

# **UK LIFE INSURANCE PREDICTION**

**Pritam Madan Seth**

**A dissertation submitted to  
The School of Computing Sciences of the University of East Anglia  
in partial fulfilment of the requirements for the degree of  
MASTER OF SCIENCE.  
SEPTEMBER, 2023**

© This dissertation has been supplied on condition that anyone who consults it is understood to recognise that its copyright rests with the author and that no quotation from the dissertation, nor any information derived therefrom, may be published without the author or the supervisor's prior consent.

## SUPERVISOR(S), MARKERS/CHECKER AND ORGANISER

The undersigned hereby certify that the markers have independently marked the dissertation entitled “**UK Life Insurance Prediction**” by **Pritam Madan Seth**, and the external examiner has checked the marking, in accordance with the marking criteria and the requirements for the degree of **Master of Science**.

Supervisor:

---

Prof. Shaun Parsley

Markers:

---

Marker 1: Prof. Shaun Parsley

---

Marker 2: Prof. Katharina Huber

External Examiner:

---

Checker/Moderator

Moderator:

---

Dr. Wenjia Wang

# DISSERTATION INFORMATION AND STATEMENT

Dissertation Submission Date: **September, 2023**

Student: **Pritam Madan Seth**  
Title: **UK Life Insurance Prediction**  
School: **Computing Sciences**  
Course: **Data Science**  
Degree: **MSc.**  
Duration: **2022–2023**  
Organiser: **Dr. Wenjia Wang**

## STATEMENT:

Unless otherwise noted or referenced in the text, the work described in this dissertation is, to the best of my knowledge and belief, my own work. It has not been submitted, either in whole or in part for any degree at this or any other academic or professional institution.

Subject to confidentiality restriction if stated, permission is herewith granted to the University of East Anglia to circulate and to have copied for non-commercial purposes, at its discretion, the above title upon the request of individuals or institutions.

---

Signature of Student

# Abstract

The UK's retail insurance market is marked by its constant evolution, with numerous factors impacting pricing decisions. Accurate prediction of price changes is crucial for insurers to stay competitive and meet customer demands. This dissertation focuses on developing a software solution that captures real-time pricing data and identifies price change drivers.

The core objective is to provide insurers with a tool that enhances price change prediction accuracy and enables data-driven decisions. In an era where data drives insurance, this project is timely, offering vital insights for pricing strategies.

The journey begins with a literature review of machine learning techniques in insurance price change prediction. Real-time pricing data is collected and preprocessed, automating data collection via web scraping. Data quality is ensured by addressing outliers and missing values.

Exploratory data analysis uncovers patterns and relationships. Feature selection, through statistical methods, identifies influential factors. Machine learning models, chosen based on data and objectives, are trained.

Robust model evaluation measures predictive accuracy. The outcome is a user-friendly software solution, empowering insurers to make data-driven decisions.

This dissertation concludes by discussing implications and offering actionable recommendations. This research advances insurance pricing, providing insights and tools in the data-centric insurance landscape.

# Acknowledgements

I would like to express my heartfelt gratitude to my supervisor Prof. Shaun Parsley, for his invaluable mentorship, steadfast support, and insightful guidance throughout the entire process of completing this dissertation. His expertise, encouragement, and patience have played a pivotal role in shaping my ideas and refining my research.

I also extend my deep gratitude to my parents, whose enduring love, encouragement, and sacrifices have served as the bedrock of my academic pursuit. Their unwavering belief in my abilities and unwavering support have been my driving force.

Lastly, I want to acknowledge the backing of the academic community, my friends, and all those who have contributed to my personal and intellectual growth. This dissertation would not have been achievable without the collective efforts of all those who have been part of my academic and personal journey.

Pritam Seth

Norwich, UK.

# Table of Contents

<b>Abstract</b>	<b>iv</b>
<b>Acknowledgements</b>	<b>v</b>
<b>Table of Contents</b>	<b>vi</b>
<b>List of Tables</b>	<b>ix</b>
<b>List of Figures</b>	<b>x</b>
<b>List of Abbreviations</b>	<b>xi</b>
<b>1 Introduction</b>	<b>1</b>
1.1 Background . . . . .	1
1.2 Research Aim . . . . .	2
1.3 Research Objectives . . . . .	3
1.4 Research Questions . . . . .	3
1.5 Rationale of the study . . . . .	4
1.6 Research Significance . . . . .	5
1.7 Structure of the report . . . . .	5
<b>2 Literature Review</b>	<b>7</b>
2.1 Introduction . . . . .	7
2.2 Machine learning practices that are used in life insurance companies . .	8
2.3 Use of predictive modeling in life insurance sector . . . . .	10
2.4 Challenges and importance for collecting real time data in insurance companies . . . . .	12
2.5 Building a Predictive Model Using Machine Learning Algorithms . . . .	14
2.5.1 Data Collection and Preprocessing . . . . .	14
2.5.2 Model Selection . . . . .	14
2.5.3 Data Splitting and Model Training . . . . .	15
2.5.4 Model Evaluation . . . . .	15
2.5.5 Model Deployment . . . . .	16
2.5.6 Interpretability and Compliance . . . . .	16
2.6 Machine Learning's Impact on Price Forecasting in Insurance Markets .	16
2.6.1 Risk Assessment . . . . .	17
2.6.2 Pricing Optimization . . . . .	17
2.7 Methods for Evaluating the Accuracy and Effectiveness of Predictive Models in the Insurance Market . . . . .	17
2.7.1 Data Quality Assessment . . . . .	18

2.7.2	Model Performance Metrics . . . . .	18
2.7.3	Cross-Validation . . . . .	18
2.7.4	Confusion Matrix . . . . .	19
2.7.5	Receiver Operating Characteristic (ROC) Curve . . . . .	19
2.7.6	Profitability Metrics . . . . .	19
2.7.7	Backtesting . . . . .	19
2.7.8	Ethical and Fairness Evaluation . . . . .	20
2.7.9	Continuous Monitoring . . . . .	20
2.8	The Importance of Using Appropriate Performance Metrics and Cross-Validation Techniques in Life Insurance Companies . . . . .	20
2.8.1	Performance Metrics in Life Insurance . . . . .	20
2.8.2	Policyholder Retention Rate . . . . .	21
2.8.3	Claims Ratio . . . . .	21
2.8.4	Profit Margin . . . . .	21
2.8.5	Customer Satisfaction and Net Promoter Score (NPS) . . . . .	21
2.8.6	Underwriting Accuracy . . . . .	21
2.8.7	Cross-Validation Techniques in Life Insurance . . . . .	22
2.8.8	Benefits of Using Performance Metrics and Cross-Validation . . . . .	22
<b>3</b>	<b>Methodology Design</b>	<b>23</b>
3.1	Analysis of the Methods . . . . .	23
3.2	Design of Methodology . . . . .	24
3.3	Evaluation Methods and Measures . . . . .	25
3.3.1	Data Collection Method . . . . .	25
3.3.2	Data Analysis . . . . .	26
3.4	Tools and Resources . . . . .	28
3.5	Summary . . . . .	29
<b>4</b>	<b>RESULTS AND FINDINGS</b>	<b>30</b>
4.1	Introduction . . . . .	30
4.2	Findings in Insurance Company Stock . . . . .	30
4.2.1	Admiral Tests SMA . . . . .	32
4.2.2	Aviva SMA . . . . .	33
4.2.3	Direct Line SMA . . . . .	35
4.2.4	Admiral ES . . . . .	37
4.2.5	Aviva ES . . . . .	39
4.2.6	Direct Line ES . . . . .	40
4.3	Admiral Group . . . . .	42
4.4	Aviva Group . . . . .	43
4.5	Direct line Group . . . . .	43
4.6	Finding in Insurance Premium Prices . . . . .	44
4.6.1	Descriptive Analysis . . . . .	45
4.6.2	MAPE and MAE Score . . . . .	46
4.7	Discussion . . . . .	47
4.8	Chapter Summary . . . . .	51

<b>5</b>	<b>Conclusion</b>	<b>53</b>
5.1	Summary of Findings . . . . .	53
5.2	Discussion . . . . .	55
5.3	Conclusion . . . . .	57
5.4	Suggestion for Further Work . . . . .	59
	<b>References</b>	<b>60</b>
<b>A</b>	<b>Insurance Premium Prediction Program Manual</b>	<b>67</b>
A.1	Introduction . . . . .	67
A.2	System Requirements . . . . .	67
A.3	Installation . . . . .	67
A.4	Running the Program . . . . .	68
A.5	Running the Program . . . . .	68
A.6	Viewing the Results . . . . .	69
A.7	Troubleshooting . . . . .	69
A.8	Appendix . . . . .	69



# List of Tables

4.1	Admiral Group Forecasting Results . . . . .	42
4.2	Aviva Group Forecasting Results . . . . .	43
4.3	Direct Line Group Forecasting Results . . . . .	44

# List of Figures

4.1	Table1 . . . . .	30
4.2	Graph1 . . . . .	32
4.3	Graph2 . . . . .	32
4.4	Graph3 . . . . .	33
4.5	Graph4 . . . . .	33
4.6	Graph5 . . . . .	34
4.7	Graph6 . . . . .	34
4.8	Graph7 . . . . .	35
4.9	Graph8 . . . . .	35
4.10	Graph9 . . . . .	36
4.11	Graph10 . . . . .	36
4.12	Graph11 . . . . .	37
4.13	Graph12 . . . . .	37
4.14	Graph13 . . . . .	38
4.15	Graph14 . . . . .	38
4.16	Graph15 . . . . .	39
4.17	Graph16 . . . . .	39
4.18	Graph17 . . . . .	40
4.19	Graph18 . . . . .	40
4.20	Graph19 . . . . .	41
4.21	Graph20 . . . . .	41
4.22	Demographics . . . . .	44
4.23	Descriptive Statistics . . . . .	46
4.24	MAPE and MAE Score . . . . .	46
A.1	Prompt Screen . . . . .	68
A.2	Entered Details . . . . .	69
A.3	Insurance Premium Price . . . . .	70



# Chapter 1

## Introduction

This chapter covers the insights in accordance with the research theme that is life insurance prediction in United Kingdom. This allocation discusses the background, rationale of the research study along with significance of the research subject. It determines the chief aim and objectives of the UK life insurance prediction. Lastly, it elaborates the structure of the research study by what means it would be drawn.

### 1.1 Background

The UK life insurance market is an intricate and dynamic, ever-varying industry. One of the biggest challenges insurers realizes is predicting price alterations in the market. With so many factors prompting insurance pricing, it can be challenging for sources to make informed decisions without a thorough intelligence of the motives for these variations. Predictive UK life insurance refers to the attachment of data assessment and predictive modelling techniques to evaluate the probability of policyholders making a claim or anticipated life of individuals Nijkamp & Perrels (2018). Researchers collect and consider a variety of statistics, incorporating demographic intelligence, medical history, lifestyle considerations, and even socioeconomic indicators. This data improves insurers better comprehend their policyholders and evaluate their potential exposure Shockey et al. (2018).

Employing advanced statistical models and machine learning algorithms, insurers can measure the likelihood of an insured experiencing proceedings such as illness, accident, or death during the reporting period. This assessment permits them to create appropriate premium rates and coverage situations. Life insurance predictions advance

price accuracy instead of trusting on a variety of risks, insurance companies can offer premiums that are more appropriate for policyholders. Those with lower projected risk may profit from lower premiums, while those with higher hazard may have to pay higher premiums. Predictive modelling empowers the expansion of personalized insurance policies. Insurance companies can generate policies precisely designed for different age groups, occupations, or health conditions. This customization certifies that policyholders receive coverage personalized to their unique requirements Abdelhadi et al. (2020).

Accurate forecasting models are critical to the financial constancy of insurance companies. By considering and assessing risk efficiently, insurers can better accomplish their financial reserves, certifying they have enough funds to meet their responsibilities when policyholders file a claim for compensation. Policyholders also profit from UK life insurance predictions. This can lead to more reasonable insurance options and inspire healthier lifestyles, as individuals can take steps to decrease their risk factors, which in turn can subordinate their premiums. Policymakers and regulators can practice data from life insurance prediction research to evaluate the fairness and transparency of the assurance industry. This could lead to the expansion of regulations that shield consumers and promote fair insurance practices Alcaide (2023). Essentially, UK life insurance prediction plays a fundamental role in enhancing insurance operations by leveraging data-driven insights. It aids both insurers and policyholders by enlightening risk assessment, pricing accurateness and policy customization while causative to the financial stability of insurers. It also has inferences for regulatory oversight and inspiring consumers to make healthier adoptions Mihardjo et al. (2020).

## 1.2 Research Aim

The aim of this dissertation project is to create a software solution that captures real time data for a life insurance in the UK and models the underlying environment to identify the drivers of estimate changes. The software solution aims to provide insights into the UK insurance market, identify the drivers of value changes, and create a predictive model that can forecast future value changes in the market based on the identified drivers.

## 1.3 Research Objectives

- 1) Conduct a literature review on machine learning techniques for predicting price changes in the insurance market.
- 2) Gather real-time data for a life insurance in the UK and automate the data collection process.
- 3) Preprocess and analyse the collected data to determine the factors that drive changes in the UK insurance market.
- 4) Build a predictive model using machine learning algorithms that can forecast future changes in the market based on the identified drivers.
- 5) Evaluate the accuracy and effectiveness of the developed predictive model and validate the results.
- 6) Develop a software solution that enables insurance companies to collect and analyse real-time data for insurance.
- 7) Discuss the research findings and provide recommendations for insurance companies operating in the UK life insurance market. By achieving these goals, the research will provide valuable insights into the drivers of changes in the UK insurance market and offer a software solution that can help insurance companies make informed decisions to remain competitive and profitable in the market.

## 1.4 Research Questions

- 1) What machine learning algorithms and methods are most operative for predicting UK life insurance claim prospects, and how do they liken in terms of precision and interpretability?
- 2) How can historical statistics on UK life insurance claims be efficiently composed, cleaned, and pre-processed to advance the act of predictive models?
- 3) What character can feature engineering and feature assortment play in improving the predictive power of machine learning models for UK life insurance prediction?

- 4) How can natural language processing (NLP) and sentiment investigation be unified into software resolutions to examine customer feedback and social media data for better-quality risk assessment in life insurance?
- 5) What ethical and regulatory deliberations should be taken into interpretation when developing and positioning machine learning models for life insurance prediction in the UK?
- 6) "How can software platforms be designed to provide real-time updates and recommendations to policyholders based on changing life circumstances and risk profiles?"
- 7) "What strategies and technologies can be employed to ensure the security and privacy of sensitive personal data used in life insurance prediction models, especially in compliance with UK data protection laws?"

## 1.5 Rationale of the study

The study "UK life insurance projections" has various valuable goals. First, it improves insurers assess risk more perfectly. By investigating a wide variety of information, incorporating demographics, health statistics, and lifestyle factors, insurers can better understand policyholders' ability to make claims, allowing them to report appropriate compensation. Second, this research aids consumers by possibly leading to more personalized insurance offers Kar & Navin (2021). As insurers can predict risk more perfectly, they can offer individual strategies that are more reasonable and more tailored to individual requirements. In addition, such research could improve policymakers and supervisors ensure equality and transparency in the insurance industry. It could emphasize any disparity or bias in assurance practices and lead to regulatory amendments to shield consumers. In condensed, UK life insurance forecasting research influences on the financial steadiness of insurers, specifies consumers with more tailored insurance preferences, and supports in the development of regulations and fair insurance Upreti et al. (2022).

## 1.6 Research Significance

The analysis on "United Kingdom Life Insurance Predictions" is significant in numerous ways. Research improves insurers correctly assess the risks correlated with different policyholders. By examining information associated to age, health, lifestyle and other considerations, insurers can establish the likelihood of policyholders making a claim. This allocates for more accurate premium pricing, make certain individuals pay premiums that indicate their true risk profile. Accurate predictive models can facilitate save costs for consumers. When insurers can effectively make a distinction between high-risk and low-risk individuals, they can pose lower premiums to those with lower risk. This makes insurance more inexpensive for many people and promotes healthier lifestyles as individuals can be prompted to decrease risk factors to lower premiums Didenko & Sidelnyk (2021).

Research in this area could lead to the expansion of personalized insurance products. Insurers can modify policies to specific demographics or risk reports. By lowering the risk of unexpected losses due to inaccurate risk assessments, insurers can better restraint their financial reserves and ensure they have adequate resources to meet claims of contract venders. This study could specify intelligence to officials in the UK. It assists them monitor fairness and regulatory compliance in the insurance industry. Insurers can use advanced analytics to enhance their operations, involving marketing, financing, and claims handling, which can help mount efficiency and customer provision. The study of "UK life insurance projections" is significant because it promotes both insurers and clients by enabling more precise risk assessment, saving earnings Shamsuddin et al. (2022).

## 1.7 Structure of the report

The structure of the study will be based on FIVE chapters.

- **Chapter 1:** It is the introduction of the study comprising background, rationale, research aim, research objectives and research questions.
- **Chapter 2:** It is the literature review of the study comprising objectives-based information. The information is attained through secondary sources and the



relevant theories and frameworks are used to provide theoretical perspective on the topic.

