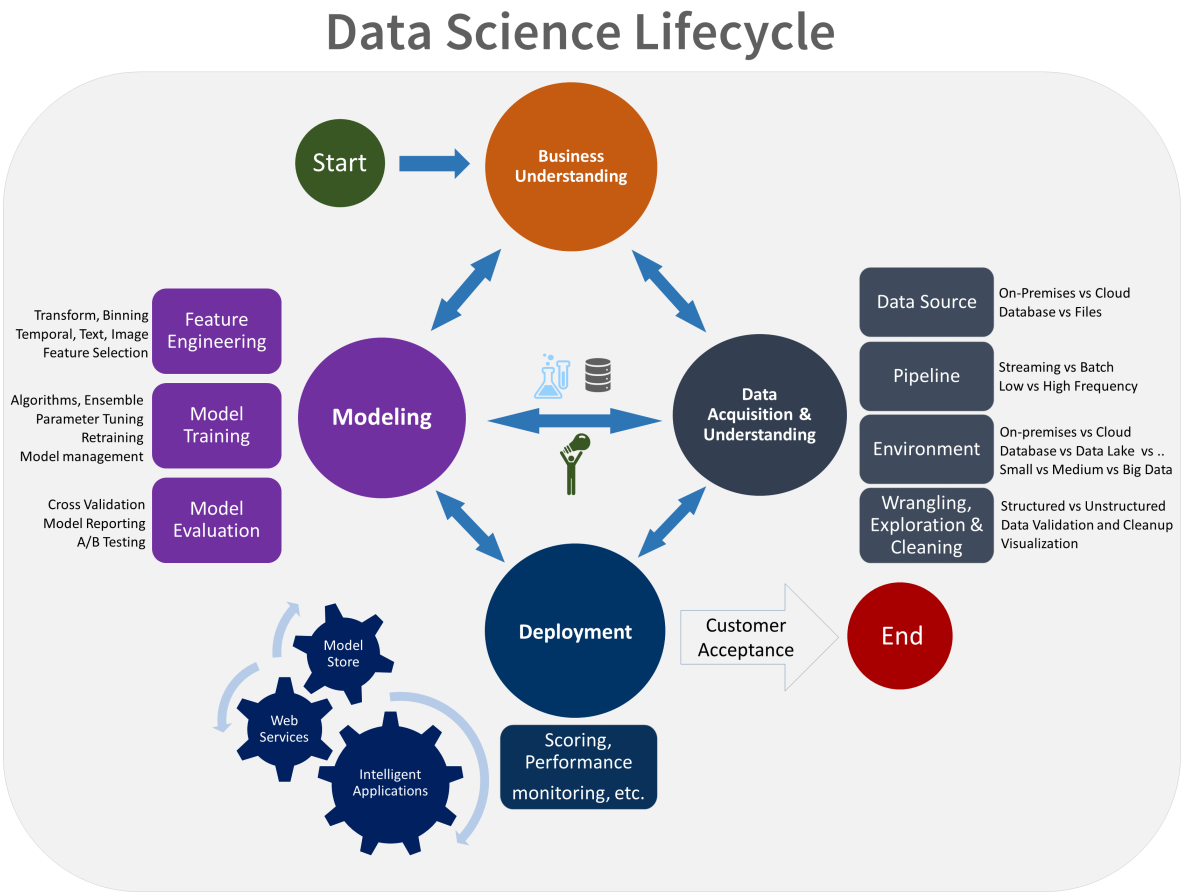

Figure 1: TDSP Work Methodology

In fact, the figure above represents all the important and crucial phases of the TDSP Methodology that we will be following through our project.

This method is aimed at driving operational efficiency, enhancing decision-making, and improving overall customer satisfaction.

2.8 Conclusion

The comprehensive study of the existing customer segmentation practices at the Lloyd Assurances agency revealed significant challenges, including the absence of centralized customer data, reliance on manual processes, and inconsistencies in data handling.

These challenges needed the formulation of clear business objectives, emphasizing the need for improved efficiency, automation, consistency, and enhanced decision-making capabilities.

Translating these business objectives into data science objectives aims to effectively address the identified challenges. The proposed metrics highlight the importance of a balanced approach in achieving the project's goals.

In essence, this chapter underscores the imperative for the Lloyd Assurances agency to adopt automated systems and data science methodologies to overcome existing challenges and enhance their customer segmentation and profiling capabilities, ultimately improving service delivery and customer satisfaction.

CHAPTER 3

Data Acquisition and Understanding

3.1 Introduction

In any machine learning project, the foundation for successful analysis and model development lies in the careful acquisition and understanding of data. This chapter will guide through the critical stages of identifying, collecting, and thoroughly comprehending the insurance customer data that forms the basis of our segmentation project.

We will explore the methodologies used to gather relevant customer data, such as policyholder information, claims history, and demographic details, addressing the challenges posed by data from diverse sources, varying formats, and complexity.

Our focus will be on the techniques and strategies employed to ensure the data's quality, relevance, and integrity, setting the stage for efficient preprocessing and effective machine learning analysis. This understanding is essential for building robust customer segmentation models tailored to the specific needs of the insurance domain.

3.2 Data Acquisition

In our project to develop an optimized customer segmentation model for the insurance sector, the first critical step is acquiring relevant data from the agency's internal systems and specific applications used by Lloyd's Assurances. This section outlines the strategic methodologies employed to gather data from these key platforms, focusing on life insurance contracts, while addressing challenges related to data integration and completeness.

Our primary objective is to consolidate data from various sources into a single, comprehensive file that contains all the necessary client information. This unified dataset will serve as the foundation for building a robust customer portfolio, enabling the agency to effectively segment its clients and tailor its services accordingly.

3.2.1 Data Sources

For the purpose of this project, we will rely on three primary internal applications used by Lloyd's Assurances to collect and manage client data. These platforms—InsurancePass, KYC Lloyd Assurances, and Lloyd AssuranceVie—each serve different functions within the company's data ecosystem. Given our focus on life insurance contracts, we will prioritize data collection from KYC Lloyd Assurances and Lloyd AssuranceVie, integrating them to form a comprehensive customer profile for life insurance clients.

3.2.1.1 InsurancePass

InsurancePass is a platform developed by an external company to centralize vehicle-related data, such as vehicle policies, claims, and customer profiles for vehicle insurance. While this system holds valuable information for the agency's broader client base, its relevance to life insurance is limited. As a result, the data from InsurancePass will not be a primary focus for this project, though it may serve as a supplementary source for future initiatives.

3.2.1.2 KYC Lloyd Assurances

KYC (Know Your Customer) Lloyd Assurances is a critical application designed to compile detailed customer data and profiles, ensuring compliance with regulatory requirements and enhancing the agency's understanding of its clients. This platform gathers a wide array of client information, including demographic details, financial status, risk profiles, and historical interaction with the agency.

For this project, we will leverage KYC Lloyd Assurances as a key source of client data, focusing on the information necessary for life insurance segmentation. Our data acquisition strategy will involve extracting client demographics, policy ownership history, and any relevant customer interactions to form a clear and comprehensive profile.

3.2.1.3 Lloyd AssuranceVie

Lloyd AssuranceVie is the dedicated platform for managing life insurance contracts, covering policy data, personal client details, and contract-specific information. This application stores critical data related to life insurance policies, including contract terms, policyholders' personal data, beneficiaries, claims history, and payment details.

Given that our project focuses on life insurance, Lloyd AssuranceVie will be the primary source of data. Our strategy will involve extracting relevant data from this platform to build a detailed portfolio of life insurance clients. We will integrate this information with data from KYC Lloyd Assurances to ensure a full picture of each customer's profile, policy history, and insurance-related activities.

3.2.2 Data Consolidation Strategy

The ultimate goal of our data acquisition process is to consolidate data from KYC Lloyd Assurances and Lloyd AssuranceVie into a unified file containing all the necessary information to segment the agency's life insurance client portfolio. This involves aligning the datasets from both platforms, ensuring consistency in customer profiles, and addressing potential data discrepancies.

