



**To investigate the use of big data technologies and techniques that can be
used to solve data analysis problems in small companies of the UK**

By

Madhusudan Ghumare

Supervisor: Michael Dacey

Project submitted as part of the requirements for the award of MSc in Data
Science and Analytics

University of Wales Trinity Saint David

September 2023

DECLARATION

I, Madhusudan Sureshrao Ghumare declare that I am the sole author of this Project; that all references cited have been consulted; that I have conducted all work of which this is a record, and that the finished work lies within the prescribed word limits. This work has not previously been accepted as part of any other degree submission.



Signed:

Date: 15/09/2023.....

FORM OF CONSENT

I Madhusudan Sureshrao Ghumare hereby consent that my Project, submitted in candidature for the M.Sc. Data science and Analytics degree, if successful, may be made available for inter-library loan or photocopying (subject to the law of copyright), and the title and abstract may be made available to outside organisations.



Signed:

Date: 15/09/2023.....

Acknowledgement

I want to sincerely thank my adviser for their consistent advice, perceptive criticism, and ongoing support throughout the process of writing this dissertation. Their advice and support were crucial in guiding and improving the research. I also want to express my gratitude to my peers and colleagues for their insightful comments and contributions that improved the study. My deepest gratitude goes out to all of the participants who voluntarily provided their thoughts and experiences to enable this research. Finally, I want to express my gratitude to my family and friends for their constant support and confidence in me.

Executive Summary

This study digs into the use of big data technology and approaches to handle data analysis difficulties in small enterprises in the (UK). When dealing with complicated databases, small businesses sometimes face resource limits and a lack of knowledge. The project is to investigate how solutions based on big data might be used to overcome these constraints, providing useful insights for improved decision-making. The investigation includes identifying common analytical issues in small UK businesses, evaluating relevant technologies such as Spark, Hadoop, and evaluating techniques such as data mining, machine learning, and predictive analytics. Small organisations may uncover previously undiscovered correlations, patterns, and trends by employing these modern technologies. The study also identifies problems associated with the application of big data solutions in small organisations, including security, data privacy, and scalability concerns. The report offers advice customised to assist small UK businesses in selecting, implementing, and optimising big data solutions that meet their specific needs and goals. The ultimate purpose of this project is to give practical insights that will assist small companies in the United Kingdom to harness the revolutionary potential of big data, allowing them to make educated choices and gain a competitive advantage in their particular sectors.

The study's primary goals concentrate around analysing common difficulties faced by small businesses and developing solutions applicable to datasets similar to those observed in large-scale organisations using big data. The research evaluates the use of statistical methods and instruments to datasets in small company contexts using empirical assessment. The study assesses the effectiveness of these strategies and principles in smaller organisational contexts, taking into account resource constraints and operational dynamics. The study's conclusions emphasise the importance of data-driven decision-making in today's corporate context. While bigger organisations have adopted big data technologies, small firms are just now beginning to see their benefits. Both large and small organisations face comparable obstacles when using big data analytics. The research also highlights the importance of data analysis methodologies and ideas in providing actionable insights from varied datasets, supporting better decision-making, and improving operational performance. In these circumstances, the study calls for the use of big data methodologies to address data-related issues and capitalise on the possibilities of decision-making based on data.

Furthermore, the study investigates the transformational influence of big data on data analysis procedures, emphasising the importance of tailoring these approaches to the particular restrictions and possibilities inherent in small company environments. It assesses the benefits and challenges of using advanced data analysis techniques inside smaller organisations, emphasising the significance of empirical validation, specialised solutions, and customised adaption in order to maximise effectiveness. Based on the study's results, the paper provides a series of thorough suggestions for small firms looking to effectively exploit big data solutions. These recommendations include the establishment of a strong data governance framework, the use of data visualisation and machine learning, thorough data pre-processing, and tailored method adaptation. By following these advice, small companies may use large amounts of information to inform their strategy, improve decision-making, and achieve an edge in their particular industries.