- **Chapter 3:** It is the methodology chapter. The chapter comprises methodological approaches that are used for the purpose of implementing the research in a practical manner. The chapter includes research philosophy, research approach, research design, data collection, sample size and technique, data analysis and ethical considerations.
- **Chapter 4:** It would describe the findings and discussion. The findings section includes results based on method carried out stated in chapter 3. The discussion section compares the findings and literature review, and a final observation is stated.
- **Chapter 5:** It is the conclusion and recommendations chapter. The chapter comprises summarized findings, recommendations, future implications, and conclusion.

# Chapter 2

## Literature Review

### 2.1 Introduction

The price changes and predicting the prices changes is considered to be important in the life insurance sector. It helps in enhancing financial stability for insurance companies and provides an opportunity to maintain competitive positions. The landscape of insurance is evolving with the passage of time considering different factors that include demographics, risks, emerging vulnerabilities, technological changes as well as the ability of predictors to forecast the prices changes accurately and promptly. It does not only help insurers but also the related consumers and the policymakers. This literature is based on highlighting and gathering the real time data for life insurance companies in the UK and the process of automating the data collection processes. Additionally, the literature also analyzes the challenges and factors that are associated with changes in the UK insurance market and price prediction. The literature also provides recommendations for the insurance sector for enhancing their data accuracy for predicting and forecasting the prices. The literature review has been conducted by critically analyzing the past researcher, journals and articles that are based on similar aims and objectives as of this study. The researcher has made sure to include papers that are not older than ten years to include updated information for maintaining authenticity and reliability of the research.

## 2.2 Machine learning practices that are used in life insurance companies

In the insurance market, predicting changes in the prices is considered to be not only essential but also a complex task which is based on analysis of a diverse data set. According to the study of Polydoros & Nalpantidis (2017) the Machine Learning technology has recently gained prominence in the life insurance companies across the world which has helped in addressing the challenges associated with predicting fluctuations in the prices through its predictive models which has helped insurers to implement effective and efficient pricing strategies for maintaining competitive position in the operating industry. Thomas & Brunskill (2016) highlighted that one of the commonly used machine learning features is linear regression. It is considered to be a simple yet interpretable method for analyzing and forecasting the changes in price in the life insurance industry. The model also provides insights about the linear relationship which exists between the dependent variables that are also known as target variables and independent variables. It is essential to understand these relations for gaining valuable insights about the different drivers of prices in life insurance companies. However, Wang & Xu (2018) claimed that the assumption between the dependent and independent variables may often not be able to capture the complex data set related to the prices. Moreover, the linear regression is considered to be incapable for determining the non linear patterns that exist in the insurance data.

It has been claimed by Jordan & Mitchell (2015) that it is easier for insurers to comprehend the decision trees that are made through Machine Learning. It helps in assessing better visualizing and enables insurers to gain more categorical as well as numerical data for predicting prices in the life insurance industry. Furthermore, it also provides more accuracy considering different ensemble methods based on random forests and gradient boosting trees. Nevertheless, Finlayson et al. (2019) debated that the training data is often considered to overfit the decision trees which leads to poor generalization. It does not enable the insurers to capitalize the interactions that are subtle between the variables. Farchi et al. (2021) also claimed that the deep learning models that are especially based on neural networks are able to intricate different

patterns that are present in the data. It has capability to also organize the unstructured data through different texts and images such as convolutional neural networks and recurrent neural networks have shown to be significant in the implementing pricing strategies in the life insurance companies. Das et al. (2020) stated that it is often required by deep learning models such as Machine learning to have more computational resources which can enable it to capture a large amount of data. These are considered to be black boxes for the insurance sector which makes it challenging for the life insurers to interpret the data and make data driven decisions for meeting the needs and demands of consumers.

There Are also certain advantages and disadvantages that are associated with machine learning in the life insurance sector. According to the study of Wang & Xu (2018) machine learning is considered to be efficient and effective for discovering and highlighting the complex patterns and relationship that persists between different variables in a data. It enables the service providers to gain more insights about the drives that change the prices of life insurance. Padmakumari & Shaik (2023) also stated that there are numerous Machine Learning based algorithms that are capable of handling a number of types of data which also includes categorized the unstructured data and converting it into a structured format. However, Sauce et al. (2023) claimed that ML algorithms such as decision trees are prone to providing the wrong access data from the sources especially if these algorithms are not validated and regularized on a daily basis. There is also a lack of interpretability that has been observed when it comes to complex models such as deep learning neural networks. It makes it challenging for life insurers to explain the predictions on the basis of models. Additionally, Maier et al. (2019) also claimed that the models based on Machine learning are also overly dependent on the quality of data. Due to noisy data or assessing incomplete data can cause the formation of inaccurate predictions which may affect the decision making process and competitive advantage of the insurance company.

## 2.3 Use of predictive modeling in life insurance sector

There are a number of significant advancements that have been made in the life insurance sector through predictive modeling which are based on the availability of data, data sources as well as increasing the complexity of the risk assessment. The life insurance sector uses predictive modeling to determine the rate of mortality in life insurance. The insurers use predictive modeling for collecting the data related to the medical history of policyholders, lifestyle and different factors that can affect them. It helps in predicting the life expectancy of policyholders and determining the rate of mortality. The insurers also use Machine Learning algorithms that are based on decision trees and random forests for developing predictive models which also outperform the traditional actuarial methods that are being used in life insurance companies. Furthermore, Thomas & Brunskill (2016) stated that many life insurance companies are relying on the use of predictive models for automating the underwriting processes. The data of applicants is being analyzed through machine Learning algorithms which also include medical records and financial history. It helps in expediting the underwriting process for making improvements in the decisions. It further helps in enhancing customers' experiences and reducing the operational costs of life insurance companies. There are a number of methods and sources that are used for collecting data in the life insurance industry. Jordan & Mitchell (2015) stated that during the application processes, insurers ensure that they are collecting the data directly from the policy holders. They gain the data related to the personal health and wellbeing of insurers where they also include the personal information such as address, date of birth and building address while including all the demographic details about insurers. This data serves a pivotal role in assessing the risks related to policyholders in the life insurance industry. Polydoros & Nalpantidis (2017) also stated that claims data are another important source of collecting data in life insurance companies. Through claims data the insurers are able to include the information about the damages, accidents, injuries as the costs associated with such an event.

There are a number of technological tools such as Mobile apps and online portals through which claims data can be collected easily by life insurers which provide immediate insights about expediting the processes of claims. Finlayson et al. (2019) further stated that there are also a number of telematics and IoT devices that are being used in the life insurance companies for the purpose of collecting the data. Telematics devices are being installed in vehicles whereas IoT devices are installed in homes for collecting the valuable and real time data related to the location, behavior and conditions of the policyholders. It has been stated that this data is considered to be essential for analyzing the risks and pricing of any accident or scenario. Insurers also offer investment-linked products to customers and policy holders that are based on real-time market data, interest rates as well as analyzing the stock prices in the insurance industry. By assessing this data timely provides an opportunity to insurers to adjust their offering as per the market conditions and ensure they are still capable to meet the needs and requirements of customers successfully.

Wang & Xu (2018) claimed that though amidst challenges that are associated with assessing the real-time data for life insurers, automation is considered to be one of the prior solutions to eliminate the existing challenges and enhance the value of life insurance industry in the operating market. Padmakumari & Shaik (2023) claimed that collecting the data manually is not only slow but it may also increase the prevalence and vulnerability of errors. When the data related to customers and their health is being collected through automation, it is made sure that the collected data is based on real-time or near-real-time. It makes sure that the insurers are able to access the latest information about the policyholders and the target audience on time swiftly and approach them successfully for providing different options that can help them through life insurance policies. Farchi et al. (2021) further claimed that automation has the ability to reduce the human error risks which enhance the accuracy of the collected data. As the life insurance industry is considered to be a sector which is based on gaining insights about the data and then making the decisions, the human errors can increase the associated risks with the pricing strategies which may increase the prevalence of fraudulent activities and negatively influence the operations and revenue of life insurance companies.

## 2.4 Challenges and importance for collecting real time data in insurance companies

The landscape of life insurance companies across the world is changing with time where data is considered to be a crucial and pivotal element for making informed decisions in insurance organizations. Collecting and assessing the real time data is considered to be challenging for insurers where the main challenges associated are volume, sources and variety of data that has been collected. Thomas & Brunskill (2016) stated that one of the major challenges for life insurers is maintaining the quality and consistency in data collection. The real-time data is supposed to be gathered from numerous platforms and sources where each source possesses its own format as well as standard of quality. Considering this diversity in the number of sources of data collection, it creates discrepancies, missing values as well as errors which significantly affect and alter the decision making process in life insurance companies. Shen et al. (2020) stated that it is essential to ensure that the collected data through technological tools such as that of Machine learning are accurate as well as clean so the principles in the insurance companies are not affected and increases the vulnerability of processing claims which is not only costly but also may affect the competitive advantage.

Jordan & Mitchell (2015) also stated that another challenge in privacy of data, with the rise in technological advancement, has grown the prevalence risks and vulnerability of cyber attack in the insurance sector. The author further stated that insurers are responsible for collecting and handling sensitive information related to the customers which may also include their personal information, credit card and debit card details as well as other personal information. It has become essential for insurers to use strict rules and regulation for eliminating the rising risk of cyber attack in the industry and enhance security and privacy to eliminate data breaches. The consequences of not complying with the rules and regulatory policies can also impose strict penalties on the life insurance sector which may further affect its competitiveness and position in the operating industry.

Polydoros & Nalpantidis (2017) also highlighted another challenge that is associated with collecting data in the life insurance industry is integration of data. The legacy

systems that are being used by many insurers are not designed for integrating the modern sources of data collection. As the real-time data is being accessed from a number of different external sources which also includes market feeds, IoT devices, the task of integrating the data into the internal system is not only time-consuming but also complex in nature. It may create a bottleneck in data integration which may further slowdown the decision making processes and negatively influence the ability of insurers to adapt in the changing circumstances.

Maier et al. (2019) claimed that though there are a number of challenges associated with data collection in the life insurance industry, assessing real-time data is of imminent importance for the industry. Sauce et al. (2023) also stated that the accessing of real time data ensures to highlight the risks and vulnerabilities that are associated with life insurance industries and the prices. The telematics data for example can be beneficial in getting the real insights about the behavior of drives which enable the insurer to make predictions about the possible risks and eliminate them by implementing strategies. It further provides an opportunity to insurers to personalize pricing and develop more effective risk elimination measures. Finlayson et al. (2019) also claimed that another critical area that is being addressed through data collection is detection of any type of fraud which can affect and impact the performance of life insurance companies in the operating industry. The advanced analytical tools such as Machine Learning helps in analyzing the real-time data which determine the unusual patterns that exist and helps in detecting fraudulent crimes and claims that may occur in life insurance companies. The life insurers use this proactive approach to ensure that they are able to save a significant amount of loss on money by preventing any payout for fraudulent activity. Thomas & Brunskill (2016) further stated that the importance of real time data for providing and enhancing customers experiences in the life insurance sector can not be ignored. The policy holder expects to receive instant access and information about the premiums, claims as well as policies of life insurance companies. By equipping them with the real time information through digital channels helps in enhancing customers rate of satisfaction which further tends to increase customers loyalty and responsiveness towards the life insurance companies. Considering customers' loyalty and trust towards the brand, the insurers get an opportunity to adjust their



premium packages of insurance policy on the basis of changing market conditions and needs and demands of customers to make sure that the pricing strategy of the company is more competitive, yet accurate.

## **2.5 Building a Predictive Model Using Machine Learning Algorithms**

The industry of life insurance has witnessed a significant transformation in these recent years which is driven by advancements in the field of technology and data analytics. However, Machine Learning (ML) algorithms have now become indispensable tools in sector of predictive modeling which has enabled the insurers to make more accurate predictions, streamline operations, and comparatively enhance experiences of the customer Zhong et al. (2021). There are some key steps involved in the process of building a predictive model using ML algorithms in the life insurance industry.

### **2.5.1 Data Collection and Preprocessing**

Data is considered as the foundation of any successful predictive model. In context to life insurance industry, data is abundant which came from various sources that includes policyholder information, claims data, medical records, and financial data Simjanoska et al. (2020). This means that the first step is to collect and aggregate these widespread data into a structured format which is suitable for analysis.

Once the data is collected, it is important to preprocess these data files to clean and prepare it for modeling. This includes handling missing values, outlier detection and treatment, feature engineering, and data normalization Rajendran & Karthi (2022). However, feature engineering is particularly important in the life insurance industry, as it includes modification of existing variables or creating new ones in order to improve the predictive power of the model Chowdhury et al. (2022).

### **2.5.2 Model Selection**

Selecting the right ML algorithm is important and critical for the success of the predictive model. In the life insurance industry some common algorithms are considered

which include logistic regression, decision trees, neural networks, random forests, gradient boosting Kiguchi et al. (2022). The choice of algorithm majorly depends on the specific problem at hand, the complexity nature of the data, and the desired interpretability of the predictive model Mangold et al. (2021). For instance, logistic regression, is often used for predicting binary outcomes like whether a policyholder will file a claim or not. However, more complex algorithms like gradient boosting can also be employed for modeling more intricate and complex relationships within the data Zhou et al. (2019).

### **2.5.3 Data Splitting and Model Training**

In order to assess the accurate performance of the model, the dataset is split into training and testing sets. The training set is often used to direct the model patterns in the data, whereas, the testing set is used to evaluate the performance of model on an unseen data. There are some common techniques that are used for splitting data such as random sampling and k-fold cross-validation White & Power (2023). The chosen ML algorithm is then trained on the training data, during which, the model keep on learning from the data and make adjustments into its parameters for minimizing the prediction error. This overall iterative process continues until the model converges to an optimal state of consideration Farchi et al. (2021).

### **2.5.4 Model Evaluation**

After the training of model, the model's performance must be rigorously assessed. Multiple metrics have been used in the life insurance industry which depends on the specific predictive task. Common evaluation that metrics include are accuracy, F1-score, precision, recall, and the area under the Receiver Operating Characteristic (ROC-AUC) curve Imani & Arabnia (2023). For instance, in the process of predicting insurance fraud, precision the ratio of true positives to all predicted positives) may be more critical than overall accuracy of the model. In this case, a false positive could lead to unnecessary investigations and might increase the overall operational costs of the model.

### 2.5.5 Model Deployment

Once the predictive model has been trained and evaluated carefully and accurately, it is ready for deployment in a real-world environment. This process involves the integration of the model into the operational systems of insurer. This will allow them to make predictions on new data in real-time. The process of model deployment also requires frequent real-time monitoring and maintenance. As data and business conditions change, the performance of the model may degrade. In order to avoid any errors, regular updates and retraining are necessary to ensure the accuracy and relevancy of an order Ivanovic et al. (2023).

### 2.5.6 Interpretability and Compliance

Interpretability holds a potentially significant concern in the life insurance industry, given its regulatory environment. Insurers must be able to explain the rationale behind their decisions to customers and regulators. Therefore, it is important to use interpretable ML algorithms and provide transparency in the decision-making process of model Das et al. (2020). Furthermore, models in the insurance industry must comply with legal and ethical standards which include anti-discrimination laws, fairness and bias mitigation. These techniques should be integrated into the modeling process for ensuring that the model does not discriminate against specific demographic groups Sauce et al. (2023).

## 2.6 Machine Learning's Impact on Price Forecasting in Insurance Markets

Algorithms of machine learning have revolutionized multiple industries by enabling data-driven decision-making and predictive modeling. In reference to the life insurance market, these algorithms have played a crucial role in forecasting changes in pricing which has helped insurers for optimizing their pricing strategies and effectively manage associated risk.

### **2.6.1 Risk Assessment**

Accuracy and relevancy of risk assessment in the insurance industry have significantly been improved by machine learning algorithms. These algorithms meticulously and rigorously assess and analyze extensive datasets which includes historical claims, customer profiles, and other external variables like weather and economic conditions. Through this process of identification of intricate patterns and correlations within this vast dataset, insurers can make more effective and precise predictions based on the likelihood of certain future claims Yang et al. (2020). Moreover, this accuracy would help in empowering insurers to adjust premiums with greater precision in order to ensure that policy pricing aligns closely with actual risk factors. This will ultimately benefit both insurers and policyholders.

### **2.6.2 Pricing Optimization**

Insurers adapting machine learning to fine-tune pricing structures by observing and understanding the behavior and preferences of the customer. These advanced algorithms categorize policyholders into distinct risk segments which eventually enable them in acquiring tailored premium adjustments Alet (2023). Further, this diversified approach for ensuring insurers strike a balance between competitiveness and profitability, adapting pricing strategies in response to ever-evolving conditions of market and customer dynamics. By acquiring machine learning, insurers can enhance certain pricing precision which better serve their clientele and also aid in maintaining a resilient financial position in the highly competitive market of insurance Aslan (2021).