By combining customer demographic details, life insurance policies, and associated personal data into a single comprehensive dataset, we will establish a robust foundation

for the next steps of our project—data preprocessing and machine learning model development. This consolidated data will enable the agency to better understand its clients and optimize customer segmentation and profiling efforts, ultimately improving service delivery and business outcomes.

3.3 Data Extraction

In our efforts to compile relevant data for customer segmentation within the insurance sector, we employed a streamlined approach to extract and consolidate data from the two key platforms, KYC Lloyd Assurances and Lloyd AssuranceVie.

Utilizing Excel reporting tools, we successfully extracted the necessary client and policy data from both platforms, transforming it into structured reports. Through the use of Excel's advanced functionalities, we managed to merge these datasets into a single, unified table that contains all the essential technical and personal columns for our analysis.

The consolidated table includes key technical columns such as:

- Policy Number
- Effective Date
- Expiry Date
- Case Status
- Insured Capital
- Premium
- Premium Frequency
- Mathematical Provision
- Product Type
- Payments Received

Additionally, we incorporated important personal information related to the insured individuals, such as:

- Policyholder
- Insured Person
- Insured Person's Date of Birth
- Insured Person's Title
- Insured Person's Address
- Insured Person's City
- Insured Person's Occupation

By merging these datasets into one comprehensive file, we now have a complete view of the agency's life insurance client portfolio. This consolidated data serves as a strong foundation for further analysis and the development of our customer segmentation models, ensuring that we capture both the technical aspects of each policy and the personal profiles of the insured clients.

This structured dataset will enable us to conduct a more in-depth analysis of client segments, helping the agency tailor its offerings and improve decision-making related to life insurance policies.

3.4 Data Understanding

The data understanding phase is a crucial step in any data science project, as it lays the groundwork for effective analysis and model development. In this project, we employed Python, leveraging libraries such as *Pandas*, *Matplotlib*, *Seaborn*, and *Scikit-learn*, to gain deep insights into the consolidated dataset, which was extracted from the two key platforms: *KYC Lloyd Assurances* and *Lloyd AssuranceVie*.

3.4.1 Step 1: Data Loading and Preliminary Inspection

We began by loading the dataset using *Pandas*, followed by a quick overview of its structure:

- **Previewing the Data**: The `head()` function was used to examine the first few rows of the dataset, providing an initial glimpse into the feature types (numerical and categorical).
- **Shape and Missing Values**: The `shape` attribute gave us the dataset's dimensions, while `isnull().sum()` was used to detect missing values in each column. This step informed our data cleaning strategies.

3.4.2 Step 2: Statistical Summary

A statistical summary of the dataset was generated to understand the distribution and variability of numerical features:

- **Descriptive Statistics**: We used the `describe()` function to calculate key metrics such as the mean, standard deviation, minimum, and maximum values. This summary helped us detect outliers and gain a better understanding of key features like *Premium Amount* and *Insured Capital*.

3.4.3 Step 3: Distribution Visualization of Numerical Features

To visualize the distribution and detect outliers in key features, we employed *Matplotlib* and *Seaborn*:

- **Boxplots**: Boxplots were used to explore variables such as *Premium Amount* and *Insured Capital*. These visualizations provided a clear way to detect outliers and abnormal patterns in the data. The figure below represents the distribution of these features and highlights potential anomalies:

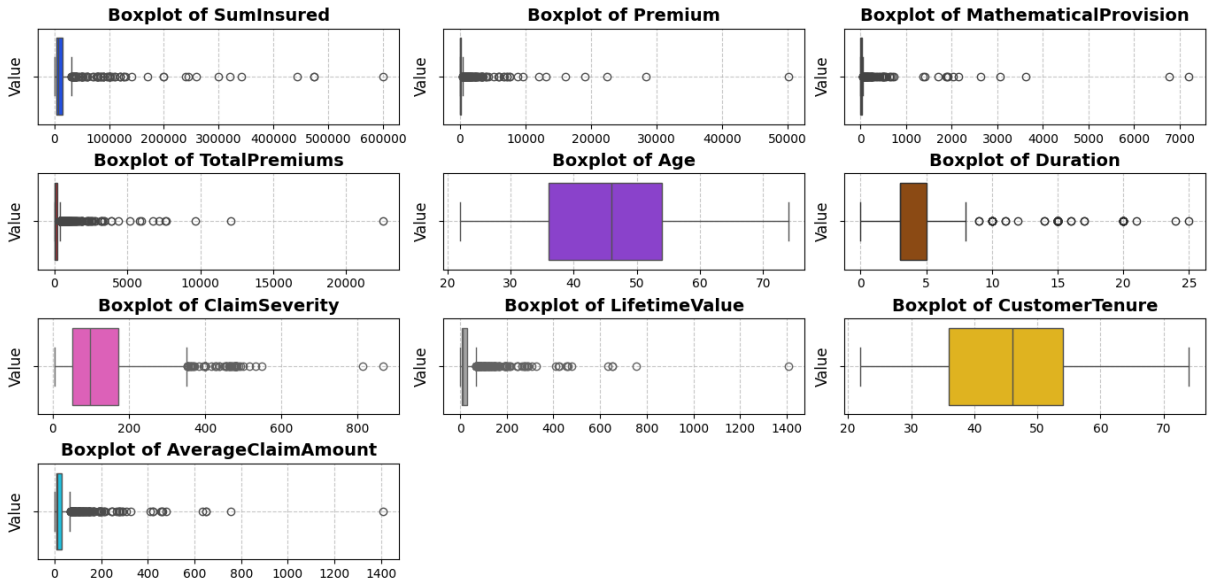

Figure 2: Boxplot highlighting outliers in Premium Amounts and Insured Capital.

3.4.4 Step 4: Correlation Analysis

We calculated the correlation matrix for numerical features to understand the relationships between them:

- **Correlation Matrix**: The correlation matrix was generated using `corr()` to investigate the relationships between features such as *Premium Amount*, *Insured Capital*, and *Age of the Insured*. Strong correlations were identified and visualized using a heatmap, revealing important dependencies between variables.

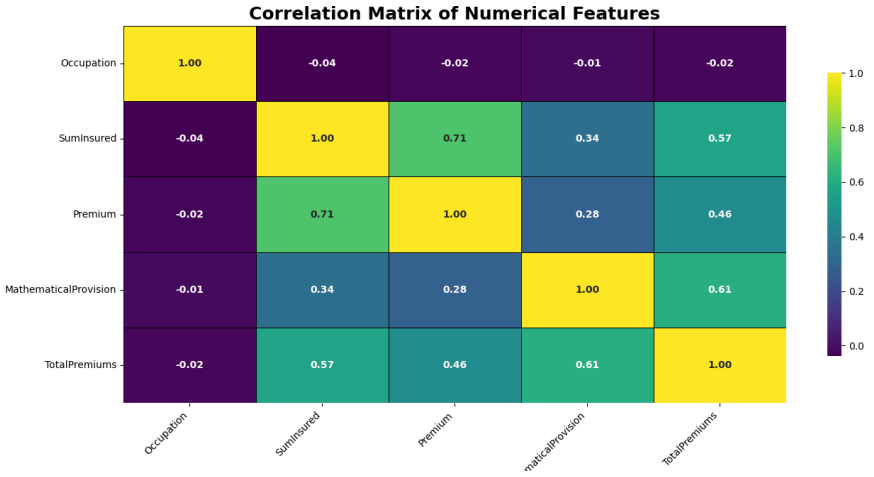

Figure 3: Heatmap of the correlation matrix for numerical features.

3.4.5 Step 5: Numerical Feature Analysis

For numerical variables, we performed frequency analysis to understand the distribution and spread of the data:

- **Frequency Distribution**: Numerical columns such as *Age* and *Claim Severity* were analyzed using boxplots to visualize their distributions, detect any outliers, and check for imbalances or skewness in the data.