The result and discussions that resulted from a careful analysis of hotel cancellation data. Patterns of duplication and null values are handled through data preparation. Exploratory Data Analysis (EDA) reveals information on correlations, such as the influence of lead time on cancellations. Strong predictive capabilities are shown by model assessments of Random Forest, Decision Tree, and Logistic Regression. Strategies for improving booking procedures, reducing cancellations, and improving hotel administration are revealed in this chapter. The comprehension of consumer behaviour and model efficacy are improved by the insights. Validation on fresh data is advised, nevertheless, for use in actual situations. Overall, this chapter contributes to the conversation on hotel cancellations by outlining relevant operational implications for the hospitality sector.

Table of Contents

Acknowledgement	3
Executive Summary	4
List of Figures	6
Chapter 1: Introduction	10
Overview of the study	10
Background of the Research	10
Problem Statement	11
Research Aim	11
Research Objectives	12
Research Questions	12
The Rationale of the study	12
Scope of the Study	13
Significance of the Study	13
Chapter 2: Literature Review	15
Introduction	15
Theoretical Framework	15
Conceptual Framework	28
Chapter Summary	28
Chapter 3: Research Methodology	29
3.1 Introduction	29

3.2 Dataset used	29
3.3 Tool Used	29
3.4 Data pre-processing	30
3.4.1 Null Values	30
3.4.2 Duplicate values	31
3.4.3 Label encoder	32
3.5 Algorithms Used	33
3.5.1 Random Forest Classifier	33
3.5.2 Decision Tree	34
3.5.3 Logistic Regression	34
3.6 Model Evaluation	35
3.6.1 Confusion matrix	35
3.6.2 Precision	35
3.6.3 Recall	36
3.6.4 F1-score	37
3.7 Summary	37
Chapter 4: Data Analysis and Discussion	38
Introduction	38
Result from coding	39
Data pre-processing	39
EDA	40

Model Evaluation	44
Summary	53
Chapter 5: Conclusion and Recommendation.....	54
Conclusion	54
Recommendation	56
Research Limitation	59
Future Consideration.....	59
Chapter 6: Reflection.....	61
References.....	65

List of Figures

Figure 1: Conceptual framework	28
Figure 2: Null values.....	39
Figure 3: Duplicate values	40
Figure 4: Corelation	40
Figure 5: Number of bookings	41
Figure 6: Types of Hotels.....	42
Figure 7: Resort vs City Hotel	43
Figure 8: Classification report of Random Forest classifier	44
Figure 9: Classification report of Decision tree classifier	45
Figure 10: Classification report of logistic regression	46

Figure 11L: Figure 1 Big Data business	49
Figure 12: Distribution of main development priorities	51

To investigate the use of big data technologies and techniques that can be used to solve data analysis problems in small companies of UK

Chapter 1: Introduction

Overview of the study

The research investigates the use of big data technology and approaches to address data analysis difficulties faced by small enterprises in the United Kingdom (UK). Small businesses sometimes encounter resource limits and a lack of competence when dealing with huge and complicated databases. It seeks to learn how big data solutions may be used to solve these challenges and give significant insights for better decision-making [1]. However, it will delve into a variety of topics, such as identifying common data analysis problems in small UK businesses, evaluating suitable technologies for big data analysis such as Hadoop, Spark and evaluating techniques such as machine learning, data mining, and predictive analytics. Small businesses may draw significant correlations, trends, and patterns by employing these advanced technologies. It will look at various obstacles connected with the implementation of big data solutions in small businesses, such as data privacy, security, and scalability [2]. This research aims to give practical insights for capitalising on big data's revolutionary potential, allowing small businesses in the UK to make informed choices and achieve a competitive advantage in their particular industries.

Background of the Research

The expansion of digital technology in recent years has resulted in unprecedented data creation across numerous industries, including small businesses in the UK. This data explosion, dubbed "big data," poses both obstacles and opportunity for small firms. Big data technology and methodologies have gained recognition as strong solutions for addressing these firms' data analysis concerns [3]. Regardless of size, small firms in the UK face data-related issues such as gathering, storing, analysing, and drawing valuable insights from massive volumes of data. Traditional data analysis approaches are frequently insufficient for dealing with the complexity and amount of data that small businesses today face. Big data technologies include a variety of tools and frameworks, such as Hadoop and Spark distributed computing platforms, as well as processing solutions [4]. These technologies help small organisations to efficiently store, handle, and analyse data despite limited resources. Adoption of big data technology in the

context of small UK businesses can assist overcome resource restrictions and level up competition with larger rivals. These organisations may obtain greater insights into client preferences, market trends, and operational constraints by utilising these tools and strategies. When deploying big data solutions, small companies must carefully assess their unique goals, available resources, and the possible problems connected with data privacy, security, and compliance.