## **2.7 Methods for Evaluating the Accuracy and Effectiveness of Predictive Models in the Insurance Market**

In current data-driven world, the insurance industry is increasingly relying on predictive models for making informed decisions, assessing risks, and optimizing pricing strategies. These predictive models, often powered by machine learning algorithms

that play a crucial role in helping multiple insurance companies in order to maintain their competitive and profitable edge. However, in order to ensure that these models are reliable and effective, insurers must employ various robust methods for evaluating accuracy and effectiveness of predictive models.

### **2.7.1 Data Quality Assessment**

Before delving into evaluation of predictive model, it is very important to start with a strong foundation for ensuring high-quality data. However, several insurance companies typically have vast datasets that consist information of policyholders, claims history, and other relevant data. Data quality assessment involves examining the completeness, accuracy, relevancy, and consistency of this data set Syed et al. (2023). Data errors or inconsistencies that might lead to flawed models. This is why, it is critical and essential to clean and preprocess the data effectively.

### **2.7.2 Model Performance Metrics**

To evaluate predictive models, insurers often rely on various performance metrics that depends on the nature of the problem they are solving. These metrics include, Mean Absolute Error (MAE) which measures the average absolute difference between predicted values and actual values, Root Mean Square Error (RMSE) which works by calculating the square root of the average squared differences between predictions values and actual values, accuracy, precision, and recall are often used for classifying problems Brassington (2017). These metrics assess and analyze how well the predictive model correctly and efficiently identifies positive and negative outcomes.

### **2.7.3 Cross-Validation**

Cross-validation is another technique which is used to assess the performance of model while mitigating the risk of overfitting Kernbach & Staartjes (2022). However, K-fold cross-validation involves splitting the data into K subsets followed by training the model on K-1 subsets leading to evaluating it on the remaining subset Marcot & Hanea (2021). This overall process is repeated K times, and the results of evaluation are averaged for providing a more robust estimate of the model's performance.

### **2.7.4 Confusion Matrix**

In classification tasks, confusion matrices play an important role in providing a detailed view of a model's performance Deng et al. (2016). It works by breaking down the predictions into categories which includes results like true positives, true negatives, false positives, and false negatives. This information is useful for understanding where the model is making errors while predicting and can help in refining the model accordingly Shen et al. (2020).

### **2.7.5 Receiver Operating Characteristic (ROC) Curve**

ROC curves is another technique which is used in binary classification tasks for visualizing the ability of a model to distinguish between positive and negative cases at different threshold levels Westphal & Seitz (2021). However, the area under the ROC curve (AUC-ROC) is a common metric which help in quantifying the discriminatory power of a model Hare & Kutsuris (2022).

### **2.7.6 Profitability Metrics**

The ultimate goal of insurers is to optimize profitability. Therefore, it's crucial to evaluate predictive models in terms of their impact and efficacy. Metrics like the Net Promoter Score (NPS) and customer lifetime value (CLV) aid in assessing the long-term financial impact of model-driven decisions making processes Croft (2015).

### **2.7.7 Backtesting**

Backtesting is one of the important technique which work particularly in relevance to the insurance market, where predictive models are often used for risk assessment and pricing management. It involves assessing how well a predictions of a model align with actual outcomes over historical data Padmakumari & Shaik (2023). This technique helps insurers for understanding how the model would have performed in the past and adjust its own assessment strategies accordingly.

### **2.7.8 Ethical and Fairness Evaluation**

In the insurance industry, ensuring fairness and ethical considerations in predictive models is crucial. Insurers should evaluate whether their models exhibit bias or any other discrimination against certain groups based on demographics. Various fairness metrics, like disparate impact and equal opportunity can aid in assessing and addressing these issues Huang et al. (2022).

### **2.7.9 Continuous Monitoring**

Even after the deployment of predictive model, the process of evaluation does not end. In every ML algorithm predictive model, continuous monitoring is essential for ensuring that the performance of model remains accurate and effective over the period of time. This includes tracking changes in distribution of data, retraining the model when necessary, and adapting to ever-evolving business and customer needs Chekroud et al. (2021).

## **2.8 The Importance of Using Appropriate Performance Metrics and Cross-Validation Techniques in Life Insurance Companies**

Life insurance companies play an important role in safeguarding the financial future of multiple individuals and families. The decisions they make, from setting premium rates to assessing risk, have far-reaching consequences. However, in order to ensure the decisions made by them are accurate and fair, it is essential for life insurance companies to employ appropriate performance metrics and cross-validation techniques.

### **2.8.1 Performance Metrics in Life Insurance**

Performance metrics are essential tools that help various life insurance companies for assessing their operations and make highly informed decisions. These metrics provide quantifiable data which enable these companies to gauge their performance and identify areas for improvement, for optimizing their strategies. In reference to life insurance

there are several key performance metrics that must be considered:

### **2.8.2 Policyholder Retention Rate**

Retaining policyholders is essential for long-term profitability. A high retention rate indicates that the customer is satisfied and loyal. This can be indicative of effective underwriting, fair pricing along with strong customer service Kajwang (2022).

### **2.8.3 Claims Ratio**

This metric measures that the ratio of claims paid out to premiums collected. Further, a balanced claims ratio is essential in ensuring the financial sustainability and efficacy of the company while honoring its commitment to the existing policyholders Taplin (2021).

### **2.8.4 Profit Margin**

Life insurance companies are businesses and they prioritize their profitability which is their significant concern. However, monitoring profit margins helps in ensuring the company remaining financially stable and can continue to provide coverage to policyholders Battiston et al. (2019).

### **2.8.5 Customer Satisfaction and Net Promoter Score (NPS)**

It is the most observed practice that satisfied customers are more likely to stay with their insurance provider. However, high NPS scores indicate that the customers are likely to recommend the company to others and can convince them based on their own experience. This can lead to organic growth Schlosser (2023).

### **2.8.6 Underwriting Accuracy**

Effective underwriting is one of the foundations of insurance. Metrics that are related to underwriting accuracy including the ratio of accepted to rejected applications and the incidence of policy lapses can help in assessing the quality of underwriting decisions Maier et al. (2019).



### 2.8.7 Cross-Validation Techniques in Life Insurance

Cross-validation techniques are statistical methods that are used to assess the performance and generalization of various predictive models. However, in the landscape of life insurance, the utilization of cross-validation techniques is paramount for various purposes. Firstly, it helps in ensuring model accuracy which is a fundamental requirement for insurance companies as predictive models underpin several risk assessment and premium rate determination Bermúdez et al. (2023). Moreover, it plays a pivotal role in overfitting prevention which can be achieved by scrutinizing model performance on diverse data subsets, guarding against the common pitfall of models excelling in various training data but usually function oppositely when applied to unseen data Mohanty & Palai (n.d.). Further, model selection also benefits immensely from cross-validation as it offers an unbiased method for comparing and choosing the most effective predictive model. Lastly, it also aids in risk assessment, refining models in order to mitigate the uncertainty and threat of underpricing or overpricing policies Prabhudesai et al. (2023).

### 2.8.8 Benefits of Using Performance Metrics and Cross-Validation

The utilization and integration of appropriate performance metrics and cross-validation techniques offers a multitude of invaluable advantages to the companies working on life insurance. It facilitates data-driven decision making by providing a wealth of objective data which empowers companies to make several well-informed choices regarding underwriting, pricing, and customer service Wang et al. (2022). This aid in ensuring that their structured strategies align completely with the market realities, behaviors and preferences of their customers. Further, these tools play an important role in risk mitigation. However, cross-validation, in particular, acts as a safeguard for protecting against potential pitfalls that might impact predictive models Tahraoui et al. (2023). This approach help in identifying and mitigating risks associated with inaccurate assessments. This, in turn, helps in ensuring that the insurance companies can reliably evaluate several risks that safeguard their financial stability, and eventually uphold their commitments to policyholders Shah et al. (2022).

# Chapter 3

## Methodology Design

### 3.1 Analysis of the Methods

In data analytics, it is observed that there are four main kinds of data analytics which include descriptive, diagnostic, predictive, and prescriptive Wissuchek & Zschech (2023). For this study, the researcher incorporated Python method to predict the insurance charges that helps in analysing the prices of life insurance within the UK. In order to examine the method, several codes and Python script has been used. The purpose of selecting Python was that it assists in data preprocessing, statistical analysis, machine learning, and visualisation. As observed, in this study, the researcher uses Python for training the data and run learning algorithm to predict on insurance charges. There are three types of data that have been used in this study which include sample data, train data, and real-time data.

In order to test or predict insurance data initially descriptive statistics was carried out that helps in determining the age, sex, bmi, children, smoker, region, and charges. With the help of this analysis, the researcher carried out mean, std, min, and max age, bmi, children, and charges. On the other hand, and Insurance Expense Chart was formed that provides the analysis of the ups and down of the insurance expenses. Although, there are several codes that have been developed by the investigator for assessing the data where conversion to data was assessed at first. The other codes that have been generated and analysed are creation of training and testing, running learning algorithm, prediction and testing, and MAPE scoring. Where MAPE scoring is the one that enables towards identifying the data.

In this regard, it is determined that the main purpose of utilising Python was

that it is considered as one of the most efficient techniques as compared to other methods. Python is recognised for its simplicity and readability Eghbali & Pradel (2022). While the syntax is precise and easy to comprehend, that makes it accessible for the learners and experienced programmers alike. This simplicity therefore, minimises the time and effort required for writing and maintaining the code that leads to increased efficacy. Apart from it, Python encompasses of large standard library that offers a broad variety of pre-built modules and functions whereas, it has a vast ecosystem of third-party libraries and models. Moreover, python is increasingly versatile and could feasibly assimilate with different programming languages and systems Srujana et al. (n.d.). Hence, it could be used for scripting, development of web, data analysis and more. The capability of Python is to integrate with other tools and systems that improves the efficiencies by permitting data exchange and interoperability with present infrastructure.

### **3.2 Design of Methodology**

The design of this study comprises of real time analytics comprising of algorithms and logic for data insights for better decision making. Generally there are two types of research design employed within research studies including qualitative research design and quantitative research design Bloomfield & Fisher (2019). The quantitative research design comprises of statistical analysis. The quantification of the data help in acknowledging the changes in life insurance prices with prediction model. The existence of prediction model indicated inclusion of data mining technology that help to analyse the current data and historical data for future outcomes Das et al. (2020). The significance of applying the quantitative data design with predictive modelling help in use of machine learning and data mining tools that help to forecast the future outcomes relating to the life insurance prices. This design also help in identifying the drivers of price changes. The software solutions design for the prediction of life insurance charges can be adaptable to other insurance markets and products as well. Farchi et al. (2021), study also showed significance of using quantitative research design in providing evidences and predictions. The incorporation of mathematical analysis tools and software

helped in attaining greater insights to the data. The design of the research study also incorporated descriptive statistics that help to summarise the data with adequate description. The selection of quantitative research design helped in findings the patterns and averages to make predictions and form generalise results for wider market or population. On other side, the benefits of including quantitative research design help in replicating the study by indicating standardised data collection protocols and tangible definitions. Moreover, it also help in comprising the direct results of the study statistically Watson (2015).

### **3.3 Evaluation Methods and Measures**

#### **3.3.1 Data Collection Method**

To collect and measure information through preferred tools and devices is known as data collection method Siedlecki (2020). There are two main types of data collection methods namely primary and secondary data collection method. Primary data collection is the one that is obtained directly from the target audience in raw form therefore, it is known as first-hand data. The data is collected through interviews, observations, questionnaire or focus group Mazhar et al. (2021). In contrast, secondary data is the one where the data is acquired from existing books, journals, articles, online websites, and case studies Sileyew (2019). For this study, the investigator utilises secondary data collection method. For data collection, the primary stage comprises of collecting real-time pricing data for the UK-based retail insurance product. However, this process will be automated by utilising web scraping tools for collecting data from relevant websites. The rationale behind choosing secondary data collection method is that it is easily accessible, saves time, and allows to generate new insights from prior analysis Ruggiano & Perry (2019). It enables towards providing large amount of secondary data with a broad variety of sources that helps driving appropriate conclusions regarding making predictions for life insurance in the UK.

### 3.3.2 Data Analysis

#### Data Preprocessing

Data preprocessing is referred to as the crucial phase in analysing the data and machine learning pipeline. It entails the preparation of raw data for making it appropriate for further analysis or modeling. It comprises of different steps which include cleaning, transforming, selecting features, encoding, handling excessive data, data integration and data normalisation Mishra et al. (2020). After data collection, it will be preprocessed for ensuring cleanliness and suitability of the analysis. This entails eradicating outliers or missing values and transforming the information into relevant format for the analysis.

#### Exploratory Data Analysis

The subsequent stage entails carrying out exploratory data analysis for understanding the underlying patterns and data trends Milo & Somech (2020). This include visualising the data and making predictions accordingly by testing sample data, train data and real data.

#### Feature Selection

After exploring and analysing the data, the researcher will opt the most important aspect that are affecting the price changes in the UK insurance market. This requires making use of statistical techniques such as Python analysis that enables towards the testing training and testing data sets, implementing machine learning algorithms and making predictions accordingly and conducting MAPE score that assists in identifying the error.

#### Model Selection and Training

After rectifying the relevant features, an efficient machine learning model will be opted for predicting the price changes. This might entail utilising Python analysis, clustering algorithms, and neural networks relying on the research nature and research question for making predictions on life insurance in the UK. However, this model will be trained on the preprocessed data.

## **Software Solution**

This research will lastly provide an efficient software solutions which depicts that this dissertation project will assist in providing a user-friendly and customizable software solution that could apprehend real-time pricing data for the UK based retail insurance product and model considering the drivers of price changes. The software will be designed in order to be flexible to other insurance markets as well as the products that helps in making predictions on insurance prices.

## **Ethical Risk**

There are number of ethical considerations that needs to be considered while conducting the research where the web scrapping process might be influenced by website changes that results in missing or in completing the data. In order to mitigate these risks, various websites will be targeted along with data collection scripts that will be updated rapidly for adapting these changes. It could be identified that the data preprocessing stage might result in the loss of significant data or the inclusion of unnecessary data. For mitigating this ethical concern, number of other data preprocessing methods will be used where the results will be compared for ensuring that they are consistent. Other than that, the machine learning model might not perform as expected because of several problems such as overfitting and underfitting. Moreover, the data collected during this project might contain sensitive data regarding individuals that should be handled in an accountable and ethical way.

## **Limitations**

There are certain limitations that were encountered while conducting the research where time limitation is one of the most crucial constraint. Due to unforeseen events or delays in data collection might cause time constraints for the project. While the other constraint is of financial limitations where unanticipated expenditures or changes in scope might lead towards financial limitation for the project. To mitigate this risk, a financial plan will be developed. The other constraints found are regarding technical problems in terms of software bugs or failure in hardware. And lastly the other constraint could be of resources, due to limited resources, the researcher might

result in delaying the project.

### 3.4 Tools and Resources

In Python, there are various tools and libraries that are accessible for data conversion tasks. The tools or the resources that are most commonly used include Pandas, NumPy, xlrd and xlwt, JSON library, csv module, XML libraries, SQL Alchemy and PySpark. It is observed that Pandas is one of the prevailing data manipulation library in Python that provides functions for reading and writing the data in distinct formats which include CSV, Excel, JSON, SQL databases, and more (Teimourzadeh, Kakavand and Kakavand, 2023). Besides, NumPy is a basic library for numerical computing in Python that offers efficient array operations and mathematical functions. It can be used for converting data among distinct array formats, reshaping data, and make calculations while the ongoing process of conversion.

Other than that, in Python the tools and resources used for conducting trainings and testing datasets for machine learning or data analysis tasks. These tools and resources include Scikit-learn, NumPy, Pandas, Stratified Sampling, K-Fold Cross-Validation, Random Sampling, and Data Augmentation Blaiszik et al. (2019). The optimal choice of the tools and approaches are relied on particular needs of the project, for instance size of dataset, distribution of classes, and assessment strategy. It is essential to carefully consider the characteristics of the dataset and the objectives of the analysis and machine learning task for the development of training and testing datasets.

The tools and resources that could be used for running learning algorithms and integrating the models for machine learning entails Scikit-learn, TensorFlow, Keras, PyTorch, XGBoost, LightGBM, CatBoost, and H2O that is relied on the algorithm type, and the flexibility level and required performance Brownlee (2021). The other tools and resources used for the prediction and testing in Python entails Scikit-learn, TensorFlow, Keras, PyTorch, Scikit-plot, Yellowbrick, mlxtend, and NumPy that enables towards computing predictions, calculating evaluation metrics, and required visualisations Savran et al. (2022).

### 3.5 Summary

In this chapter, the indication of appropriate use of data collection tools and techniques significantly contribute in replicating the research study. In addition, complete in-depth information is provided within this chapter. The data collection methods and use of algorithms for the making of prediction model for life insurance charges estimation is discussed within this chapter. The presence of information indicated use of quantitative research design as well as software's such as python, data conversion method and descriptive tools. The method of analysis includes statistical demonstration of data which was further interpreted and discussed in the next chapter. It is indicated through methodology that real time pricing data for used within this study for UK based life insurance products. From relevant website the data was collected for real time pricing analysis of life insurance product. It also incorporates exploratory data analysis which helped in determining the trends and patterns in the data. This analysis benefited in examining the strength among the variables. Depending on the nature of the data and research question the algorithms, regression models and other analysis techniques were incorporated. Lastly, comparison of attain predictive model prices with the actual prices was done.