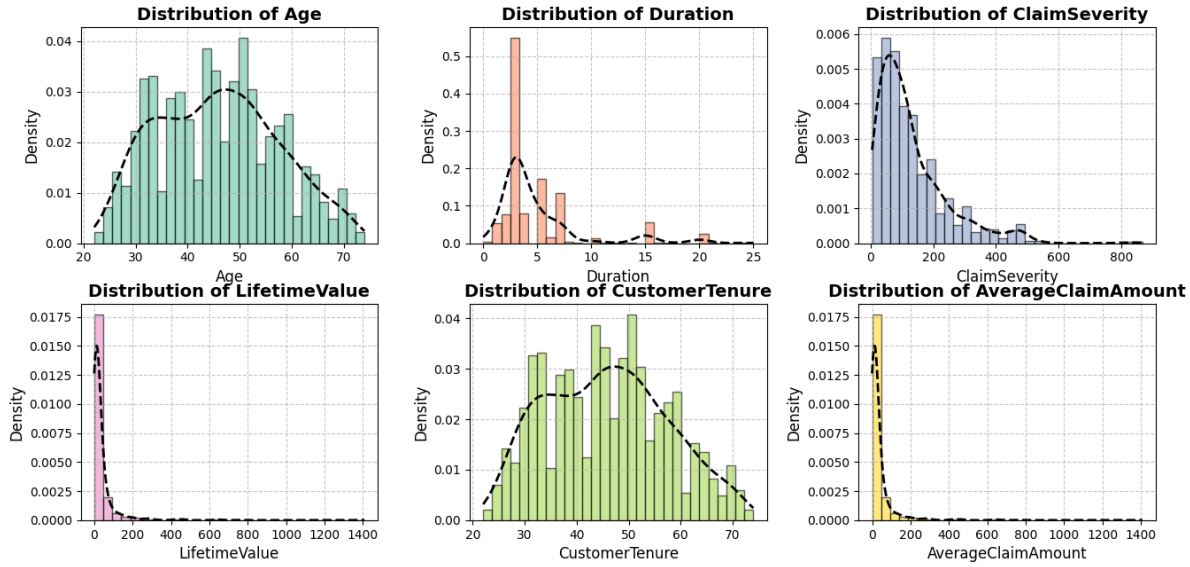

Figure 4: Frequency Distribution Histogramms

Through these steps, we obtained a comprehensive understanding of the numerical data, allowing us to identify key trends, patterns, and potential anomalies. This solid foundation of data understanding will guide the subsequent phases of data preparation, analysis, and modeling.

3.5 Data Preparation

In the Data Preparation phase of our project, we undertook a series of steps to refine and transform the dataset, ensuring its suitability for robust analysis and modeling. This phase was critical to maintaining data quality and enhancing the predictive power of our machine learning models. Below are the key techniques and transformations applied:

3.5.1 Data Extraction and Filtering

Initially, we extracted and filtered the dataset from relevant sources, focusing on retaining only the most pertinent columns and rows for analysis. Duplicate entries were identified and removed to ensure data integrity.

3.5.2 Outlier Detection and Removal

Outliers were identified using the Z-score method, where data points that fell beyond a threshold of 3 standard deviations from the mean were flagged as potential outliers. This was particularly useful for features such as *Premium Amount* and *Sum Insured*, which showed high variability. These outliers were carefully removed to prevent skewing the analysis and model training.

3.5.3 Missing Value Imputation

To handle missing data, we applied mean imputation for numerical features, ensuring the dataset remained complete without introducing bias. Missing categorical values were handled through mode imputation to retain the most common category in each feature.

- **Mean Imputation**: For numerical features such as *Premium Amount* and *Insured Capital*, missing values were replaced with the mean value of the respective column.
- **Mode Imputation**: For categorical features such as *Occupation* and *Policy Type*, missing values were filled using the most frequent category.

3.5.4 Feature Engineering

We engineered several new features to enrich the dataset and provide more predictive power for our models. These include:

- **Age**: Calculated as the difference between the `EffectiveDate` and `DateOfBirth`, converted to years. This feature helps in understanding the age distribution of policyholders and its impact on insurance needs and risk profiles.

- **Duration**: Derived from the difference between the `ExpirationDate` and `EffectiveDate`, also converted to years. It provides insight into the length of time a policy has been active.
- **Gender**: Created by applying a lambda function to the `Title` column, classifying policyholders as 'Male' or 'Female'. This feature helps in demographic analysis.
- **AverageClaimAmount**: Computed by dividing `TotalPremiums` by `Duration` (plus one to avoid division by zero), giving insights into claim amounts per year.
- **ClaimSeverity**: Defined as the ratio of `SumInsured` to `Premium`, helping assess the severity of claims.
- **LifetimeValue**: Calculated as the ratio of `TotalPremiums` to `Duration`, aiding in customer valuation.

3.5.5 Categorical Variable Encoding

For categorical variables, we used `LabelEncoder` to convert them into numerical format suitable for machine learning models. This was particularly effective for ordered categorical variables like *Occupation* and for transforming all other categorical features.

3.5.6 Data Standardization

Numerical features were standardized to ensure uniformity across the dataset and improve the performance of machine learning models. Standardization transforms features to have a mean of 0 and a standard deviation of 1, which is crucial for algorithms sensitive to feature scales such as k-means clustering or gradient boosting.

- **StandardScaler**: Applied to features such as *Premium Amount*, *Insured Capital*, and *Claim Severity* to ensure they have a consistent scale, improving the accuracy and convergence speed of our models.

After applying all the data preparation techniques, the dataset was now free from outliers, missing values were imputed, and both numerical and categorical features were

transformed as needed. These steps laid the foundation for more accurate analysis and model building in the next phases of the project.

3.6 Tools Used for Data Handling

- **Pandas** is a powerful Python library for data manipulation and analysis. It provides data structures like DataFrames that are ideal for handling and analyzing structured data. With Pandas, we can clean, transform, and perform statistical analysis on large datasets efficiently.

- **NumPy** is a fundamental package for numerical computing in Python. It offers support for large, multi-dimensional arrays and matrices, along with a collection of mathematical functions to operate on these arrays. NumPy is essential for performing numerical calculations and handling large-scale data operations.

- **Matplotlib** is a widely-used Python library for creating static, animated, and interactive visualizations. It provides a variety of plotting functions that enable us to visualize data trends, distributions, and relationships effectively. Matplotlib is crucial for presenting data insights through graphs and charts.

- **Seaborn** is built on top of Matplotlib and provides a high-level interface for drawing attractive and informative statistical graphics. It simplifies the creation of complex visualizations such as heatmaps, pair plots, and categorical plots, making it easier to explore and interpret data.

- **Scikit-learn** is a comprehensive library for machine learning in Python. It includes tools for data preprocessing, feature engineering, and model building. Scikit-learn's algorithms for clustering, classification, and regression will be utilized for analyzing client segments and predicting insurance-related outcomes.

- **TensorFlow** and **PyTorch** are leading frameworks for developing and training machine learning and deep learning models. TensorFlow is known for its flexibility and extensive support for deep learning applications, while PyTorch offers a dynamic computational graph and ease of use. Both frameworks will be considered

for advanced modeling and predictive analytics.

3.7 Difficulties and Challenges Encountered

This section outlines the primary challenges faced during the data understanding and preparation phase of our project. Despite employing advanced tools and methodologies for data extraction, several obstacles impacted the accuracy, reliability, and efficiency of our efforts.

The challenges included limitations due to the relatively small dataset size, constraints imposed by the specific nature of the data, and the complexities of data integration and quality assurance. Addressing these issues was crucial for maintaining the integrity and effectiveness of the data preparation process.

1. **Limited Dataset Size:** The portfolio post-data gathering contained only around a thousand contracts. As we are working on an agency level, this relatively small dataset posed limitations in terms of its representativeness and statistical power. To address this, we considered data augmentation techniques to enrich and expand the dataset if necessary.