Problem Statement

The problem statement of the study revolves around the issues faced by small businesses in the UK as a result of the boom in digital data generation. These companies have challenges in successfully gathering, storing, analysing, and drawing insights from the massive and complicated statistics they amass. Conventional data analysis methodologies fail to manage and extract value from this data flood. As a result, small businesses struggle to make informed decisions, improve customer engagement, and optimise operations, placing them at a competitive disadvantage when compared to bigger competitors. The research will also look into how big data technology and methodologies may be used strategically to solve the unique data analysis needs of small firms in the UK. The study's goal is to discover appropriate big data technologies, such as distributed computing platforms, that can effectively handle and analyse data within the limits of low resources that small businesses sometimes face [5]. Furthermore, the study intends to assess the suitability of data mining, machine learning, and machine learning-based predictive analytics approaches for extracting important insights from accumulated data, allowing for better decision-making and operational efficiency. However, significant roadblocks such as confidentiality, security, and compliance must be considered to ensure that the deployment of big data solutions corresponds with the particular demands and problems.

Research Aim

To develop solutions to problems related to the analysis of data that has some characteristics that mimic those of big data.

Research Objectives

- To analyse existing problems encountered by small companies to determine solutions that can be applied to datasets that produce issues similar to those encountered in organisations with big data applications
- To evaluate the application of data analysis techniques and tools to the investigation of a small company dataset
- Critically evaluate the techniques and principles applied to determine if they are effective in smaller organisation contexts

Research Questions

- What are the existing problems encountered by small companies to determine solutions that can be applied to datasets that produce issues similar to those encountered in organisations with big data applications?
- What are the data analysis techniques and tools to the investigation of a small company dataset?
- What are the techniques and principles applied to determine if they are effective in smaller organisation contexts?

The Rationale of the study

The rationale behind conducting the research is motivated by the rising importance of big data technologies and methodologies as potential solutions to data analysis difficulties encountered by small enterprises in the UK. With the digital world continuously expanding, these organisations accumulate massive volumes of data, which presents both possibilities and challenges. As the competitive landscape grows more data-driven, small companies must leverage the strength of their data to make informed decisions, improve customer engagement, and optimise operations [6]. This topic was chosen because bigger organisations frequently have the resources to embrace and implement big data methods, small firms have resource limits which restrict their capacity to successfully analyse and extract information from their data. This study aims to close the gap by examining specialised methods to integrating big data technologies and methodologies in the context of small UK businesses. This research intends to help small firms to exploit their data as a strategic asset by identifying cost-effective and

scalable solutions. The research investigating the problems which is consistent with wider business trends, such as the democratisation of data analytics and technology adoption, which is becoming increasingly crucial for firms of all sizes. This research contributes to a fairer business climate by addressing the unique issues encountered by small businesses, supporting innovation and growth even when resources are limited. Overall, the study's results can help small firms in the UK leverage the power of big data, enhancing their competitiveness and sustainability in an ever-changing market context.

Scope of the Study

The goal of this research is to thoroughly evaluate the use of big data technologies and approaches to address data analysis difficulties in small businesses in the United Kingdom. The study will dive into the whole range of data-related difficulties that these firms confront, including data gathering, storage, processing, analysis, and actionable insights. The research will identify and evaluate several big data technologies such as Spark, Hadoop, and taking into account their appropriateness and practicability for adoption in resource-constrained contexts characteristic of small firms [7]. Furthermore, the study will dig into investigating their utility in extracting important insights that may guide decision-making, consumer engagement tactics, and operational improvements. The study will include a literature review, case studies, and surveys or interviews with small company owners and experts in the sector. By working within these constraints, the research hopes to give practical knowledge and suggestions that small businesses in the UK can implement to successfully exploit big data for a competitive edge.