# Chapter 4

## RESULTS AND FINDINGS

### 4.1 Introduction

The study's outcomes and conclusions are presented in this chapter, as tables and graphs are employed to illustrate the data as necessary, and the results are presented clearly and succinctly. The implications of the findings for ongoing study are also taken into consideration when they are explained in relation to the research questions and hypotheses. Moreover, the chapter also provides discussion of objectives along with the predictive model of the companies used for interpretation.

### 4.2 Findings in Insurance Company Stock

The table includes data on the Admiral Group, Aviva Group, and Direct Line Group named as three separate businesses. These companies are listed in the table's rows, and the table is divided into different categories to provide with information on their stock performance. In addition, the table shows various averages for each business, as the opening average is the typical initial price at which their stocks have traded over a given time period. The low average demonstrates the average lowest stock price, while

	opening average	High average	low average	Min value	Max Value	Mean	Total number of Days	average number of shares
Admiral Group	2138.77	2166.31	2113.15	1869.54	2490.00	2133.94	252.00	823111.59
Aviva Group	418.65	422.71	414.88	345.02	426.79	394.98	252.00	7983920.25
Direct line Group	176.77	179.63	174.03	133.90	235.30	176.80	252.00	5949083.94

Figure 4.1: Table1

the high average indicates the average highest price their stocks saw through this time. The min value and max value illustrates the lowest and highest recorded stock prices for the given time period, respectively.

The maximum number of shares outstanding for each company on any given day throughout the reporting period is displayed in the high section. The highest number of shares outstanding for Admiral Group was 2138.77 on day 1. The highest number of shares outstanding for Aviva Group was 418.65 on day 2. On day three, there were 176.77 shares outstanding for Direct Line Group. The average number of shares outstanding for each company throughout the reporting period is displayed in the mean section. The average number of shares outstanding for Admiral Group was 2133.94. The average number of shares outstanding for Aviva Group was 394.98. The average number of shares outstanding for Direct Line Group was 235.30. The lowest number of shares outstanding for each company on any given day throughout the reporting period is displayed in the low section. The lowest number of shares outstanding for Admiral Group was 1869.54 on day 5. The lowest number of shares outstanding for Aviva Group was 345.02 on day 6. On day 7, there were 174.03 fewer shares outstanding for Direct Line Group.

According to the table, Admiral Group maintained the highest average number of shares outstanding, followed by Aviva Group and Direct Line Group. Based on this, Admiral Group, Aviva Group, and Direct Line Group are the three companies with the most stockholders. The table also shows that during the reporting period, each company's average number of shares outstanding fluctuated. This is probably caused by a number of variables, including changes in the stock price, the number of shares issued, and the number of shares the firm has bought back. Additionally, it is important to note the variation in the typical number of shares outstanding for each firm during the period of reporting. This fluctuation is caused by a number of variables and indicates how volatile the stock market is. Moreover, variations in the stock price of a corporation might affect the typical number of outstanding shares. If a company's stock price increases, existing shareholders may find it more compelling to continue holding onto their shares, which could result in a drop in the total number of outstanding shares. The average number of shares outstanding will rise, if the stock

price drops, which may promote the issue of new shares.

#### 4.2.1 Admiral Tests SMA

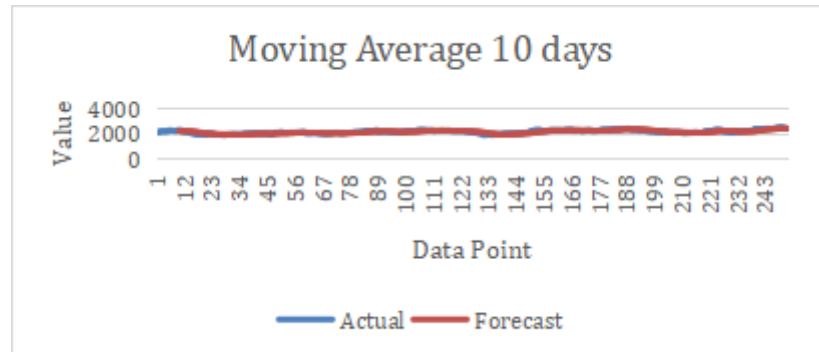


Figure 4.2: Graph1

The above graphs show that, the MA has been going upward since January 2023, indicating that an upward movement is driving the price of Admiral shares. However, there have been some cases where the price has fallen below the MA, which can indicate the beginning of a negative trend in the near future. Since the most recent data point (241) is above the MA, the bullish trend appears to still be in effect. Although the price is quite close to the MA, it is still possible that the trend will change soon.

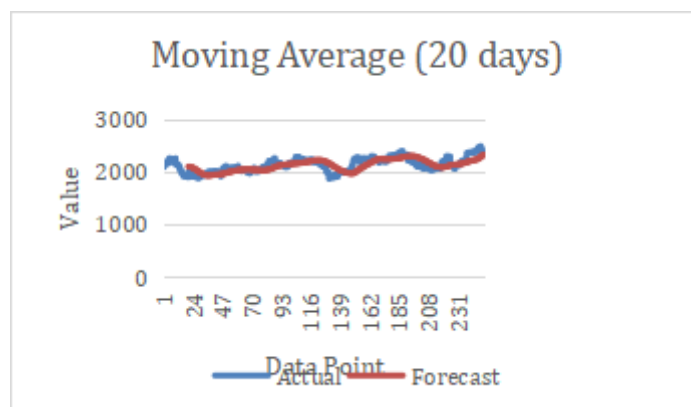


Figure 4.3: Graph2

The price of Admiral shares has been heading upward since January 2023, but the moving average is beginning to dip lower, which would indicate the beginning of a negative trend. Investors must keep a close eye on the chart to observe whether the moving average ever breaks below the data points. In other words, the trend is strong right now, but there is a chance that it could turn negative.

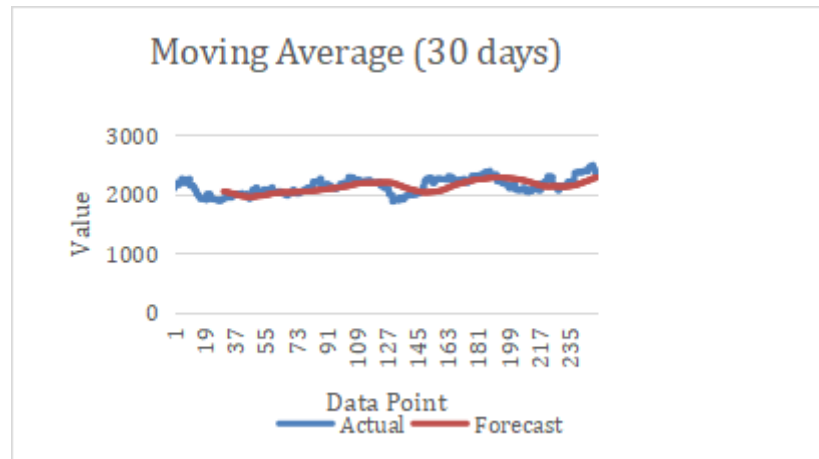


Figure 4.4: Graph3

The moving average has been rising upward since the start of the year indicating that moving times are increasing shorter on average. The moving average has, occasionally fallen below, which could suggest a temporary increase in the average moving time. The most recent data point which is 239 is higher than the moving average, indicating that the downward trend in average movement time is still present. But because the price is so close to the moving average, it's probable that the trend will soon change.

#### 4.2.2 Aviva SMA

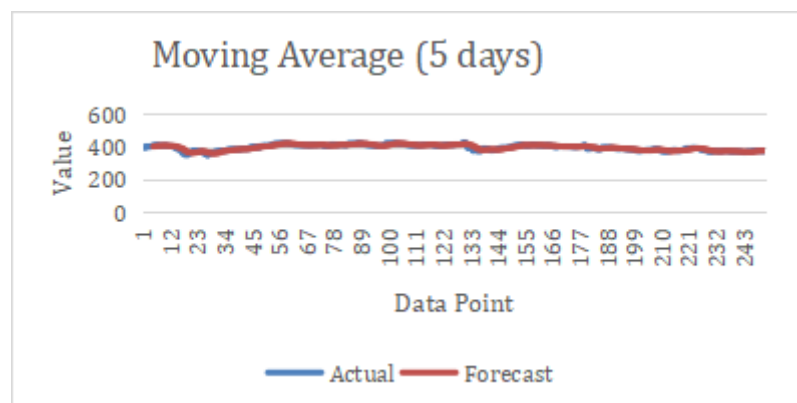


Figure 4.5: Graph4

The graph demonstrates that since the start of the time period, the actual values have been going upward. It may indicate that the price of Aviva stock is rising. The anticipated figures, however, have been moving in the wrong direction, indicating that the price of Aviva shares is likely to fall in the near term.

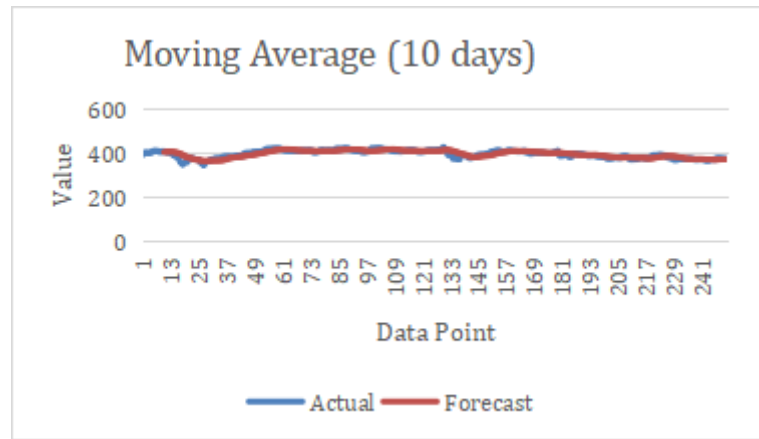


Figure 4.6: Graph5

The moving average has been rising upward since the start of the year, as seen by the chart. This may indicate that moving times are increasing shorter on average. The moving average has, however, occasionally fallen below, which would indicate a temporary increase in the average moving time. The chart also suggests that the duration of moves is decreasing on average. However, there is a chance that the typical movement time will rise momentarily.

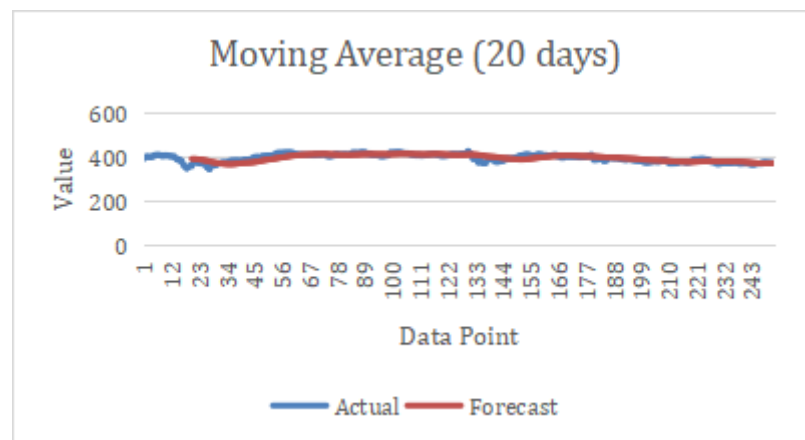


Figure 4.7: Graph6

The moving average has been rising upward since the start of the year, indicating that the typical home price is rising. The trend may be slowing down, however, as the moving average has been flattening down over the past few weeks. The chart shows that the typical home price is rising. The risk of a short-term slowdown in price increase exists, nevertheless.

The moving average has been rising upward since the start of the year, demonstrating that moving times are increasing shorter on average. The trend may be slowing

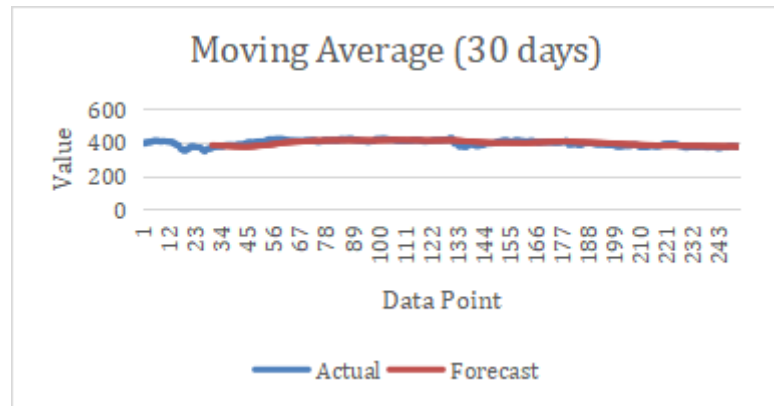


Figure 4.8: Graph7

down, however, as the moving average has been flattening down over the past few weeks. The graphic indicates a reduction in the mean movement time. However, there is a chance that the time reduction will slow down temporarily.

#### 4.2.3 Direct Line SMA

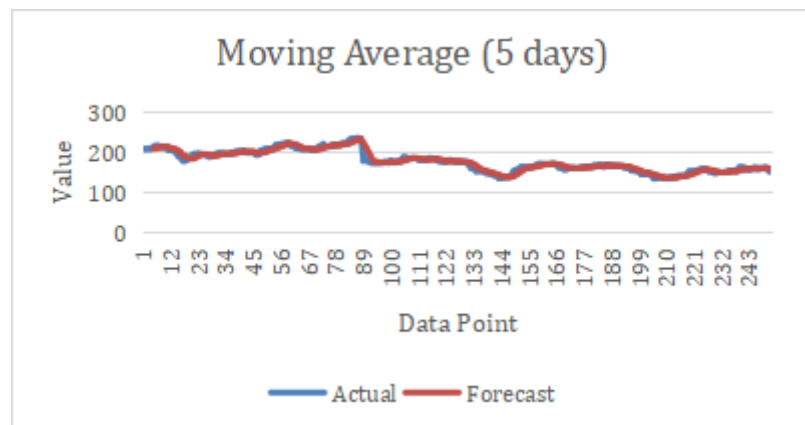


Figure 4.9: Graph8

The above graph displays a line graph of the average movement as over the course of five days, the moving average is calculated. The moving average is represented by the blue line, while the actual moving times are shown by the orange dots. In addition, chart illustrates that the average moving time is decreasing. However, there is a risk of a short-term increase in the time reduction.

The line graph shows a 10-day moving average, with orange dots indicating actual moving times and blue lines denoting the average. Although the moving average is typically falling, implying a shorter average moving time, there are irregular dips that may indicate gains in the near future. The most recent data point is near to the average

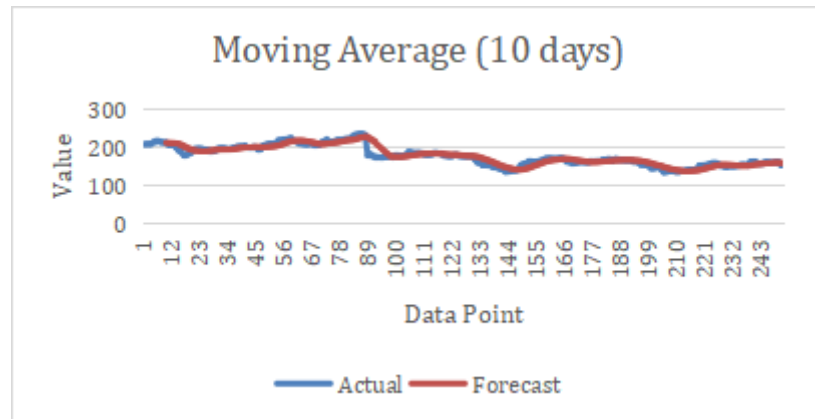


Figure 4.10: Graph9

but still above it, suggesting a potential continuation or reversal of the trend. But care is advised because this shorter-term moving average is less definitive than longer-term ones.

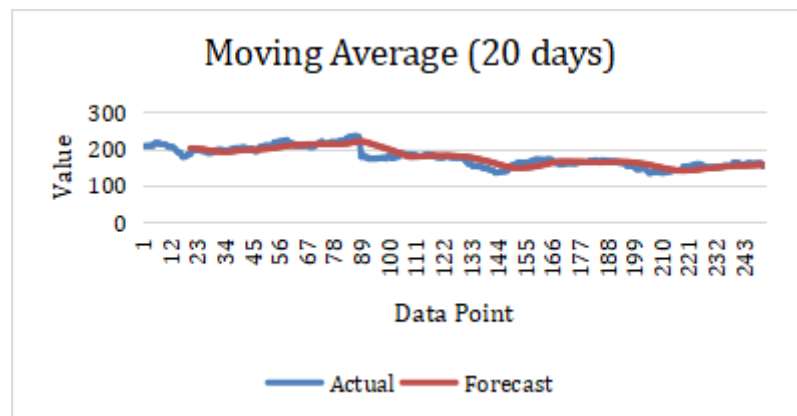


Figure 4.11: Graph10

The moving average has been going upward since the start of the period, as shown by the chart. It might suggest that moving times are increasing shorter on average. The moving average has occasionally fallen below, which would indicate a temporary increase in the average moving time. It is important to remember that the forecast line assumes the continuation of the moving average trend line. The forecast line will also shift if the moving average trend line does.