2. **Diverse Data Formats and Inconsistencies:** The financial documents extracted from various platforms exhibited diverse formats and inconsistencies. This diversity complicated the extraction and integration processes, leading to challenges in maintaining uniformity and accuracy across the dataset.

3. **Complex Data Integration:** Integrating data from multiple sources, such as KYC Lloyd Assurances and Lloyd AssuranceVie, presented challenges in aligning and consolidating information. Variations in data structures and formats required careful handling to create a cohesive and comprehensive dataset.

4. **Quality Assurance and Validation:** Ensuring the accuracy and reliability of the extracted data was hindered by the absence of robust quality assurance mechanisms. The manual verification processes used were time-consuming and susceptible to human error, emphasizing the need for more effective automated validation approaches.

These challenges highlight the need for a thoughtful approach to data handling, encompassing effective data augmentation strategies, careful integration, and rigorous quality assurance processes to enhance the overall quality and usability of the dataset.

3.8 Conclusion

In this chapter, we have navigated through the critical phases of data acquisition, preparation, and analysis, integral to the development of our insurance clustering solution. We systematically identified, collected, and understood the insurance data, tackling challenges related to data diversity, volume, and complexity.

Our journey began with strategic data acquisition, where we meticulously gathered insurance data from various sources. Through robust data extraction techniques and consolidation, we assembled a comprehensive dataset, essential for effective clustering and analysis.

In the data preparation phase, we applied feature engineering techniques to enhance our dataset. By creating new columns that fit the technicality of the business, we enriched the dataset with valuable features. These enhancements facilitated more nuanced analysis and improved the accuracy of our clustering algorithms.

Transitioning to the analysis phase, we leveraged statistical analysis and Python's powerful visualization libraries to gain insights into the insurance data. Through exploratory data analysis (EDA), we identified key patterns and trends within the dataset. Visualization tools enabled us to present these findings effectively, aiding in the development of our clustering models.

Ultimately, this chapter has established a solid foundation for our insurance clustering solution. By combining meticulous data preparation with insightful statistical analysis and visualization, we have paved the way for a sophisticated approach to clustering insurance data. This comprehensive approach ensures that our clustering models are well-informed and capable of driving strategic insights for enhanced decision-making in the insurance sector.

CHAPTER 4

Modeling

4.1 Introduction

In this phase, we transition from data acquisition to the development of predictive models aimed at enhancing our understanding of client dynamics within Lloyd Assurances. Utilizing our dataset, we will employ advanced machine learning algorithms and statistical techniques to extract actionable insights that inform effective client segmentation strategies. Our primary focus is to identify and analyze distinct customer groups, enabling Lloyd Assurances to tailor its offerings and optimize engagement with diverse client segments.

This chapter delves into the intricacies of the Modeling phase, where innovation converges with data science. By uncovering hidden patterns and trends in client behavior, we empower Lloyd Assurances to make informed, data-driven decisions that foster improved efficiency, profitability, and sustainable growth in the insurance market.

4.2 Dimensionality Reduction

In our analysis, we implemented Principal Component Analysis (PCA) as a crucial step in the dimensionality reduction process. The primary motivation for using PCA is to simplify the dataset while retaining as much variance as possible. High-dimensional datasets can pose challenges for clustering algorithms, such as increased computational complexity and difficulty in visualizing data distributions. By reducing the number of features, we aim to mitigate these issues, making our data more manageable and interpretable.

To perform PCA, we first encoded the categorical variables in our dataset using Label Encoding. This step transformed categorical features, such as 'Gender,' 'City,' and 'Product,' into numerical formats suitable for analysis. Next, we standardized the numerical features using the StandardScaler, ensuring that all variables contributed equally to the PCA transformation. The numerical features included 'SumInsured,' 'Premium,' 'MathematicalProvision,' 'TotalPremiums,' 'Age,' 'Duration,' 'ClaimSeverity,' 'LifetimeValue,' 'CustomerTenure,' and 'AverageClaimAmount.'

Subsequently, we applied PCA to the standardized dataset, specifying three principal components for analysis. This allowed us to capture the most significant variance in the data while reducing its dimensionality. The transformed dataset now includes three principal components (PC1, PC2, and PC3), which represent the original features in a condensed form.

The results of the PCA transformation are illustrated in Figure 5, showcasing how the data points are distributed in the new feature space defined by the principal components.

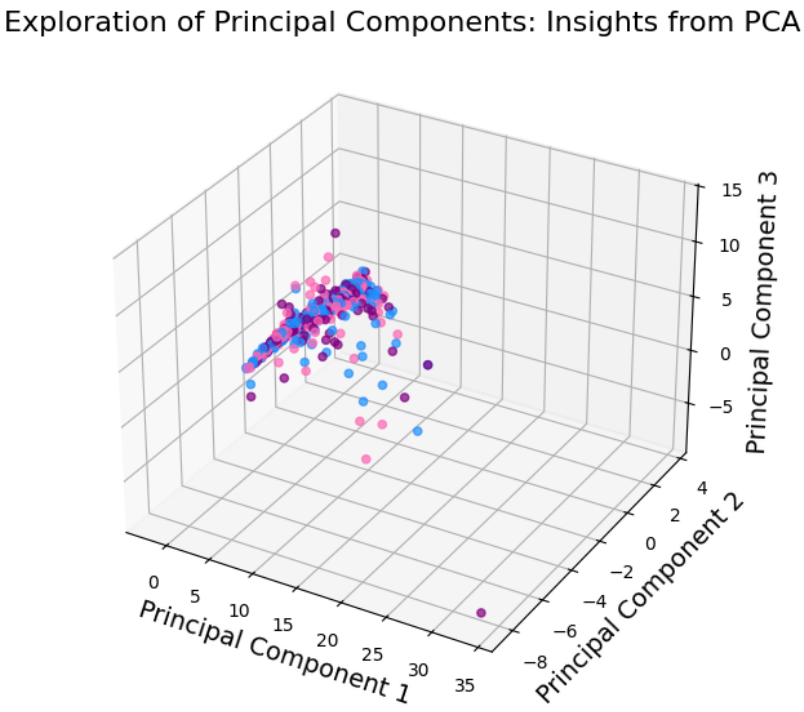

Figure 5: 3D Visualization of Data Distribution in Principal Component Space

The 3D PCA plot visualizes the distribution of data projected onto the first three principal components, with Principal Component 1 capturing the largest variance.

The majority of the data is tightly clustered, indicating shared characteristics, while some outliers are present. The distinct colors suggest potential groupings or classes within the dataset, showing how these groups vary across the principal components.

The significant spread along PC1 suggests it explains most of the variance, while PC2 and PC3 capture less variance, providing insights into the structure of the data in reduced dimensions.

4.3 Model Selection and Building

In the Model Selection phase of our project, we embark on a strategic journey to identify and implement unsupervised learning models that align with our key business objectives for client segmentation at Lloyd Assurances. Our approach involves systematically evaluating various clustering algorithms, each designed to uncover distinct client groups based

on their characteristics and behaviors.

Our goal is to identify the most effective clustering model for segmenting insurance clients, enabling us to derive actionable insights that inform targeted marketing strategies and improve customer engagement.

This collaborative effort emphasizes our commitment to delivering value and driving innovation, positioning the agencies of Lloyd Assurances for enhanced efficiency, profitability, and strategic growth in a competitive insurance landscape.

4.3.1 K-Means Clustering

For our data segmentation objective, we have implemented the K-Means clustering algorithm. This method is particularly effective for partitioning our insurance dataset into distinct groups based on feature similarity.

K-Means clustering offers the advantage of simplicity and efficiency, allowing us to analyze and interpret data patterns effectively. To determine the optimal number of clusters, k, we utilized the Elbow method, which assesses the within-cluster sum of squares for various k values. Following this analysis, we identified $k = 3$ as the optimal choice, leading to meaningful segmentation of our dataset.