Significance of the Study

The importance of this study lies in its ability to address the critical data analysis difficulties faced by small businesses in the UK and to provide real answers that help fuel their growth and efficiency. The value of extracting important insights from gathered data cannot be stressed as the corporate landscape becomes increasingly data-driven [8]. This research is required to bridge the gap between the availability of accessible big data technology and the limited resources that small enterprises often have. Because of the inherent restrictions that small businesses confront, such as budget limits, a lack of specialised employees, and limited technology infrastructure, research in this field is critical. These obstacles frequently impede their capacity to fully realise the potential of big data. This research is a thorough guide that

emphasises customised ways to implement and utilise big data technologies and methodologies, assuring their practical usefulness in small company situations.

The advantages of this study extend to a wide range of stakeholders. Small company owners and managers will acquire insights into how to successfully manage their data assets, which will lead to better decision-making, increased customer engagement tactics, and streamlined operations. Furthermore, the study adds to the corpus of information on big data deployment in small company settings, perhaps encouraging future research in related disciplines [9]. The study's findings can also aid policymakers and industry professionals by informing policy suggestions and activities targeted at supporting creativity and development among small businesses, therefore contributing to the UK's broader economic environment.

Chapter 2: Literature Review

Introduction

A literature review is a succinct summary and assessment of the body of scholarly literature on a particular subject. It entails locating, evaluating, and synthesising pertinent sources, including books, journals, and academic papers. The review's goal is to give readers a thorough grasp of the state of the field's current knowledge by identifying important trends, controversies, and knowledge gaps. It adds to academic discourse and helps in the formulation of research questions or hypotheses by critically evaluating alternative views, methodology, and findings. A solid literature review serves as the foundation for research efforts, allowing researchers to situate their findings within the larger academic debate and to show that they are aware of earlier studies and their significance.

Theoretical Framework

To analyse existing problems encountered by small companies to determine solutions that can be applied to datasets that produce issues similar to those encountered in organisations with big data application

Data-driven decision-making has become crucial to organisational performance in the modern business environment. Small businesses are increasingly realising the potential rewards of using data to spur growth and innovation, while larger companies have established themselves as early consumers of big data applications [10]. However, when dealing with data-related problems, small and large businesses face comparable difficulties that might restrict development and prevent the best use of available resources. To remedy these concerns for datasets causing similar problems within larger organisations using big data applications, this article will explore the current issues small businesses face while managing data-related challenges.

Small businesses frequently face budgetary, human resources and technological capacity restrictions. These restrictions make it more challenging for them to efficiently collect, store efficiently, and process data [11]. For instance, it might not be possible financially to purchase cutting-edge data analytics tools and hire qualified data specialists. Small businesses need help with financial, human resources, and technological constraints restricting their ability to handle

data effectively. Hiring skilled data specialists and acquiring advanced analytics resources can be cost-impossible. As a result, these businesses challenge to fully utilise data for informed decision-making and gain an advantage in the market.

Small businesses might need more inside information to handle duties utilising data efficiently. This covers the gathering, preparation, analysis, and comprehension of data. Legal and moral issues may arise from inadequate awareness of data security and privacy safeguards [12]. Both small and large enterprises need help to ensure knowledge accuracy and reliability. Small businesses may collect data manually or from various sources, which can result in inaccuracies and errors [13]. Poor data quality might skew analyses and make it more challenging to make wise decisions. In the final analysis, maintaining data reliability and accuracy is a challenge faced by businesses of all sizes. Smaller companies' dependency on human data collecting and several sources might lead to inaccuracies and imprecisions. The integrity of analyses is affected by such reduced data quality, which also obstructs the way to making informed choices. To overcome these obstacles, serious attempts must be made to establish reliable data confirmation processes and standardised data collection techniques, ensuring that small and large businesses can obtain accurate data for sound decision-making.

Expanding small businesses face the challenge of scaling data infrastructure. Expansion frequently burdens ageing systems when dealing with increasing information quantities, leading to inefficiencies and decreased performance [14]. This issue must be resolved to maintain operational effectiveness and promote informed decision-making. By changing to more advanced systems, you can have the adaptability you need to keep up with the growing demand for data and continued growth. By using adaptive technologies, businesses can guarantee flawless data management, take advantage of possibilities, and keep their competitive edge in the ever-changing market [15].