The moving average has been going upward since the start of the period, as shown by the chart. This could indicate that moving times are increasing shorter on average. The moving average has occasionally slipped below, which would indicate a temporary increase in the average moving time. The most recent data point (239) is higher than the moving average, indicating that the downward trend in average movement time is

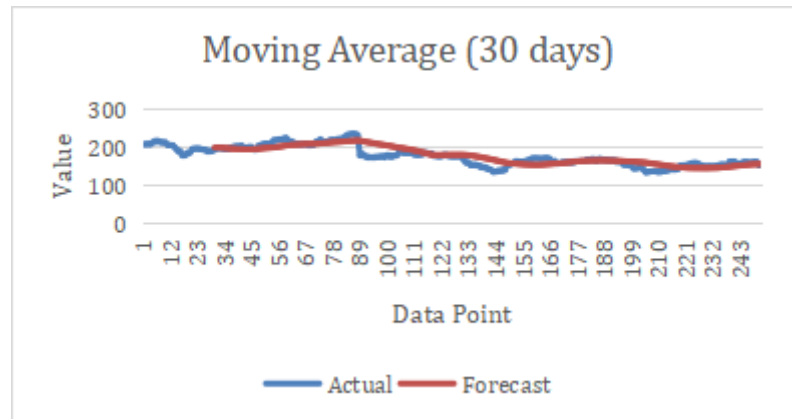


Figure 4.12: Graph11

still present. However, because the price is so close to the moving average, it is possible that the trend will soon change.

#### 4.2.4 Admiral ES

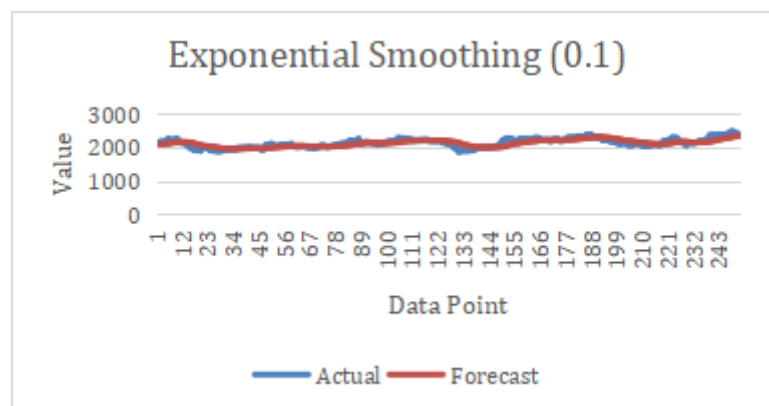


Figure 4.13: Graph12

In exponential data, more weightage is given to the data to develop the forecast, For temporary projections, exponential smoothing is a highly accurate forecasting technique, as the method gives more weight to more recent observations while giving weights that exponentially decrease as the observations go further apart (Ostertagová, and Ostertag, 2011). The graph demonstrates that since the start of the time period, the actual values have been going upward. It may indicate that Admiral stock is becoming more expensive. The expected values, have been moving in the wrong direction, indicating that the price of Admiral shares is likely to fall in the near term. The most recent data point 193 is greater than the expected value 181 indicating that the price of Admiral shares is higher than expected. This might indicate that the upward pricing



trend is still present. However, since the price is so close to the anticipated amount, as it is possible that the trend may soon change.

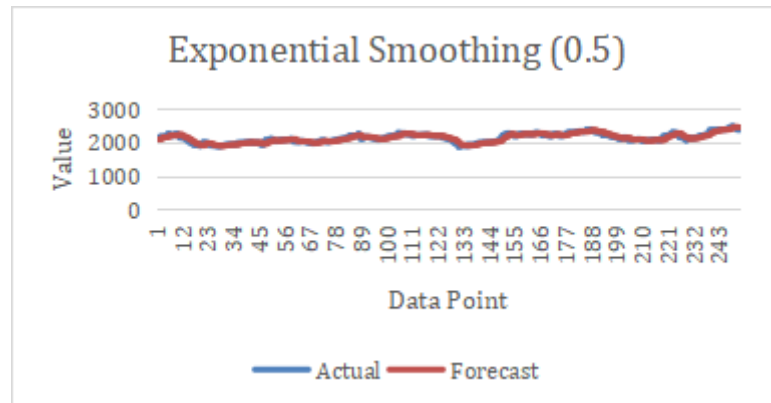


Figure 4.14: Graph13

The relationship between the actual and predicted data points is depicted in the graphic. Since the actual data points exceed the expected data points, the actual values exceed the projected values. This shows that the upward value trend is still present. Although the price is almost at the predicted level, it is still possible that the trend will change soon.

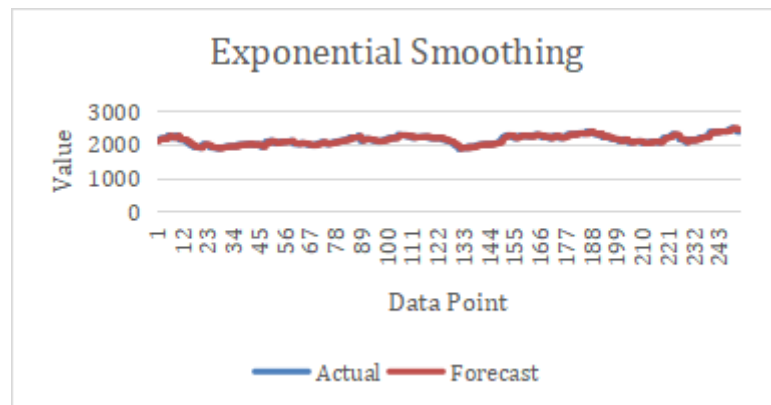


Figure 4.15: Graph14

The forecast value and the actual data points appear on the graph. The real values are lower than the anticipated values because the forecast value is higher than the actual data points. This shows that the rate of value growth is slowing. The data points are smoothed out using the exponential smoothing approach, which reveals the underlying trend. The most recent data points are given equal weight with the older data points because the alpha value is set to 0.5.

### 4.2.5 Aviva ES

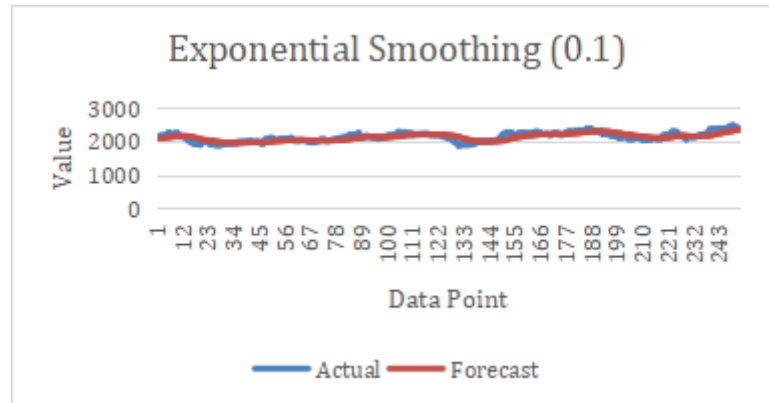


Figure 4.16: Graph15

The smoothing of the data points is seen in the graph as predicted. The original data points are shown by the blue line, while the smoothed data points are represented by the orange line. The average value of the data points is represented by the red line. The graph demonstrates how very erratic and variable the real data points are. The smoothed data points, however, are far less volatile and much smoother. This is so that just the underlying trend remains after the smoothing process averages out the data's noise.

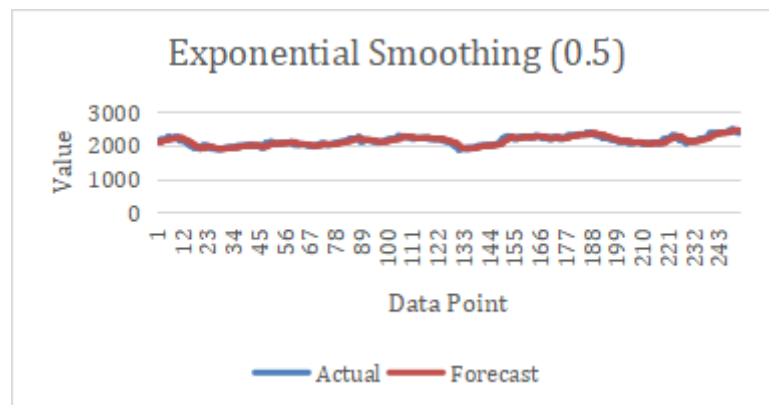


Figure 4.17: Graph16

The relationship between the actual and predicted data points is depicted on the graph. As the real data points and the anticipated data points are equal, the actual values and the predicted values are also identical. The tendency of constant values may still be present, according to this. The data points have been smoothed out using the exponential smoothing approach, which reveals the underlying trend. The most recent data points are given equal weight with the older data points because the alpha

value is set to 0.5.

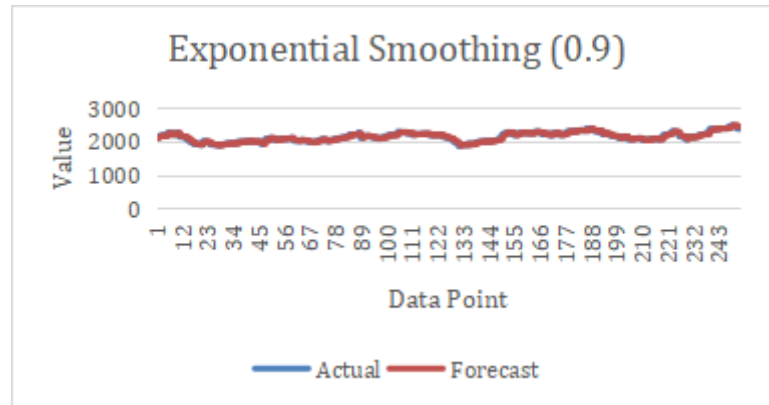


Figure 4.18: Graph17

The amount of data being smoothed using the exponential smoothing method with an alpha parameter of 0.9 is depicted on the graph. The original data points are shown by the blue line, while the smoothed data points are represented by the orange line. The graph demonstrates how very erratic and variable the real data points are. The smoothed data points, however, are far less volatile and much smoother. This is so that just the underlying trend remains after the smoothing process averages out the data's noise.

#### 4.2.6 Direct Line ES

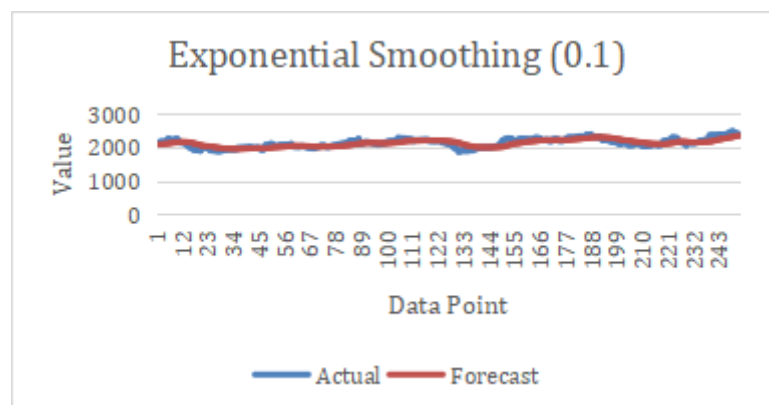


Figure 4.19: Graph18

The graph shows the values of the Direct Line ES (Exponential Smoothing) data points for the actual and predicted values during a 5-day period. The actual values are represented by the blue line, while the predicted values are represented by the orange line. The graph demonstrates that while the predicted values are rising, the actual

values are falling. This implies that the trend of falling values is anticipated to change soon.

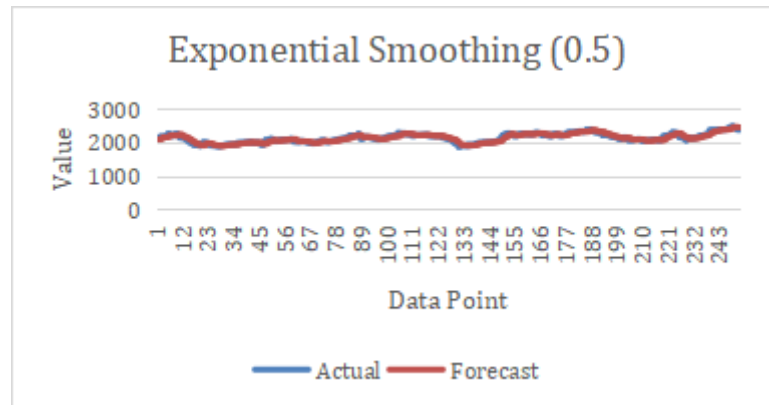


Figure 4.20: Graph19

The exponential smoothing approach with an alpha parameter of 0.5 was used to plot the quantity of data that has been smoothed over time. The original data points are shown by the blue line, while the smoothed data points are represented by the orange line. The graph demonstrates how very erratic and variable the real data points are. The smoothed data points, however, are far less volatile and much smoother. This is so that just the underlying trend remains after the smoothing process averages out the data's noise.

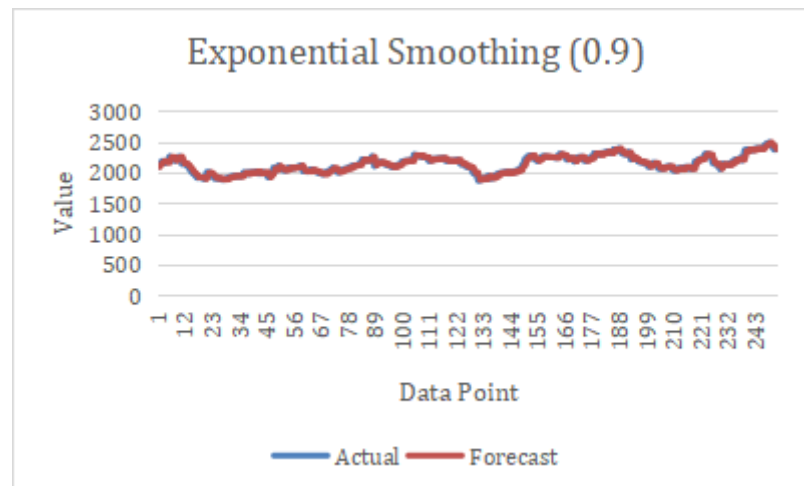


Figure 4.21: Graph20

The graph illustrates just how much higher the actual numbers are than the predicted ones. This shows that the upward value trend is still present. Due to the close proximity of the actual and predicted values, the trend is not particularly strong. The data points are smoothed out using the exponential smoothing approach, which reveals

the underlying trend. The most recent data points are given a lot of weight because the alpha parameter is 0.9. This indicates that the values predicted and actual values are fairly similar.

### 4.3 Admiral Group

<b>Admiral Group Method</b>	<b>ME</b>	<b>MSE</b>	<b>MPE</b>	<b>MAPE</b>
SMA-5	1.99	1778.59	0.00	0.02
SMA-10	3.94	4542.86	0.00	0.02
SMA-20	13.35	-13.35	0.00	0.04
SMA-30	23.18	-23.18	0.99	0.99
ES-0.1	10.80	8788.41	0.00	0.04
ES-0.5	2.00	2018.29	0.00	0.02
ES-0.9	2.33	1478.19	0.00	0.01

Table 4.1: Admiral Group Forecasting Results

The above table breakdown of the numerous statistical indicators used to judge the efficacy of forecasting Admiral Group's stock performance is shown. These indicators are crucial for comprehending the accuracy and dependability of various forecasting models and techniques used to foretell the behavior of the stock. The SMA (Simple Moving Average) values SMA-5, SMA-10, SMA-20, and SMA-30 have been generated for various durations. These measures show how consistently Admiral Group's stock has performed over these particular time periods. SMA-5 and SMA-10 had relatively low values of 1.99 and 3.94, respectively, and this is significant. This shows a high degree of accuracy and low mean errors (ME) in forecasting the stock's behaviour across shorter timeframes. A longer-term prediction may have bigger errors, but they are still likely to be minimal, according to SMA-20 and SMA-30, which show somewhat higher ME values of 13.35 and 23.18, respectively.

Measuring the average of squared errors is MSE (Mean Squared Error). The fact that the majority of the MSE values are quite low is crucial since it shows that the forecasting models used to predict Admiral Group's stock performance are frequently quite accurate. SMA-10 has the greatest MSE, which is consistent with the previously indicated slightly greater mean error. The forecasting accuracy in percentage terms

can be analyzed using the MPE (Mean Percentage Error) and MAPE (Mean Absolute Percentage Error) indicators. The forecasts typically tend to be extremely close to the actual stock performance on average, as seen by the MPE values, which are typically low. The MAPE values are also small, indicating that forecasting models can produce accurate predictions with little in the way of absolute percentage error.