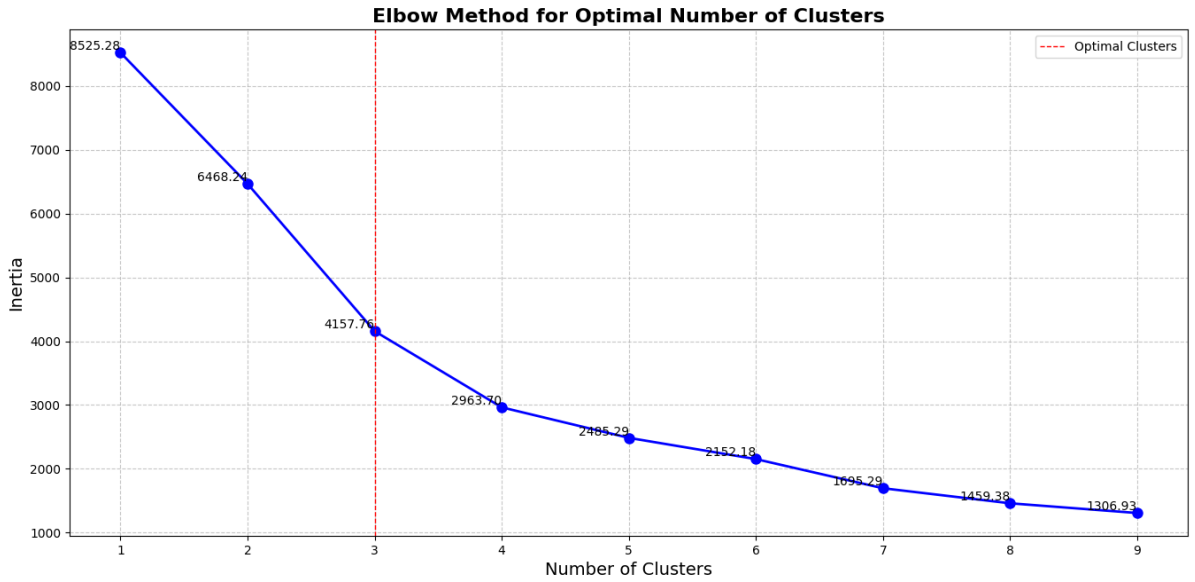

Figure 6: Visualization of the Elbow Method

By applying the K-Means algorithm, we iteratively assigned data points to the nearest cluster centroid until convergence was achieved. This approach enabled us to uncover underlying trends within the data, providing valuable insights that can inform our decision-making process.

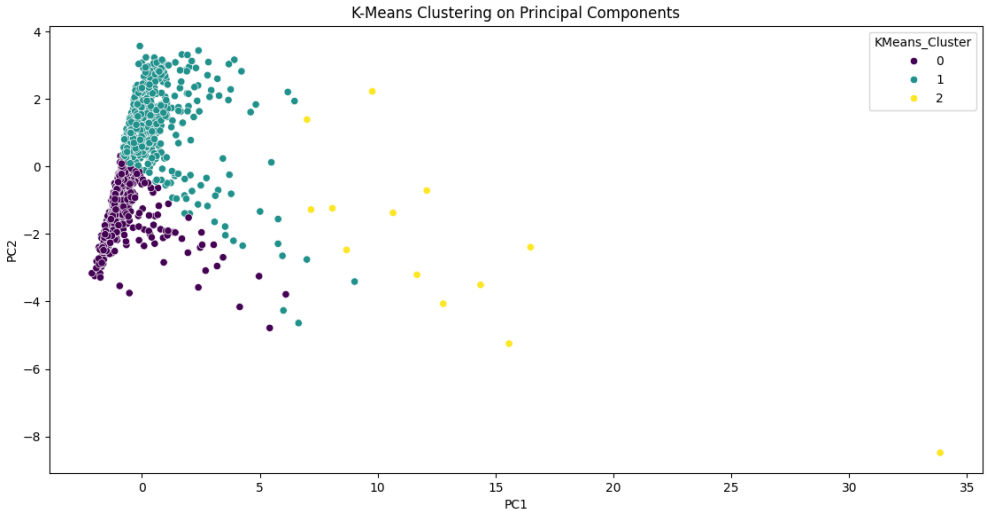

Figure 7: K-Means Clustering Results with $k = 3$

Our strategic implementation of K-Means clustering underscores our commitment to leveraging data-driven methodologies for enhanced operational effectiveness.

4.3.2 DBSCAN Clustering

For our analysis of the insurance dataset, we implemented the DBSCAN (Density-Based Spatial Clustering of Applications with Noise) algorithm. This model is particularly advantageous for identifying clusters of varying shapes and sizes while effectively handling noise and outliers within the data.

DBSCAN operates by grouping together points that are closely packed while marking as outliers points that lie alone in low-density regions. This approach allows us to discover clusters without requiring a predefined number of clusters, offering flexibility in our segmentation strategy.

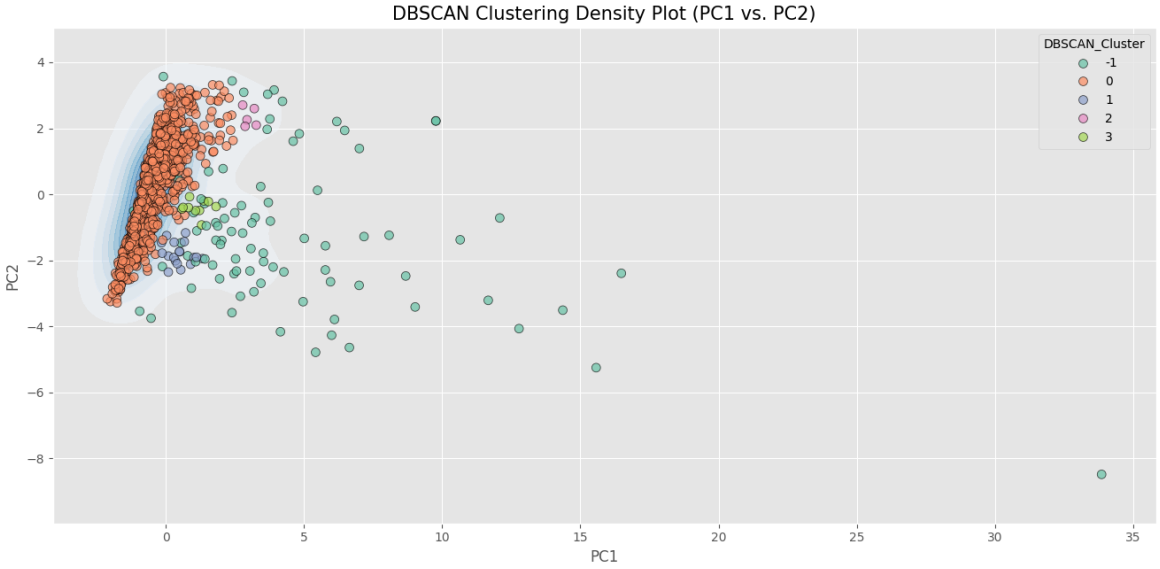

Figure 8: DBSCAN Clustering Results Visualization

The graph represents a DBSCAN clustering result, projected onto the first two principal components (PC1 vs. PC2).

With this figure, we see that DBSCAN effectively separates denser regions from sparser ones, identifying both compact clusters and scattered noise points, which suggests that the data contains varying density regions.

4.3.3 Hierarchical Clustering

To complement our clustering analysis, we also employed Hierarchical Clustering, a technique that builds a hierarchy of clusters either through agglomerative (bottom-up) or divisive (top-down) approaches. This model is particularly useful for revealing the nested structure of our insurance data and providing a more nuanced understanding of the relationships between different data points.