Getting an overall business view requires integrating data from many sources and functions. Small businesses may have disconnected systems that are difficult to communicate with, making it challenging to understand operations [16]. In summary, establishing a comprehensive business perspective depends critically on collecting data from various sources and organisations. Due to their separate systems, small businesses frequently need help to get a consistent operating view. For informed choices and operational optimisation, these knowledge

gaps must be closed. Businesses of any size can gain valuable insights, improve efficiency, and adapt to changing market needs by facilitating seamless data exchange and integration. A critical initial move in unlocking the full potential of organisational data and fostering sustainable growth in the current competitive business climate is pursuing the development of integrated data systems.

A data management framework's implementation aids in maintaining the accuracy and consistency of data. Small businesses can set rules for gathering, storing, and utilising data. Large businesses can profit from centralised data governance in the context of big data applications to guarantee consistency across departments. Establishing a data governance structure is crucial to maintaining the reliability and integrity of data [17]. Setting up comprehensive instructions for data collection, storage, and use is necessary to provide a solid foundation for small businesses. Larger firms benefit from centralised data governance in the ample data space, maintaining consistency across many departments and apps. Such a structured strategy optimises the decision-making process and protects data integrity. Big or small businesses that emphasise data governance are better positioned to negotiate the challenges of contemporary data management and succeed in their endeavours over the long term.

Small businesses can overcome their lack of in-house experience by outsourcing data-related work or working with outside experts. Similarly, massive companies might collaborate with specialised data consulting firms to address complex big data fears that are outside the scope of their internal resources [18].

Small enterprises might use data integration technologies to allow smooth data transfer between diverse systems. This increases access to information and accuracy. Robust integration solutions allow a unified view of the company operations in larger businesses where data circulates across several divisions [19].

To sum up, businesses of all sizes can significantly profit from adopting data integration alternatives. These tools facilitate data flow between systems in small businesses, improving consistency and accessibility. Effective integration solutions give larger firms, which struggle with separate databases across branches, a unified view of the whole operational picture.

Businesses of all sizes may improve processes, gain a competitive edge, and make better decisions by promoting efficient data integration [20]. Organisations may unlock the full potential of their data assets and enhance efficiency and innovation in their specific fields by adopting data integration as a strategic priority.

To protect sensitive information, small businesses can employ essential data security features like protection and access controls. Large enterprises must set up thorough security procedures and consistently audit their systems in the setting of big data, where security issues can have far-reaching effects.

The presentation of data makes complex information easier for small businesses to understand. Similarly, significant businesses can use cutting-edge data visualisation tools to provide insights to stakeholders at various levels in a way that makes sense. Data visualisation is an effective tool for improving comprehension in small and large businesses. Illustrations make complex information simpler for small businesses, facilitating efficient processing. The simple presentation of insights to stakeholders at various levels is made possible in larger businesses by advanced data visualisation tools. Businesses of all sizes can promote openness, support decision-making, and develop stronger connections with data by utilising visual aids. Integrating data visualisation strategies ultimately enables businesses to realise the total value of their data, enabling them to make informed choices and improve communication in today's data-driven business environment.

Small enterprises can use machine learning algorithms to develop insights and automate data analysis procedures. In big data, businesses may use machine learning models to process large amounts of information swiftly, find trends, and make predictions in the present. Small and large businesses can benefit substantially from machine learning algorithms.

Small organisations can use such algorithms to automate insights development, speed data analysis, and improve operational efficiency. Organisations profit from machine learning's fast processing speeds in the context of massive data, which makes it possible to identify complex patterns and make recommendations in real-time [21]. Machine learning is carefully included to encourage data-driven decision-making, which drives firms toward innovation and

competitiveness. Adopting machine learning is crucial in maximising data use and facilitating continuous growth across various company sizes as technology advances.

It can be critically evaluated that dealing with data-related worries poses similar obstacles for small and large businesses. The basic concepts regarding information quality, resource limitations, knowledge, and integration are still applicable even though the size of the challenges may vary. The ideas provided here are inspired by techniques used in larger organisations with big data applications, and startups can learn from them. Smaller companies can better handle data-related difficulties by adopting cloud computing, enacting data governance, utilising integration tools, and prioritising data security. More significant firms may simultaneously optimise their big data strategy by including automation, machine learning, and visualizations aspects in their methods.