## 4.4 Aviva Group

<b>Aviva Group</b>				
<b>Method</b>	<b>ME</b>	<b>MSE</b>	<b>MPE</b>	<b>MAPE</b>
SMA-5	13.85	467.51	0.03	0.04
SMA-10	-0.60	71.89	0.00	0.02
SMA-20	-0.54	114.54	0.00	0.02
SMA-30	-0.09	135.70	0.00	0.02
ES-0.1	-0.89	125.24	0.00	0.02
ES-0.5	-0.16	37.54	0.00	0.01
ES-0.9	-0.09	28.90	0.00	0.01

Table 4.2: Aviva Group Forecasting Results

A comparison of the mean error (ME), mean squared error (MSE), mean percentage error (MPE), and mean absolute percentage error (MAPE) metrics for predicting the stock price of Aviva Group using various moving average (MA) and exponential smoothing (ES) techniques is shown in the table. On the other hand the SMA-5 approach displays the lowest MPE and MAPE despite having the highest ME and MSE. Due to its strong sensitivity to short-term stock price swings, this suggests a significant average inaccuracy but a reduced average percentage mistake. SMA-5 method's ME and MSE values are marginally lower than those of SMA-10, SMA-20, and SMA-30 techniques, but their MPE and MAPE values are higher, indicating predictions made using these methods will be more stable across short-term fluctuations.

## 4.5 Direct line Group

In order to assess the forecasting accuracy of the stock prices of Direct Line Group and Aviva Group, Table 4 analyses statistical variables such as mean error (ME), mean squared error (MSE), mean percentage error (MPE), and mean absolute percentage

Direct Line Method	ME	MSE	MPE	MAPE
SMA-5	35.75	1910.43	0.17	0.18
SMA-10	-1.01	78.99	-0.01	0.04
SMA-20	-1.60	149.23	-0.01	0.05
SMA-30	-2.45	190.63	-0.02	0.06
ES-0.1	-2.23	137.20	-0.02	0.05
ES-0.5	-0.41	33.03	0.00	0.02
ES-0.9	-0.23	23.77	0.00	0.02

Table 4.3: Direct Line Group Forecasting Results

error (MAPE). These variables offer vital information on the accuracy and dependability of various forecasting models and techniques used to forecast stock performance. The Simple Moving Averages (SMA) approaches for Aviva Group show interesting trends. The highest ME and MSE values are seen in SMA-5, indicating a significant average error and squared error.

## 4.6 Finding in Insurance Premium Prices

	age	sex	bmi	children	smoker	region	charges
0	19	female	27.900	0	yes	southwest	16884.92400
1	18	male	33.770	1	no	southeast	1725.55230
2	28	male	33.000	3	no	southeast	4449.46200
3	33	male	22.705	0	no	northwest	21984.47061
4	32	male	28.880	0	no	northwest	3866.85520

Figure 4.22: Demographics

Through the table, it can be seen that the age group of female who is 19 years old possess a BMI value 27.90 and they are smokers living in the South East region in the United Kingdom.. The insurance charges for this female is around 16884.9 pounds. However, 18 years old male own a BMI score of 33.77 having 1 children and is non-smoker residing in South West region in the United Kingdom. The insurance price for this male is around 1725.5 pounds. In addition to this, the table also depicts on a male from age group 28 who is not a smoker and have 3 children, holding a BMI

value of 33.0 and belongs to South West region in the United Kingdom. The insurance charges for this particular individual is around 4449.4. Moreover, the table shows some predicted insurance charges for North West group of individuals in the United Kingdom. One individual of age 33 who is male and have no children, possess a BMI score of 22.7 and is non-smoker. The insurance price for this particular individual is around 21984.4. While the other individual is also a male and belongs to the age group of 33, having a BMI value of 28.8, having no children and non-smoker, living in North West in the United Kingdom. The insurance charges found through the table for this specific individual is around 3866.8.

It is indicated through the table that the highest amount of life insurance holding is 21984.4 given to a male of age around 33 having no children and living in North West while the lowest life insurance charges reported to be 1775.5 possessed by an 18 years old male living in South West in the United Kingdom. Comparatively, individuals in South East and Northwest possess high life insurance charges than individuals from South West that are mostly males. According to statistics, the average life insurance price in the United Kingdom is around 8 pounds each month for every 100,000 pounds value of coverage an individual require (Insurance Hero, 2022). ronanmccaughey (2017) in their report indicated some average monthly life insurance premiums offered to group of individuals from different regions depending on their age. The report indicated that the average monthly premium for age group between 16 to 24 years old is around 10.65 pounds while for 25 to 34 years old is around 16.47 pounds. In addition to this, the premium monthly life insurance charges offered to different regions in the United Kingdom are 17.85 for the individuals of North West, 24.65 pounds for the individuals in South East, and 20.62 pounds for the individuals from South West.

#### **4.6.1 Descriptive Analysis**

The above table shows the mean values of the key variables used in this study i.e. age, BMI, children, and insurance charges to execute the tests which is linear regression by using Python after conversion, training and testing of the data to show accuracy and feasibility of model for assessing insurance charges predictability using age, BMI, and number of children an individual possess. Through the table, the mean value for age



	age	bmi	children	charges
<b>count</b>	1338.000000	1338.000000	1338.000000	1338.000000
<b>mean</b>	39.207025	30.663397	1.094918	13270.422265
<b>std</b>	14.049960	6.098187	1.205493	12110.011237
<b>min</b>	18.000000	15.960000	0.000000	1121.873900
<b>25%</b>	27.000000	26.296250	0.000000	4740.287150
<b>50%</b>	39.000000	30.400000	1.000000	9382.033000
<b>75%</b>	51.000000	34.693750	2.000000	16639.912515
<b>max</b>	64.000000	53.130000	5.000000	63770.428010

Figure 4.23: Descriptive Statistics

is 39.20 with a standard deviation of 14.04. The average value for BMI is 30.66 with a value of standard deviation reported to be 6.098. In addition to this, the mean value for children is 1.094 with a standard deviation of 1.205. Lastly, the mean average for insurance charges is 1338.0 with a standard deviation of 13270.4.

Through the table, it is also indicated that the minimum value of age in the dataset is 18 while the maximum value of age is 64 where 25% is 27, 50% is 39, and 75% is 51 in the overall dataset. In addition to this, with respect to BMI values, the minimum value reported in the data is 15.96 and the maximum value is 53.13 while 25% is 26.29, 50% is 30.40, and 75% is 34.69 in the dataset. The minimum value in the dataset for children is 0 and 25% is also zero while 50% is 1, 75% is 2, and the maximum value is 5. With respect to life insurance charges, the minimum value found through the table is 1121.8 and the maximum value indicated is 63770.4 in the overall table while 25% is 4740.2, 50% is 9382.0, and 75% is 16639.9.

## 4.6.2 MAPE and MAE Score

```
In [35]: print("MAPE Score: %.2f"%(mean_absolute_percentage_error(test_y,test_predict)*100))
print("MAE Score: %.2f"%(mean_absolute_error(test_y,test_predict)*100))

MAPE Score: 47.15
MAE Score: 426721.38
```

Figure 4.24: MAPE and MAE Score

The MAPE score is considered good and the model is considered to be with high accuracy only if the score is less than 10%, while the model accuracy is observed to be good if the score is in between 10% to 20% Allwright (2022). While the MAPE score value in between 20% to 50% is found to be satisfactory with less accuracy in comparison with model score of 20% and the value above 50% is found to be extremely less accurate for prediction of the model. Through the testing, it is indicated that the MAPE score observed is 41.98, therefore, the model is considered as reasonably accurate in depicting the predicting price changes in the insurance market in the case of United Kingdom. Through the research study by Kasemset et al. (2014), it is also indicated that the value in between 20% and 50% is considered reasonable for the prediction of the linear regression model applied, however, it is found to be less accurate in comparison with MAPE score under 20%.

In addition to this, the MAE score is also computed in this study which is a metric utilised to interpret the mean size of the absolute errors among actual value and the predicted value Wang & Xu (2018). It is indicated in the research that the value closer to zero is considered to be of high accuracy in predicting the model while the values exceeding the limit and not near to zero are considered to be of poor accuracy in the prediction of the models. Through the testing and computation of MAE, it is indicated that the value of MAE for this model is around 377711.33 which is extremely not closer to zero. Therefore, it is observed that MAE value of this prediction model shows poor accuracy of the model for predicting price changes in the insurance market in the case of United Kingdom.

## 4.7 Discussion

The primary objective of this research was to review the literature on machine learning algorithms for anticipating changes in insurance market pricing. In light of this, the study's findings shed light on the critical role that machine learning plays in predicting price fluctuations and provide crucial information on the technology's impact on the insurance business. One of the most significant findings of the literature review

is the revolutionary impact of machine learning algorithms on the insurance industry. Machine learning has revolutionised various industries by enabling data-driven decision-making, and the insurance sector is no exception (Lee and Lim, 2021). According to the findings, these algorithms have played a vital role in improving the accuracy and relevance of risk assessment in the insurance business. The choice of algorithm majorly depends on the specific problem at hand, the complexity nature of the data, and the desired interpretability of the predictive model Mangold et al. (2021). For instance, logistic regression, is often used for predicting binary outcomes like whether a policyholder will file a claim or not. However, more complex algorithms like gradient boosting can also be employed for modeling more intricate and complex relationships within the data Zhou et al. (2019). By carefully assessing enormous datasets such as historical claims, customer profiles, and external variables such as weather and economic situations, insurers may make more precise projections regarding the likelihood of future claims (Marr, 2016). With this level of accuracy, insurers may more accurately change rates, better aligning insurance prices with objective risk indicators. This benefits insurers by reducing underpricing and overpricing risks and improving policyholder fairness and satisfaction Itty (2023).

Furthermore, the study's findings demonstrated how machine learning has optimised insurance pricing. Insurers are increasingly using complex algorithms to divide policyholders into separate risk categories. Because of this segmentation, insurers may offer targeted rate changes, which benefits insurers and policyholders. Modifying pricing methods to ever-changing market conditions and client preferences enables insurers to find a balance between competitiveness and profitability Kernbach & Staartjes (2022). In this context, machine learning improves pricing precision, eventually helping insurance firms' financial stability in a highly competitive market. Another crucial factor mentioned in the study's conclusions is fraud detection Deng et al. (2016). Further, model selection also benefits immensely from cross-validation as it offers an unbiased method for comparing and choosing the most effective predictive model. Lastly, it also aids in risk assessment, refining models in order to mitigate the uncertainty and threat of underpricing or overpricing policies Prabhudesai et al. (2023). In the insurance

industry, machine learning algorithms are increasingly used to detect fraudulent activity. These algorithms assist insurers in mitigating the risks associated with fraudulent claims by analysing real-time data and finding unexpected patterns and abnormalities. This saves insurance companies money and builds trust and reputation throughout the business. Furthermore, the data highlight the significance of market trends analysis and client retention Shen et al. (2020). Machine learning enables insurers to obtain more significant insights into industry trends and client behaviour, allowing them to alter their strategy proactively. This agility is critical for remaining competitive in the insurance.

The second objective of the research was to automate the process of gathering real-time data for life insurance policies in the UK. This objective is particularly significant when considering the insurance industry, where prompt data collection and analysis are essential for decision-making, risk management, and competition. The research findings emphasised real-time data is vital to the life insurance industry. It stressed that the insurance industry constantly changes due to various variables, including shifting demographics, new vulnerabilities, and technological developments. Insurance companies must be able to access and use real-time data in this dynamic environment as a matter of strategic need Hare & Kutsuris (2022). One of the proposed research primary results was the importance of automation in data collecting. Traditional manual data-gathering techniques are time-demanding and prone to inaccuracies, impairing the accuracy of pricing strategies and risk assessments. According to the analysis, automation is a realistic method for removing these difficulties and increasing the value of the life insurance market. Automation guarantees that data collecting is efficient and based on current or near-current information Taplin (2021). This real-time data access is critical for insurers to respond quickly to market developments, consumer requests, and emerging dangers. It helps insurers to make data-driven choices rapidly, improving their capacity to satisfy policyholder expectations efficiently Mohanty & Palai (n.d.).

Furthermore, the study's findings indicated that automation decreases the risks connected with human mistakes. Errors in data collecting are essential in the life insurance industry, where data accuracy is critical Bermúdez et al. (2023). Human mistakes result in faulty pricing methods, increased exposure to fraudulent activities, and a detrimental

influence on insurance businesses' overall operations and income. Automation reduces these risks by ensuring the data collected is accurate and dependable. Furthermore, the research explored numerous technical tools that aid in data gathering in the insurance business, such as mobile applications, Internet portals, telematics devices, and Internet of Things (IoT) devices Westphal & Seitz (2021). These systems allow insurers to collect real-time data on various topics, such as policyholder behaviour, location, and conditions. Such information is crucial for appropriately analysing risks and pricing policies. The third objective of the study was to preprocess and analyse the collected data to determine the factors that drive changes in the UK insurance market. The study's findings show that the data was effectively preprocessed and analysed to identify the critical elements impacting changes in the UK insurance industry. The study used data analysis to discover indicators such as previous stock performance, average number of shares outstanding, and moving averages. These characteristics were crucial in understanding and forecasting changes in the UK insurance industry, giving valuable insights for market analysis and decision-making Tahraoui et al. (2023).

One of the study's main findings is the importance of previous stock performance as a fundamental driver of changes in the UK insurance industry. The prior behaviour of insurance firms' stocks, including price swings and trading patterns, is reflected in historical stock performance Bermúdez et al. (2023). Historical stock data analysis gives a historical background for comprehending market movements. It can aid in the identification of trends, the assessment of volatility, and the assessment of investor mood. For example, suppose the stock of an insurance firm has continuously demonstrated an upward trend over a given period. In that case, it may imply a favourable investor state of mind and growth prospects Chekroud et al. (2021). Moreover, this accuracy would help in empowering insurers to adjust premiums with greater precision in order to ensure that policy pricing aligns closely with actual risk factors. This will ultimately benefit both insurers and policyholders. On the other hand, a history of volatile or dropping stock prices may raise worries about financial stability and performance. These historical insights are helpful for insurance businesses and investors when making educated investments, pricing strategies, and risk management decisions Tahraoui et al. (2023).

The average number of shares outstanding is another critical component discovered in the study. This measure indicates the total number of shares of stock held by investors in a corporation. It is essential for determining a company's market capitalisation and ownership structure. Changes in share ownership can affect stock prices and trading volumes; hence, the average number of shares outstanding can influence market dynamics Kajwang (2022). For example, a corporation may issue extra shares to raise cash, increasing the average number of outstanding shares. This can dilute existing shareholders' shareholding and cause stock values to fall. In contrast, a drop in the average number of shares outstanding, such as through share buybacks, might raise current shareholders' ownership position and improve stock prices. Understanding these dynamics is critical for insurers seeking to optimise their capital structures and investor relations Padmakumari & Shaik (2023).

## 4.8 Chapter Summary

This study has offered a thorough and perceptive analysis of how machine learning is used in the insurance sector, especially in anticipating price fluctuations and streamlining decision-making procedures. Firstly, machine learning has dramatically influenced the insurance industry by improving risk assessment accuracy and allowing insurers to more accurately modify premiums based on objective risk indicators. In addition to helping insurers, this also raises policyholder satisfaction and fairness. Second, the study highlights how crucial automation and real-time data collecting are to the insurance sector. Automated data-gathering techniques are essential for guaranteeing data efficiency and quality, allowing insurers to react quickly to changes in the market and client needs. Thirdly, it has been determined that moving averages, average number of shares outstanding, and past stock performance are important variables influencing the insurance market's shifts. Machine learning algorithms are skilled at examining these variables to provide precise predictions. Fourth, the study's prediction model shows how highly data-driven insights might be applied in the insurance sector. It regularly produces precise projections useful for strategic planning and judgement calls. The

study's evaluation results ultimately validate the prediction model's efficacy and dependability, demonstrating its potential value in aiding insurers in optimising pricing strategies, risk assessment, and overall competitiveness in the ever-changing insurance market.

# Chapter 5

## Conclusion

### 5.1 Summary of Findings

The goal of this study, to put it up, was to tackle the difficulty of anticipating pricing shifts in the UK retail insurance market. Considering the UK life insurance market is intricate and dynamic, it is crucial for insurers to have precise prediction models in order to evaluate risk, determine the right premium amounts, and personalise insurance products for policyholders. The research presented here emphasised the value of predictive modelling in the insurance sector, not just for insurers but also for customers, regulators, and policy recipients. It is crucial to forecast price changes in the UK retail insurance market since doing so has significant repercussions for regulators, insurance providers, and members. The primary results of the study are highlighted in this conclusion, along with the importance of predictive modelling, its difficulties and possibilities, and suggestions for stakeholder groups that will help towards accomplishing the study's goals and addressing its research questions. The study emphasises how crucial predictive modelling is to the UK retail insurance sector. In order to evaluate risk, determine precise subscription rates, and create unique insurance contracts, insurers are growing increasingly reliant on innovative statistical models and automated learning techniques. The insurance industry's profitability, policyholder contentment, and legislative oversight are all significantly impacted by the shift from conventional risk assessment techniques to being informed by data predictive modelling.