In our implementation, we utilized the agglomerative approach, starting with each data point as an individual cluster and progressively merging them based on their similarity. We evaluated different linkage criteria, including single-linkage, complete-linkage, and average-linkage methods, to determine the most effective way to measure the distance between clusters. The results were visualized using a dendrogram, allowing us to identify the optimal number of clusters based on the hierarchical relationships observed.

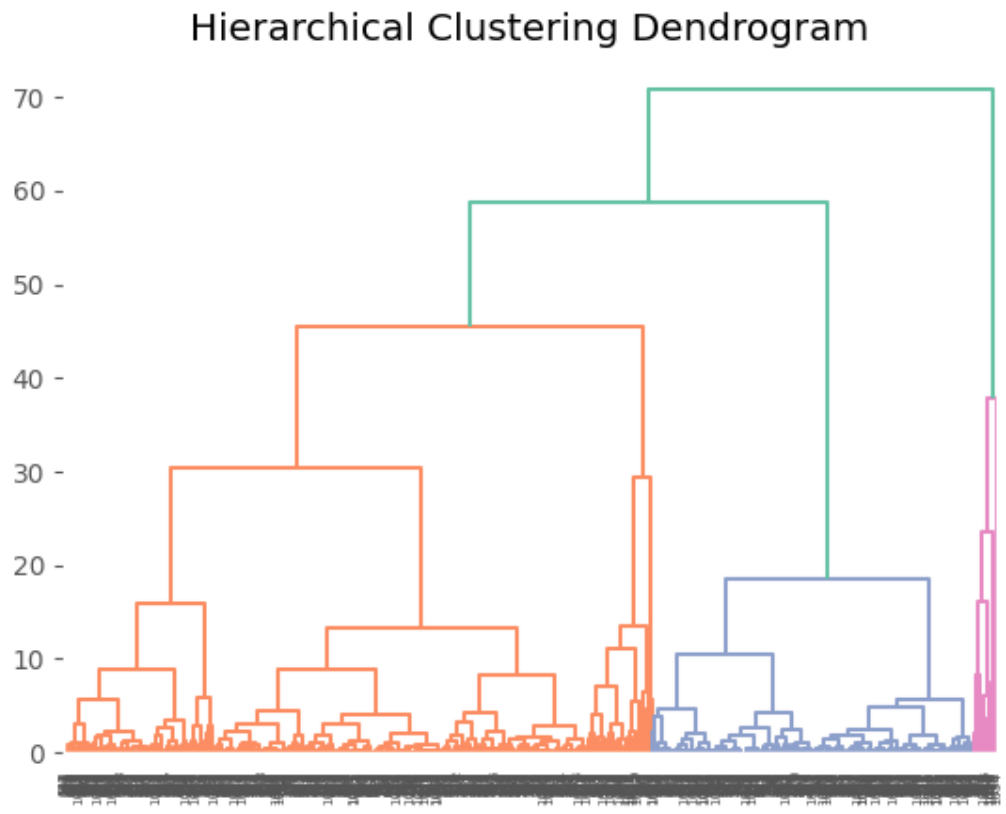

Figure 9: Hierarchical Clustering Results Visualization

Upon examining the dendrogram, we observed that three main clusters emerge at a height of approximately 60 to 70 on the vertical axis, where the largest vertical gaps occur. This suggests that a natural grouping of the data into three clusters is appropriate. Substructures within each main cluster indicate finer divisions, which can be useful if additional granularity is desired.

In summary, the dendrogram analysis indicates that our dataset can be segmented into three primary clusters. This grouping aligns with our objective to identify distinct data segments, and the hierarchical approach provides flexibility for further sub-clustering if required.

4.4 Model Performance Evaluation

In order to segment our client data effectively, we applied three clustering algorithms: K-Means, DBSCAN, and Hierarchical Clustering. Each model was evaluated using three metrics: Silhouette Score, Davies-Bouldin Index, and Inertia (Sum of Squared Distances to Centroid). These metrics help us to assess the clustering quality and identify the most suitable method for our client segmentation objectives.

4.4.1 Evaluation Metrics

- **Silhouette Score**: This metric measures how similar an object is to its own cluster (cohesion) compared to other clusters (separation). The Silhouette Score ranges from -1 to 1, with higher values indicating more well-defined and separated clusters.
- **Davies-Bouldin Index**: This index represents the average similarity ratio between each cluster and the cluster that is most similar to it. Lower values indicate better clustering, as clusters are more distinct from each other.
- **Inertia (Sum of Squared Distances to Centroid)**: Although traditionally associated with K-Means, we calculate inertia for each model as the sum of squared distances between each point and its assigned cluster centroid. Lower values suggest compact, cohesive clusters.

4.4.2 K-Means Clustering Results

K-Means clustering produced the following results:

- **Silhouette Score**: 0.4190
- **Davies-Bouldin Index**: 0.8068
- **Inertia**: 4157.7556

K-Means achieved the highest Silhouette Score among the three methods, indicating better-defined clusters. Additionally, it yielded a low Davies-Bouldin Index, signifying a high degree of separation between clusters. The inertia value is also within an acceptable range, suggesting that data points are generally close to their respective cluster centroids.

4.4.3 DBSCAN Clustering Results

The DBSCAN algorithm, being a density-based clustering method, produced the following results:

- **Silhouette Score**: 0.1982
- **Davies-Bouldin Index**: 0.7049
- **Inertia**: 2723.4063

While DBSCAN achieved the lowest Davies-Bouldin Index, indicating high inter-cluster separation, its Silhouette Score was notably lower than that of K-Means. This suggests that DBSCAN clusters were less cohesive, potentially due to the presence of noise points that were not assigned to any cluster. The low inertia value indicates compact clusters; however, the Silhouette Score implies that the clusters may lack strong cohesion.

4.4.4 Hierarchical Clustering Results

Hierarchical Clustering generated the following metrics:

- **Silhouette Score**: 0.3988
- **Davies-Bouldin Index**: 0.7973
- **Inertia**: 4288.4127

While Hierarchical Clustering performed moderately well in terms of the Davies-Bouldin Index, its Silhouette Score was slightly lower than that of K-Means. The inertia value was also relatively high, suggesting that clusters may not be as compact or cohesive as desired.

4.4.5 Final Model Selection: K-Means Clustering

Based on the results from the three evaluation metrics, we selected **K-Means clustering** as the optimal model for our client segmentation. K-Means achieved the highest Silhouette Score (0.4190) and a competitive Davies-Bouldin Index (0.8068), which together indicate well-defined and distinct clusters. Additionally, the inertia value for K-Means was sufficiently low, demonstrating that data points are closely grouped around their respective cluster centroids.

The simplicity, interpretability, and computational efficiency of K-Means further enhance its suitability for this segmentation task. Unlike DBSCAN, which is sensitive to noise, or Hierarchical Clustering, which lacks flexibility in adjusting cluster assignments post hoc, K-Means provides a robust framework for consistent client segmentation. These strengths align with our objective to derive actionable insights from the data, guiding informed decision-making for our insurance portfolio.

Therefore, we conclude that **K-Means clustering** is the most appropriate model for this project, balancing clustering quality with interpretability and computational efficiency.

4.5 Predictive Regression Analysis

In addition to client segmentation, we conducted a regression analysis aimed at predicting the premium amount for each policyholder. This analysis leverages the Random Forest Regressor, a robust ensemble learning method, to understand the underlying patterns that influence premium values. The objective of this analysis was to identify key features affecting premium predictions, enabling us to gain a deeper understanding of the factors that drive premium pricing.

4.5.1 Model Setup

The following steps outline the process of setting up and training the Random Forest Regressor:

- **Feature Selection:** We prepared the dataset by excluding non-numeric and categorical features irrelevant to prediction, including `PolicyNumber`, `Insured`, `Gender`, `City`, `Product`, and the cluster labels generated from K-Means, DBSCAN, and Hierarchical Clustering.
- **Target Variable:** We defined `Premium` as the target variable, as this is the value we aim to predict.
- **Train-Test Split:** The dataset was split into training and testing sets with a 70-30 ratio, ensuring that the model can be evaluated on unseen data for accuracy and generalization.
- **Model Initialization and Training:** We initialized the Random Forest Regressor with 100 estimators and trained it on the training dataset. The Random Forest algorithm was chosen due to its robustness and ability to capture complex interactions between features.

4.5.2 Prediction and Feature Importance

Upon training the model, we evaluated its predictions on the test set. To understand the influence of each feature on premium predictions, we computed the feature importances using the built-in functionality of the Random Forest model. This analysis helped identify which factors are most relevant to premium pricing, offering insights that could assist in policy adjustment and strategic decision-making.

4.5.3 SHAP Analysis for Interpretability

To further interpret the model, we conducted SHAP (SHapley Additive exPlanations) analysis, which provides a detailed view of each feature's impact on predictions. SHAP values quantify the contribution of each feature to the predicted outcome for each sample, enabling a comprehensive understanding of model behavior.

The following figure presents the SHAP summary plot, which provides insights into the impact and distribution of each feature's contributions to the premium prediction model.

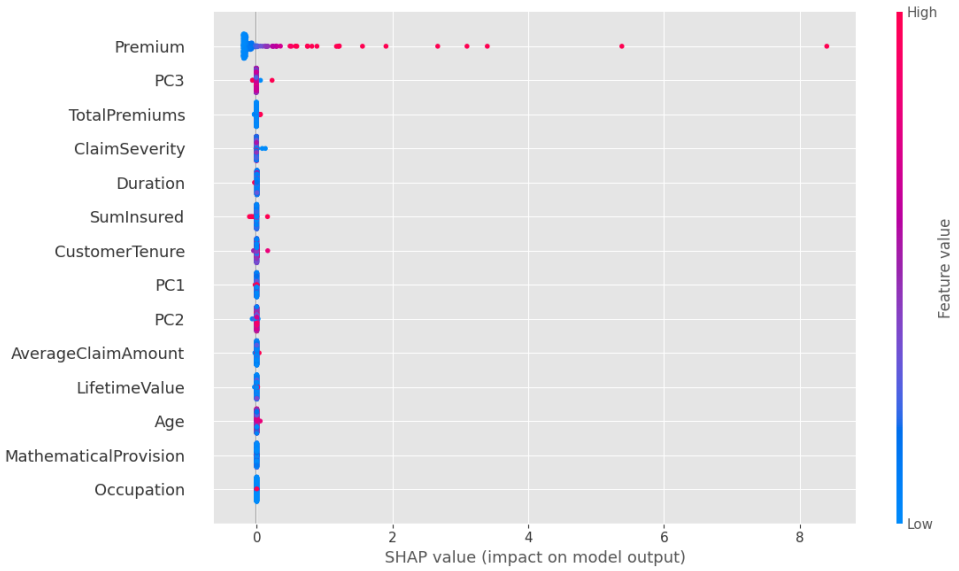

Figure 10: SHAP Summary Plot for Feature Importance in Premium Prediction

- **Premium**: As expected, this feature shows the highest SHAP values, indicating it is a dominant factor in the model's output.
- **PC3 and Total Premiums**: The principal component (PC3) and `TotalPremiums` are also influential, with considerable impact on the premium predictions. These features exhibit higher SHAP values, suggesting they significantly contribute to premium variation.
- **Claim Severity and Duration**: These features have moderate SHAP values, reflecting their roles in premium calculations. They contribute meaningfully to predictions, though with less impact than `Premium` and `PC3`.
- **Remaining Features**: Features such as `SumInsured`, `CustomerTenure`, and the remaining principal components (PC1 and PC2) exhibit lower SHAP values. Although they affect the predictions, their influence is less pronounced.

The color gradient from blue to red indicates feature values from low to high. Features like `Claim Severity` and `PC3` show that higher values tend to increase the premium predictions, as indicated by the concentration of red points on the positive side of the SHAP values.

The Random Forest Regressor, enhanced by SHAP interpretability analysis, provided valuable insights into the factors influencing premium amounts. Features with high importance in the model suggest areas where policy adjustments might be beneficial, as well as attributes that significantly contribute to premium pricing. This predictive model serves as a foundation for data-driven adjustments to premium calculation, contributing to a more optimized and fair pricing strategy.

4.6 Visualization and Understanding of Clusters

In this final phase of the modeling stage, we focus on interpreting the results obtained from the KMeans clustering algorithm. The clusters generated provide insight into the underlying patterns within the dataset. To facilitate understanding, we first add the KMeans cluster labels from the scaled data to the original unscaled dataset, allowing us to analyze the average values of relevant features grouped by these clusters.

We calculated the average values for original numerical features, including Age, Duration, Premium, Sum Insured, Average Claim Amount, Customer Tenure, and Claim Severity, for each KMeans cluster. The results are summarized in the following table, which highlights the distinguishing characteristics of each cluster.

tableau 1: Original Numerical Features Summary for KMeans Clustering

KMeans Cluster	**Age**	**Duration**	**Premium**	**Sum Insured**	**Avg. Claim Amount**	**Customer Tenure**	**Claim Severity**
0	35.60	4.75	155.90	16045.62	13.38	35.60	206.32
1	54.69	4.72	553.85	14187.17	42.51	54.69	63.87
2	53.64	14.07	7516.25	142214.29	530.06	53.64	17.03

Additionally, we computed the average values for categorical features, such as Gender, Occupation, City, and Product, grouped by the KMeans clusters, as summarized in the next table. This analysis enables us to observe trends and associations between these features and the identified clusters.

tableau 2: Scaled Categorical Features Summary for KMeans Clustering

KMeans Cluster	Gender	Occupation	City	Product
0	0.72	0.23	57.97	0.49
1	0.72	0.16	72.41	0.33
2	0.79	0.14	82.71	0.07

To visually represent the average feature values for each KMeans cluster, we created a bar chart that illustrates the mean values of the scaled features across the clusters. This visual representation aids in identifying the key differences between clusters and highlights the varying characteristics of the data points within each group.

Figure 11: Average Scaled Feature Values for KMeans Clusters

Through these analyses and visualizations, we gain a clearer understanding of the clusters produced by KMeans, enabling us to draw actionable insights and inform further decision-making processes.

4.6.1 Cluster Output Analysis

The KMeans clustering algorithm identified three distinct clusters within the dataset, each exhibiting unique characteristics that provide valuable insights into customer profiles. Below, we analyze each cluster in detail:

4.6.1.1 Cluster 0: Younger Customers

Cluster 0 is characterized by a demographic of younger customers, with an average age of approximately 35.6 years. Key characteristics include:

- Average Premium: 155.90
- Average Claim Amount: 13.38
- Claim Severity: 206.32
- Customer Tenure: 35.60 months

This cluster suggests that these customers may have less frequent but potentially less severe claims, indicating a stable customer base with lower associated risk. Their relatively low average premium also points to a demographic that is perhaps just beginning to engage with insurance products, likely valuing affordability over extensive coverage.

4.6.1.2 Cluster 1: Older, Higher-Risk Customers

In contrast, Cluster 1 encompasses older customers, with an average age of around 54.7 years. Their characteristics are as follows:

- Average Premium: 553.85
- Average Claim Amount: 42.51
- Claim Severity: 63.87
- Customer Tenure: 54.69 months

The data indicates that older customers tend to have more complex insurance needs, which may be attributed to increased health-related risks and life stage considerations.

Their higher average premium reflects the necessary adjustments made to accommodate these factors. Additionally, their tenure suggests loyalty, potentially indicating satisfaction with the services provided.