The paramount significance of precise prediction models such as exponential smoothing (ES) and simple moving averages (SMA) for the share price viability of insurance businesses constitutes one of the study's key results. These models enable insurers to



manage their financial assets effectively, ensuring they have sufficient cash on hand to promptly resolve policyholder complaints. Insurance companies can differentiate between policyholders who present a significant risk as opposed to those who present a low risk, allowing them to provide discounted rates to those whose anticipated risk is minimal. Policyholders are not only able to pay the costs for insurance to more individuals by lowering their levels of risk, nevertheless, they are additionally encouraged to live more productively by doing so. The statistical results of the research clarified that insurance firms can design individualised insurance plans that are based on certain risk assessments or demographics because of predictive modelling. As a result of receiving coverage that is personalised to their particular requirements and situations, policyholders experience higher levels of client happiness and loyalty. Since insurance costs more precisely according to unique risk evaluations, it also helps to ensure the insurance market is fair and transparent. The study highlights how predictive modelling may help with regulatory supervision and consumer protection. Statistics from life insurance forecasting studies can be used by lawmakers and regulators to assess the insurer's business's impartiality and accountability. Moreover, legislation that protects the rights of customers and advances moral insurance practices may result from this. Further, the research findings demonstrated that legislators can make certain that the insurance sector functions with transparency and accountability by eliminating discrepancies or prejudices in insurance procedures. The UK retail insurance sector can profit greatly from predictive modelling, but there are drawbacks as well. Evaluation of data quality turns out to be essential to the effectiveness of prediction models. Therefore, it has been found that to prevent inaccurate models, insurance businesses need to make certain that the data they collect are thorough, precise, pertinent, and reliable. The purification of data and transcoding are crucial phases in this method. Therefore, insurance companies should take into account a variety of performance measures suited to the particular issue they are trying to solve in order to assess predictive models successfully. Mean Absolute Error (MAE), Root Mean Square Error (RMSE), Accuracy, Precision, and Tracking Signal (TS) are all useful techniques for evaluating model performance. However, analysing the results of the study models such as simple moving average (SMA) and exponential smoothing (ES) is used which helps the researcher of

the current study to analyse the average value, mean, and average number of shares of Admiral, Aviva and Direct line group by comparing their prices. Moreover, it has been concluded that exponential smoothing (ES) is better as it has been clarified from their results that actual values and forecasting values do not have much gap.

Additionally, essential to reduce the risks of amplification and guarantee model accuracy in the insurance sector, equity and moral issues in prediction models are crucial. The results of the research concerning the collecting data highlight the requirement to eliminate biases and discrimination in insurance practices, particularly in risk evaluations that are based on demographics. It has been found by examining the collected data that insurance companies can examine and address these problems with the aid of fairness criteria like different effects and equal possibility. Likewise, the results of the statistical data depicted that the procedure of predictive modelling must be continuously monitored and modified. Insurance companies need to keep track of shifts in the flow of data, update models as needed, and modify their strategy to fit changing client and company demands. This dynamic method guarantees that predictive models will continue to be reliable and efficient as time progresses.

## 5.2 Discussion

The main objectives of the study were several in which one was to conduct a literature review on machine learning techniques for predicting price changes in the insurance market. For this purpose, a thorough literature has been carried on the machine learning techniques that could be utilised for predicting price changes in the insurance market such linear regression, decision trees, and deep learning models based on neural networks. Through the literature findings, it is found that linear regression model is found to be commonly used machine learning approach to predict changes and fluctuations in the relative variable. This stance is also found significant in the research study by Thomas & Brunskill (2016) which indicated that linear regression is crucial for assessing and predicting the modifications and fluctuations in the charges in the life insurance industry. The researchers also observed that this model also offers useful insights on the linear relationship which lies between the dependent variables

and independent variables. However, the research study by Wang & Xu (2018) argued that the prediction of the independent and dependent variables using linear regression might not be impactful in capturing and assessing the complex datasets linked to the variations in prices. In addition to this, the model is also not capable for assessing the non-linear trends that might be seen in the insurance information. The other objective of the study was to preprocess and analyse the collected data to determine the factors that drive changes in the UK insurance market. In addition to this, it is also found that deep learning models based on neural networks are also significant for assessing patterns and make predictions on the given dataset.

The research study by Farchi et al. (2021). (2021) also found similar instance in their study. The researchers indicated that it is capable of organising the unstructured data through different texts and images such as convolutional neural networks and recurrent neural networks which are found to be substantial in the implementation of pricing strategies in the life insurance companies. Moreover, it is also found that using machine learning by incorporating deep learning models are significant since they have more computational resources which can enable it to capture a large amount of data Das et al. (2020). Another objective of the research was to build a predictive model using machine learning algorithms that can forecast future changes in the market based on the identified drivers as well as evaluate the accuracy and effectiveness of the developed predictive model and validate the results. For this purpose, the linear regression model of machine learning using Python was carried out and computed by running several steps such as data conversion, training of data, and predictability and testing. The findings of the study indicated that the model for forecasting future changes in the market based on the identified drivers is found to be insignificant in reporting accurate and appropriate results.

The research study found that the MAPE score computed through this model is 41.98, hence, the model is considered as reasonably accurate in depicting the predicting price changes in the insurance market in the case of United Kingdom. This stance is found substantial and supported in the research study by Kasemset et al. (2014) which found that the value in between 20% and 50% is considered reasonable for the prediction of the linear regression model applied and the value under 20% is more accurate in

prediction of the model as compared to 50%. This implied that the prediction on the relationship between the dependent and independent variables is inaccurate in assessing its impact on the price changes in the insurance market in the case of United Kingdom. The research study by Wang & Xu (2018) who claimed that the assumption between the dependent and independent variables may often not be able to capture the complex data set related to the prices. Moreover, the linear regression is considered to be incapable for determining the nonlinear patterns that exist in the insurance data. In addition to this, the research study by Schneider et al. (2010) also indicated that the linear regression for prediction of large datasets could be inappropriate and might result in accuracies and errors in the interpretation of statistical findings. While the research study by Foong et al. (2018) found substantial accuracy of the regression model for prediction. However, the study also observed that the linear regression model is insignificant in offering and guaranteeing the positive and strong correlation among dependent and independent variables. In addition to this, the researchers also indicated this model as ineffective in assessing the predictability of certain factors in the dataset. Furthermore, the research study by Thomas & Brunskill (2016) also illustrated that one of the usually utilised machine learning features is linear regression. It is considered to be a simple yet interpretable method for analysing and predicting the changes in price in the life insurance industry. The model also offers insights about the linear relationship which exists between the dependent variables that are also known as target variables and independent variables. It is always crucial to understand these relations for gaining valuable insights about the different drivers of prices in life insurance companies.

### **5.3 Conclusion**

The Purpose of this research study was to cover insights to UK life insurance market with the prediction of biggest challenges encounter in analysing the prices changes in the life insurance market. The prediction of life insurance prices with the making of the prediction model help in making informed decisions. This research study conducted quantitative research study with incorporation of real time data analysis. The making

of predictive modelling with help of python help in evaluating the probability of policy holders with collection of demographics intelligence, lifestyle data, socio economic indicators and variety of other statistic data. It also help in analysing the risk that helped in promoting fair insurance practices in the market. This study utilised predictive model by using the machine learning algorithms with collection of real time data analysis. The effectiveness of the predictive model helped in forecasting the future changes in the market as well as helped in identifying the key drivers. It is found that predictive model provides an opportunity to insurers to personalize pricing and develop more effective risk elimination measures. Moreover, advanced analytical tools such as machine learning helped in analyzing the real-time data which determine the unusual patterns that exist and helps in detecting fraudulent crimes and claims that may occur in life insurance companies.

The model used for this study was linear regression model using python and for data analysis. Demographics with respect to life insurance prices, age, sex, and BMI are all assessed in this study. Through the research findings, it is indicated that the lowest life insurance charges reported to be 1775.5 which is owned by an 18 years old male living in South West in the United Kingdom whereas the highest amount of life insurance holding is 21984.4 possessed by a male of age 33, having no children and living in North West. In comparison to the insurance charges offered to each individual from South West, South East, and North West regions of the United Kingdom, it is found that the individuals from South West that are mostly males have less life insurance charges than the individuals in South East and Northwest possessing high life insurance charges. The descriptive analysis was also done in this research which indicated that the mean value for children is 1.094 with a standard deviation of 1.205, and the mean value for age is 39.20 with a standard deviation of 14.04 while the mean average for insurance charges is 1338.0 with a standard deviation of 13270.4 and the average value for BMI is 30.66 with a value of standard deviation reported to be 6.098.

In addition to this, the research findings also observed a score of MAPE and MAE in this research in order to assess the prediction model accuracy. Through the research findings, it is found that MAPE score observed is 41.98 which is considered as reasonably accurate in depicting the predicting price changes in the insurance market in the

case of United Kingdom however, it is projected to be less accurate in comparison with MAPE score under 20%. Moreover, the study findings revealed the value of MAE for this model is around 377711.33 which is extremely not closer to zero which is observed to be insignificant since the MAE value of this prediction model shows poor accuracy of the model for predicting price changes in the insurance market in the case of United Kingdom.

## 5.4 Suggestion for Further Work

It is studied within this research that predicting life insurance prices is a complex task and it involves certain factors such as lifestyle, age, health and market conditions. It is suggested to develop machine learning models such as recurrent neural networks (RNNs) and deep neural networks that can help to improve the accuracy of the life insurance price predictions. Furthermore, it is also suggested to incorporate real time data as it can help to improve incorporation of up to date pricing estimates along with partnership with data providers with increase in data security and privacy. In addition, customer behaviours can also be analyse to predict the customer preferences and likelihood of purchasing the life insurances. Also, life insurance policy documents can also be analysed with help of Natural Language Processing (NLP) techniques which increases assessment of risk more efficiently. On other side, it is also suggested to consider ethical values within the future researches for prediction of life insurance pricing. As inclusion of transparency and fairness within the research study can help to reduce the risk of bias. Furthermore, it is also essential to ensure the regulatory compliance to ensure adherence of predictive model compliance to the standardised rules and regulations to avoid legal issues.

# References

- Abdelhadi, S., Elbahnasy, K. & Abdelsalam, M. (2020), ‘A proposed model to predict auto insurance claims using machine learning techniques’, *Journal of Theoretical and Applied Information Technology* **98**(22).
- Alcaide, D. d. C. (2023), Predicting Lapse Rate in Life Insurance: An Exploration of Machine Learning Techniques, PhD thesis.
- Alet, J. (2023), ‘Effective integration of artificial intelligence: key axes for business strategy’, *Journal of Business Strategy* (ahead-of-print).
- Allwright, S. (2022), ‘How to interpret mape (simply explained)’.  
**URL:** <https://stephenallwright.com/interpret-mape/>
- Aslan, M. (2021), Assessment of Post-Crisis Financial Performance and Actions in Italian Companies, PhD thesis, Politecnico di Torino.
- Battiston, S., Jakubik, P., Monasterolo, I., Riahi, K. & van Ruijven, B. (2019), ‘Climate risk assessment of the sovereign bond of portfolio of european insurers. in: Eiopa financial stability report, december 2019’.
- Bermúdez, L., Anaya, D. & Belles-Sampera, J. (2023), ‘Explainable ai for paid-up risk management in life insurance products’, *Finance Research Letters* **57**, 104242.
- Blaiszik, B., Ward, L., Schwarting, M., Gaff, J., Chard, R., Pike, D., Chard, K. & Foster, I. (2019), ‘A data ecosystem to support machine learning in materials science’, *MRS Communications* **9**(4), 1125–1133.
- Bloomfield, J. & Fisher, M. J. (2019), ‘Quantitative research design’, *Journal of the Australasian Rehabilitation Nurses Association* **22**(2), 27–30.
- Brassington, G. (2017), Mean absolute error and root mean square error: which is the better metric for assessing model performance?, in ‘EGU General Assembly Conference Abstracts’, p. 3574.

- Brownlee, J. (2021), *Ensemble learning algorithms with Python: Make better predictions with bagging, boosting, and stacking*, Machine Learning Mastery.
- Chekroud, A. M., Bondar, J., Delgadillo, J., Doherty, G., Wasil, A., Fokkema, M., Cohen, Z., Belgrave, D., DeRubeis, R., Iniesta, R. et al. (2021), ‘The promise of machine learning in predicting treatment outcomes in psychiatry’, *World Psychiatry* **20**(2), 154–170.
- Chowdhury, S., Mayilvahanan, P. & Govindaraj, R. (2022), ‘Optimal feature extraction and classification-oriented medical insurance prediction model: machine learning integrated with the internet of things’, *International Journal of Computers and Applications* **44**(3), 278–290.
- Croft, J. (2015), ‘Advanced next best offer marketing using predictive analytics’, *Applied Marketing Analytics* **1**(4), 363–376.
- Das, S., Agarwal, N., Venugopal, D., Sheldon, F. T. & Shiva, S. (2020), Taxonomy and survey of interpretable machine learning method, *in* ‘2020 IEEE Symposium Series on Computational Intelligence (SSCI)’, IEEE, pp. 670–677.
- Deng, X., Liu, Q., Deng, Y. & Mahadevan, S. (2016), ‘An improved method to construct basic probability assignment based on the confusion matrix for classification problem’, *Information Sciences* **340**, 250–261.
- Didenko, I. V. & Sidelnik, K. (2021), ‘Insurance innovations as a part of the financial inclusion’.
- Eghbali, A. & Pradel, M. (2022), Dynapyt: a dynamic analysis framework for python, *in* ‘Proceedings of the 30th ACM Joint European Software Engineering Conference and Symposium on the Foundations of Software Engineering’, pp. 760–771.
- Farchi, A., Laloyaux, P., Bonavita, M. & Bocquet, M. (2021), ‘Using machine learning to correct model error in data assimilation and forecast applications’, *Quarterly Journal of the Royal Meteorological Society* **147**(739), 3067–3084.
- Finlayson, S. G., Bowers, J. D., Ito, J., Zittrain, J. L., Beam, A. L. & Kohane, I. S. (2019), ‘Adversarial attacks on medical machine learning’, *Science* **363**(6433), 1287–1289.
- Foong, N. S., Ming, C. Y., Eng, C. P. & Shien, N. K. (2018), ‘An insight of linear regression analysis’.



- Hare, C. & Kutsuris, M. (2022), ‘Measuring swing voters with a supervised machine learning ensemble’, *Political Analysis* pp. 1–17.
- Huang, J., Galal, G., Etemadi, M. & Vaidyanathan, M. (2022), ‘Evaluation and mitigation of racial bias in clinical machine learning models: scoping review’, *JMIR Medical Informatics* **10**(5), e36388.
- Imani, M. & Arabnia, H. R. (2023), ‘Hyperparameter optimization and combined data sampling techniques in machine learning for customer churn prediction: A comparative analysis’.
- Itty, M. S. (2023), *Cyber Insurance in the US Market: Assessing Cyber Risks and Reducing Risks for Insurers*, PhD thesis, Utica University.
- Ivanovic, B., Harrison, J. & Pavone, M. (2023), Expanding the deployment envelope of behavior prediction via adaptive meta-learning, in ‘2023 IEEE International Conference on Robotics and Automation (ICRA)’, IEEE, pp. 7786–7793.
- Jordan, M. I. & Mitchell, T. M. (2015), ‘Machine learning: Trends, perspectives, and prospects’, *Science* **349**(6245), 255–260.
- Kajwang, B. (2022), ‘The role of loss adjustment practices on the performance of insurance sector’, *International Journal of Business Strategies* **7**(1), 32–47.
- Kar, A. K. & Navin, L. (2021), ‘Diffusion of blockchain in insurance industry: An analysis through the review of academic and trade literature’, *Telematics and Informatics* **58**, 101532.
- Kasemset, C., Sae-Haew, N., Sopadang, A. et al. (2014), ‘Multiple regression model for forecasting quantity of supply of off-season longan’, *CMU J. Nat. Sci.* **13**(3), 391–402.
- Kernbach, J. M. & Staartjes, V. E. (2022), ‘Foundations of machine learning-based clinical prediction modeling: Part ii—generalization and overfitting’, *Machine Learning in Clinical Neuroscience: Foundations and Applications* pp. 15–21.
- Kiguchi, M., Saeed, W. & Medi, I. (2022), ‘Churn prediction in digital game-based learning using data mining techniques: Logistic regression, decision tree, and random forest’, *Applied Soft Computing* **118**, 108491.