4.6.1.3 Cluster 2: High Premium, Low Claim Severity

Lastly, Cluster 2 is notable for its average age of 53.6 years and extraordinarily high average premium payments, peaking at 7516.25. The characteristics include:

- Average Premium: 7516.25
- Average Claim Amount: 530.06
- Claim Severity: 17.03
- Customer Tenure: 14.07 months

This cluster indicates that while the premiums are significantly high, claims are less frequent and less severe. This demographic likely seeks high-value coverage for specific risks, possibly reflecting an affluent customer base that prioritizes comprehensive protection over cost considerations.

4.6.2 Summary of Analysis

In summary, the analysis reveals the following insights:

- Cluster 0 represents younger, lower-risk customers, indicating a potential market for entry-level insurance products.
- Cluster 1 consists of older, higher-risk customers with moderate premiums, suggesting a need for tailored solutions that address complex life stages and health considerations.
- Cluster 2 includes clients with exceptionally high premiums and average claim amounts but lower claim severity, highlighting opportunities for offering specialized, high-end insurance products.

These insights can guide targeted marketing strategies and risk assessments tailored to the specific needs and profiles of each cluster.

4.6.3 Visual Representation of Clusters

To better illustrate the distinctions among the three clusters, the following graph presents six representative examples, highlighting differences in average premium, claim amount, claim severity, age, sum insured and customer tenure for each cluster:

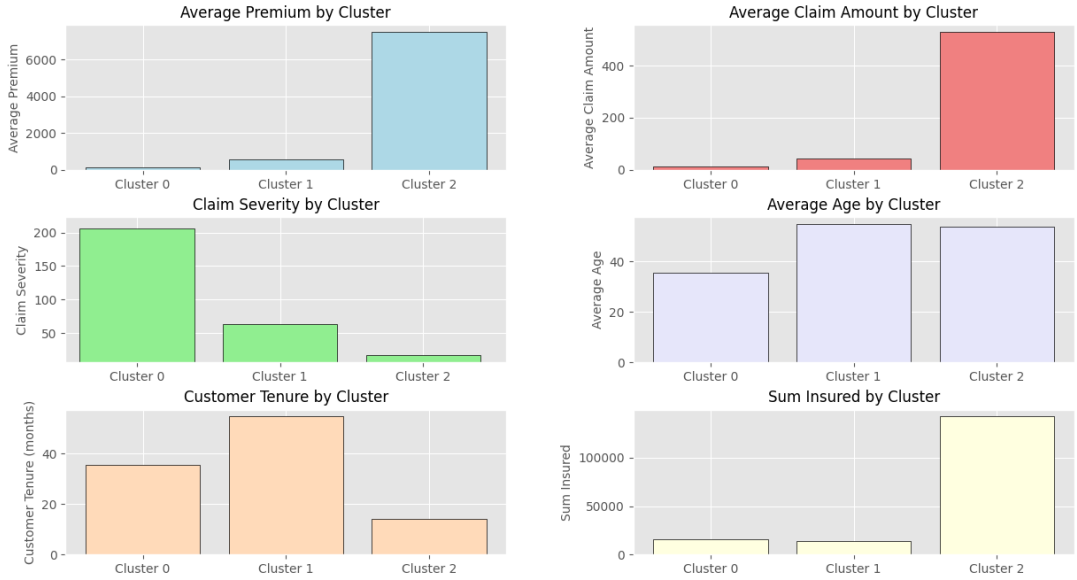

Figure 12: Comparison of Key Features Across Customer Clusters

This visual aid will enhance understanding of the varying customer profiles, supporting further strategic decision-making.

The insights derived from our analysis are crucial for formulating effective business strategies and making informed decisions. By identifying distinct customer segments based on risk profiles, premiums, claim severity, customer age, tenure, and sum insured, we can tailor our approach to maximize customer satisfaction and business profitability. For instance, we can optimize pricing strategies for each cluster, prioritize customer retention efforts for low-risk segments, target specific demographics for acquisition, and develop customized product offerings to cater to diverse needs.

4.7 Conclusion

In this chapter, we conducted a comprehensive analysis for client segmentation and premium prediction within the insurance domain. Our approach encompassed various clustering and regression techniques, allowing us to derive meaningful insights from the dataset.

For client segmentation, we evaluated three clustering algorithms: K-Means, DBSCAN, and Hierarchical Clustering. Our metrics indicated that K-Means effectively captured well-defined and distinct clusters. The analysis revealed three distinct clusters characterized by differing demographics, claim behaviors, and premium amounts, facilitating tailored strategies for marketing and risk management.

In parallel, we implemented a **Random Forest Regressor** for predictive regression analysis, focusing on the identification of key features influencing premium amounts for policyholders. The SHAP analysis provided additional interpretability, allowing us to quantify the contributions of each feature to the model's predictions. This analysis highlighted the dominant role of the premium itself, alongside significant influences from principal components and other features.

The clustering and regression analyses collectively enhanced our understanding of client behavior and premium pricing strategies. The insights gleaned from the K-Means clusters indicated opportunities for entry-level insurance products targeting younger customers, as well as specialized high-end products for affluent clients with complex needs.

In conclusion, the integration of K-Means clustering and Random Forest regression presents a robust framework for client segmentation and premium prediction.

These methodologies not only provide clarity on customer profiles and their behaviors but also empower the organization to make informed, data-driven decisions in optimizing their insurance offerings. Future work may explore further enhancements, such as incorporating additional features or employing advanced machine learning techniques, to refine these models and improve prediction accuracy.

CHAPTER 5

General Conclusion and Perspectives

In concluding this report, we reflect on our comprehensive journey through the intricacies of client segmentation within the insurance sector. This project has underscored the essential role of data-driven approaches in transforming client understanding and enhancing operational efficiencies.

Our work began with meticulous data extraction and collection, where we faced the challenges of diverse data sources and varied formats. Through careful preprocessing and exploration, we ensured the integrity and usability of the data, paving the way for effective analysis. This phase highlighted the importance of thorough data cleansing and feature engineering in fostering meaningful insights.

As we moved forward, we employed various data analysis techniques to understand the underlying patterns and characteristics of our client base. The segmentation analysis utilized K-Means clustering, which revealed distinct client groups based on demographics and behaviors, thereby facilitating targeted marketing strategies and personalized service offerings. The insights derived from these clusters empower the agency to tailor its approach to different client needs, ultimately enhancing customer satisfaction and loyalty.

In parallel, we deployed predictive modeling techniques, specifically utilizing Random Forest regression, to identify key features influencing premium amounts. This allowed

us the potential possibility for the stakeholders to create a robust framework for pricing strategies that align with client profiles and market demands.

Looking ahead, we recognize the potential for continuous improvement and innovation within our framework. The next phase could involve the deployment of decision-making solutions or tools that showcase our findings through real-time data-oriented platforms. Such a platform could include interactive reports, visualizations, and predictive analysis features, offering recommendations tailored for the agency's stakeholders. This advancement would empower decision-makers with immediate access to insights, enabling them to respond swiftly to market dynamics and client needs.

As we advance, we remain committed to exploring new horizons and driving forward the intersection of data science and insurance, ultimately aiming for a future where informed decision-making and enhanced client relationships thrive.

Bibliography

[1] S. Lloyd, "Least Squares Quantization in PCM," *IEEE Transactions on Information Theory*, vol. 28, no. 2, pp. 129–137, 1982. doi: 10.1109/TIT.1982.1056489.

[2] J. A. Hartigan and M. C. Wong, "A K-Means Clustering Algorithm," *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, vol. 28, no. 1, pp. 100–108, 1979. doi: 10.2307/2984875.

[3] A. K. Jain, M. N. Murty, and P. J. Flynn, "Data Clustering: A Review," *ACM Computing Surveys*, vol. 31, no. 3, pp. 264–323, 1999. doi: 10.1145/331499.331504.

[4] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd ed. New York: Springer, 2009.

[5] Scikit-learn, "Scikit-learn Documentation," 2022. [Online]. Available: https://scikit-learn.org/stable/.

[6] Pandas Development Team, "Pandas Documentation," 2022. [Online]. Available: https://pandas.pydata.org/docs/.