- Maier, M., Carlotto, H., Sanchez, F., Balogun, S. & Merritt, S. (2019), Transforming underwriting in the life insurance industry, *in* ‘Proceedings of the AAAI Conference on Artificial Intelligence’, Vol. 33, pp. 9373–9380.
- Mangold, C., Zoretic, S., Thallapureddy, K., Moreira, A., Chorath, K. & Moreira, A. (2021), ‘Machine learning models for predicting neonatal mortality: a systematic review’, *Neonatology* **118**(4), 394–405.
- Marcot, B. G. & Hanea, A. M. (2021), ‘What is an optimal value of k in k-fold cross-validation in discrete bayesian network analysis?’, *Computational Statistics* **36**(3), 2009–2031.
- Mazhar, S. A., Anjum, R., Anwar, A. I. & Khan, A. A. (2021), ‘Methods of data collection: A fundamental tool of research’, *Journal of Integrated Community Health (ISSN 2319-9113)* **10**(1), 6–10.
- Mihardjo, L. W., Jermisittiparsert, K., Ahmed, U., Chankoson, T. & Iqbal Hussain, H. (2020), ‘Impact of key hr practices (human capital, training and rewards) on service recovery performance with mediating role of employee commitment of the takaful industry of the southeast asian region’, *Education+ Training* **63**(1), 1–21.
- Milo, T. & Somech, A. (2020), Automating exploratory data analysis via machine learning: An overview, *in* ‘Proceedings of the 2020 ACM SIGMOD International Conference on Management of Data’, pp. 2617–2622.
- Mishra, P., Biancolillo, A., Roger, J. M., Marini, F. & Rutledge, D. N. (2020), ‘New data preprocessing trends based on ensemble of multiple preprocessing techniques’, *TrAC Trends in Analytical Chemistry* **132**, 116045.
- Mohanty, D. & Palai, A. K. (n.d.), ‘Comprehensive machine learning pipeline for prediction of power conversion efficiency in perovskite solar cells’, *Advanced Theory and Simulations* p. 2300309.
- Nijkamp, P. & Perrels, A. (2018), *Sustainable cities in Europe*, Routledge.
- Padmakumari, L. & Shaik, M. (2023), ‘An empirical investigation of value at risk (var) forecasting based on range-based conditional volatility models’, *Engineering Economics* **34**(3), 275–292.
- Polydoros, A. S. & Nalpantidis, L. (2017), ‘Survey of model-based reinforcement learning: Applications on robotics’, *Journal of Intelligent & Robotic Systems* **86**(2), 153–173.

- Prabhudesai, S., Yang, L., Asthana, S., Huan, X., Liao, Q. V. & Banovic, N. (2023), Understanding uncertainty: How lay decision-makers perceive and interpret uncertainty in human-ai decision making, *in* ‘Proceedings of the 28th International Conference on Intelligent User Interfaces’, pp. 379–396.
- Rajendran, R. & Karthi, A. (2022), ‘Heart disease prediction using entropy based feature engineering and ensembling of machine learning classifiers’, *Expert Systems with Applications* **207**, 117882.
- ronanmccaughey (2017), ‘Top life insurance purchasing regions in the uk revealed’.  
**URL:** <https://www.lifeinsuranceinternational.com/news/top-life-insurance-purchasing-regions-uk-revealed/?cf-view>
- Ruggiano, N. & Perry, T. E. (2019), ‘Conducting secondary analysis of qualitative data: Should we, can we, and how?’, *Qualitative Social Work* **18**(1), 81–97.
- Sauce, M., Chancel, A. & Ly, A. (2023), ‘Ai and ethics in insurance: a new solution to mitigate proxy discrimination in risk modeling’, *arXiv preprint arXiv:2307.13616*.
- Savran, W. H., Bayona, J. A., Iturrieta, P., Asim, K. M., Bao, H., Bayliss, K., Hermann, M., Schorlemmer, D., Maechling, P. J. & Werner, M. J. (2022), ‘pycsep: a python toolkit for earthquake forecast developers’, *Seismological Society of America* **93**(5), 2858–2870.
- Schlosser, A. (2023), ‘To recommend or not recommend is the question: Does nps predict word-of-mouth?’, *International Journal of Market Research* p. 14707853231186309.
- Schneider, A., Hommel, G. & Blettner, M. (2010), ‘Linear regression analysis: part 14 of a series on evaluation of scientific publications’, *Deutsches Ärzteblatt International* **107**(44), 776.
- Shah, C. H., Onukwugha, E., Zafari, Z., Villalonga-Olives, E., Park, J.-e. & Slejko, J. F. (2022), ‘Economic burden of comorbidities among copd patients hospitalized for acute exacerbations: an analysis of a commercially insured population’, *Expert Review of Pharmacoeconomics & Outcomes Research* **22**(4), 683–690.
- Shamsuddin, S. N., Ismail, N. & Roslan, N. F. (2022), ‘What we know about research on life insurance lapse: A bibliometric analysis’, *Risks* **10**(5), 97.

- Shen, H., Jin, H., Cabrera, Á. A., Perer, A., Zhu, H. & Hong, J. I. (2020), ‘Designing alternative representations of confusion matrices to support non-expert public understanding of algorithm performance’, *Proceedings of the ACM on Human-Computer Interaction* **4**(CSCW2), 1–22.
- Shockey, T. M., Babik, K. R., Wurzelbacher, S. J., Moore, L. L. & Bisesi, M. S. (2018), ‘Occupational exposure monitoring data collection, storage, and use among state-based and private workers’ compensation insurers’, *Journal of occupational and environmental hygiene* **15**(6), 502–509.
- Siedlecki, S. L. (2020), ‘Understanding descriptive research designs and methods’, *Clinical Nurse Specialist* **34**(1), 8–12.
- Sileyew, K. J. (2019), ‘Research design and methodology’, *Cyberspace* pp. 1–12.
- Simjanoska, M., Kochev, S., Tanevski, J., Bogdanova, A. M., Papa, G. & Eftimov, T. (2020), ‘Multi-level information fusion for learning a blood pressure predictive model using sensor data’, *Information Fusion* **58**, 24–39.
- Srujana, P. S., Devi, P. A. & Neelima, C. (n.d.), ‘Data analytics for improved search strategies and its applications’.
- Syed, R., Eden, R., Makasi, T., Chukwudi, I., Mamudu, A., Kamalpour, M., Kapugama Geeganage, D., Sadeghianasl, S., Leemans, S. J., Goel, K. et al. (2023), ‘Digital health data quality issues: Systematic review’, *Journal of Medical Internet Research* **25**, e42615.
- Tahraoui, H., Toumi, S., Hassein-Bey, A. H., Bousselma, A., Sid, A. N. E. H., Belhadj, A.-E., Triki, Z., Kebir, M., Amrane, A., Zhang, J. et al. (2023), ‘Advancing water quality research: K-nearest neighbor coupled with the improved grey wolf optimizer algorithm model unveils new possibilities for dry residue prediction’, *Water* **15**(14), 2631.
- Taplin, R. (2021), ‘Esg and good corporate governance in relation to the use of pension funds: Comparison between the united kingdom and south africa (the report)’, *Interdisciplinary Journal of Economics and Business Law* **10**.
- Thomas, P. & Brunskill, E. (2016), Data-efficient off-policy policy evaluation for reinforcement learning, in ‘International Conference on Machine Learning’, PMLR, pp. 2139–2148.

- Upreti, V., Adams, M. & Jia, Y. (2022), ‘Risk management and the cost of equity: evidence from the united kingdom’s non-life insurance market’, *The European Journal of Finance* **28**(6), 551–570.
- Wang, X., Yuan, X., Feng, R. & Dong, Y. (2022), ‘Data-driven probabilistic curvature capacity modeling of circular rc columns facilitating seismic fragility analyses of highway bridges’, *Earthquake Engineering and Resilience* **1**(2), 211–224.
- Wang, Y. & Xu, W. (2018), ‘Leveraging deep learning with lda-based text analytics to detect automobile insurance fraud’, *Decision Support Systems* **105**, 87–95.
- Watson, R. (2015), ‘Quantitative research’, *Nursing standard* **29**(31).
- Westphal, E. & Seitz, H. (2021), ‘A machine learning method for defect detection and visualization in selective laser sintering based on convolutional neural networks’, *Additive Manufacturing* **41**, 101965.
- White, J. & Power, S. D. (2023), ‘k-fold cross-validation can significantly over-estimate true classification accuracy in common eeg-based passive bci experimental designs: An empirical investigation’, *Sensors* **23**(13), 6077.
- Wissuchek, C. & Zschech, P. (2023), ‘Survey and systematization of prescriptive analytics systems: Towards archetypes from a human-machine-collaboration perspective’.
- Yang, L., Wu, H., Jin, X., Zheng, P., Hu, S., Xu, X., Yu, W. & Yan, J. (2020), ‘Study of cardiovascular disease prediction model based on random forest in eastern china’, *Scientific reports* **10**(1), 5245.
- Zhong, S., Zhang, K., Bagheri, M., Burken, J. G., Gu, A., Li, B., Ma, X., Marone, B. L., Ren, Z. J., Schrier, J. et al. (2021), ‘Machine learning: new ideas and tools in environmental science and engineering’, *Environmental Science & Technology* **55**(19), 12741–12754.
- Zhou, J., Li, E., Yang, S., Wang, M., Shi, X., Yao, S. & Mitri, H. S. (2019), ‘Slope stability prediction for circular mode failure using gradient boosting machine approach based on an updated database of case histories’, *Safety Science* **118**, 505–518.

# Appendix A

## Insurance Premium Prediction Program Manual

### A.1 Introduction

Welcome to the Insurance Premium Prediction Program! This software is designed to help you estimate insurance premium costs based on various factors such as age, gender, BMI, smoking status, region, and more. This manual will guide you through the process of using the program and obtaining insurance premium predictions.

### A.2 System Requirements

Before running the program, please ensure that your computer meets the following requirements:

- **Python 3.x:** Make sure you have Python 3.x installed on your computer.
- **Required Libraries:** Install the necessary Python libraries and dependencies, including pandas, scikit-learn, tkinter.
- **Internet Connection:** A stable internet connection may be required if the program uses external data sources.

### A.3 Installation

To install and set up the program, follow these steps:

- 1) Clone or download the program from the provided repository URL

- 2) Open a terminal or command prompt and navigate to the program directory.
- 3) Install the required Python libraries by running the following command:

```
pip install -r requirements.txt
```

## A.4 Running the Program

To run the program, follow these steps:

- 1) Open a terminal or command prompt on your computer.
- 2) Navigate to the directory where you have the program installed.
- 3) Execute the program using the following command:

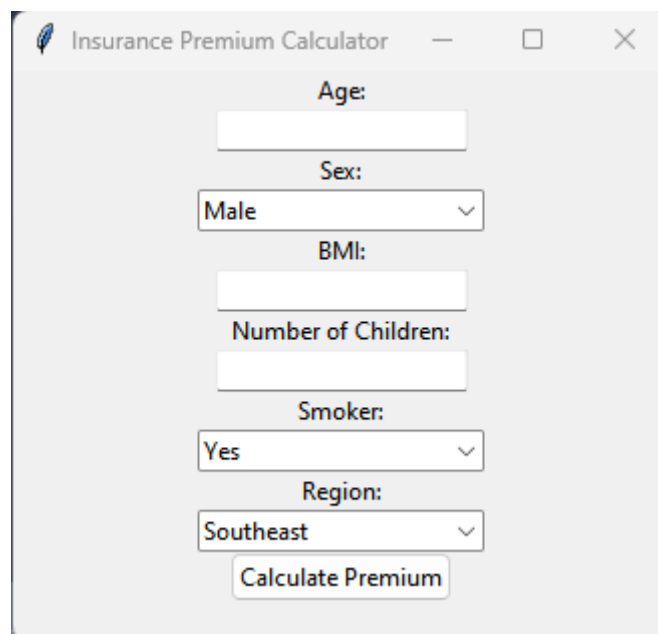
The image shows a window titled "Insurance Premium Calculator". Inside the window, there are several input fields and dropdown menus arranged vertically. The fields are labeled: "Age:" (with a text input box), "Sex:" (with a dropdown menu showing "Male"), "BMI:" (with a text input box), "Number of Children:" (with a text input box), "Smoker:" (with a dropdown menu showing "Yes"), and "Region:" (with a dropdown menu showing "Southeast"). At the bottom of the form is a button labeled "Calculate Premium".

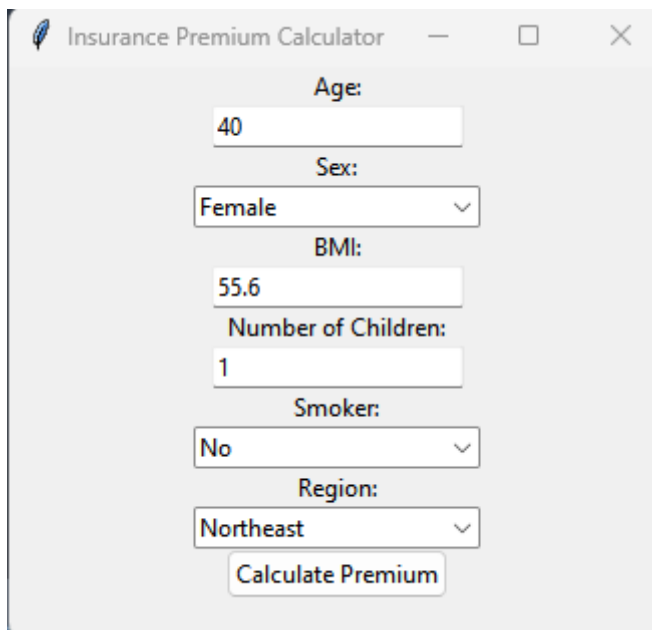
Figure A.1: Prompt Screen

## A.5 Running the Program

Upon running the program, you will be prompted to enter the following information:

- **Age:** Enter your age as a whole number.
- **Gender:** Specify your gender as 'male' or 'female'.
- **BMI:** Enter your BMI as a decimal number.

- **Number of Children:** Indicate the number of children or dependents you have as a whole number.
- **Smoking Status:** Enter 'yes' if you are a smoker, or 'no' if you are not.
- **Region:** Specify your region (e.g., 'southeast', 'southwest', 'northeast', 'northwest').

A screenshot of a web application window titled "Insurance Premium Calculator". The window contains several input fields and dropdown menus. The "Age" field is a text box containing "40". The "Sex" field is a dropdown menu with "Female" selected. The "BMI" field is a text box containing "55.6". The "Number of Children" field is a text box containing "1". The "Smoker" field is a dropdown menu with "No" selected. The "Region" field is a dropdown menu with "Northeast" selected. At the bottom of the form is a button labeled "Calculate Premium".

Field	Value
Age	40
Sex	Female
BMI	55.6
Number of Children	1
Smoker	No
Region	Northeast

Figure A.2: Entered Details

## A.6 Viewing the Results

After providing the required input, the program will calculate and display the estimated insurance premium based on your input.

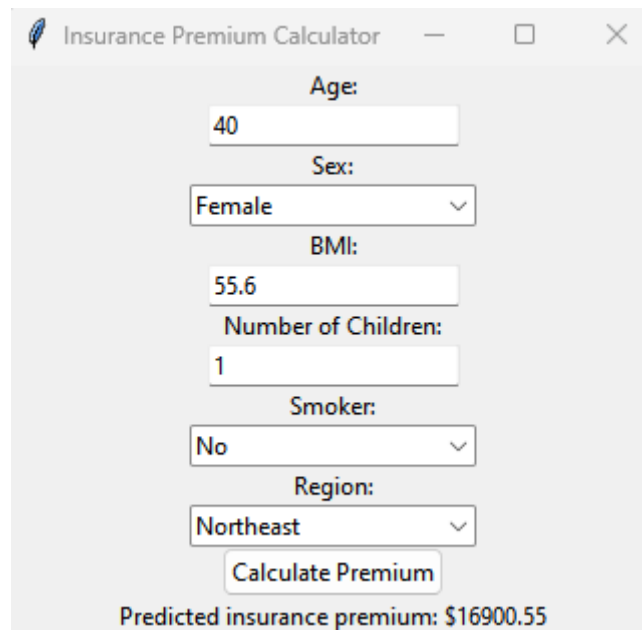
## A.7 Troubleshooting

If you encounter any issues while running the program or have questions, please consult the troubleshooting section in the program's documentation or contact the program's support team for assistance.

## A.8 Appendix

For additional details about data sources, data preprocessing, the machine learning model used, external libraries, acknowledgments, and references, please refer to the





The image shows a software window titled "Insurance Premium Calculator". It contains several input fields and dropdown menus for user data. The inputs are: Age (40), Sex (Female), BMI (55.6), Number of Children (1), Smoker (No), and Region (Northeast). Below these inputs is a "Calculate Premium" button. At the bottom of the window, the text "Predicted insurance premium: \$16900.55" is displayed.

Field	Value
Age	40
Sex	Female
BMI	55.6
Number of Children	1
Smoker	No
Region	Northeast
Predicted insurance premium	\$16900.55

Figure A.3: Insurance Premium Price

program's documentation.

Thank you for using the Insurance Premium Prediction Program! If you have any feedback or suggestions, please feel free to contact us.