To evaluate the application of data analysis techniques and tools to the investigation of a small company dataset

The process of analysing, cleaning, manipulating, and modelling data in order to identify useful details, draw conclusions, and help decision-making is known as data analysis. [22] It entails using a variety of approaches, instruments, and techniques to extract useful information from raw data. Data analysis can be used to a variety of data kinds, including structured information (tabular information found in databases and sheets) and unstructured information (text, photos, videos, and so on).[23] Data analysis enables organisations and people to take better choices based on data facts and insights. Data analysis can reveal patterns, trends, and connections that were previously unknown. Data analysis is a critical component of business intelligence. It lets businesses to gain a better understanding of their customers' actions, market trends, and operational performance.[24] This data can be utilised to develop plans, allocate resources, and remain competitive. Data analysis is the process of evaluating, washing, influencing, and analysing data in order to uncover useful features, draw conclusions, and aid decision-making[25].It comprises extracting useful details from raw data using a range of approaches, devices, and processes. Analytical techniques can be used to a wide range of data types, including structured data (tabular data that exists in spreadsheets and files) and unstructured data (text, photographs, videos, and so on). Data analysis allows businesses and individuals to make better decisions using information facts and insights. Analysis of data can reveal

previously unseen patterns, trends, and linkages. Data analysis is an important part of business intelligence. It enables companies to obtain greater knowledge of their consumers' activities, economic conditions, and operational efficiency [26].

Techniques

Data Pre-processing: It is critical to process the data before diving into the main analysis. Cleaning, converting, and organising data to guarantee its quality and dependability [27]. Techniques such as addressing missing values, detecting outliers, and data normalisation or scaling aid in preparing data for precise analysis.

Descriptive Analysis: To get preliminary insights, descriptive analysis entails summarising and visualising the dataset. The standard deviation, mean, and median are basic statistical measurements that provide a glimpse of the data's central patterns and variability. Visualisation techniques like as bar graphs, histograms that and scatter plots provide a visual representation of the distribution of data and relationships.

EDA: EDA entails more in-depth research to find patterns, correlations, and trends within a dataset. Techniques like scatter graphs, box graphs, and correlation matrices aid in analysing variables' relationships [28]. Advanced visualisation methods like heatmaps and tree maps may also be used in EDA to help find hidden insights.

Statistical Analysis: Statistical techniques are essential for extracting meaningful inferences from datasets. Hypotheses examination, t-tests, an ANOVA, or regression analysis help to determine how different factors affect a company's performance, revenue, or other important metrics [29].

Monte Carlo: When building decisions or taking certain actions, many different possibilities come out. If public transport is first priority, then there a chance to get stuck in traffic. If the walk is second priority, then there a chance to get wet in the rain or possibility to delay in the success journey. For that every man deeply analyses every decision that going to be taken or what are the benefits of the decision. Monte Carlo is a well-known tactic as the monte carlo method it is a computerised technology to generate possible models of outcomes and provide

the possibility of errors[30]. This technology provides many possibilities for the outcome and analyse every possibility of an outcome that is going to be successful. Monte Carlo used data analysts to recon advanced risk and help to better understand what might be happening in the future and take decisions accordingly. To run monte carlo small business have to arrange his data in mathematical stretcher as a spreadsheet. With the spreadsheet, a small business can get a number of variables, such as the number of sales profits and this variable will impact your outcome [31]. If the business is looking for the profit variable, then relevant input is also included, such as market spending, employee salaries or sales. If the data is exact for all your inputs, then it's easy to calculate the profit that will be left in the end. For the research investigating into big data technologies in tiny UK companies, Monte Carlo simulation is crucial. It addresses handling incomplete data, tailoring for smaller enterprises, and decision-making ambiguity. Monte Carlo assists in making educated decisions by simulating numerous scenarios, calculating risks, and allocating resources optimally. Small businesses benefit from its versatility, which ensures better financial planning and risk management. Additionally, it thrives with faulty data, which makes it appropriate for contexts with limited resources.

Analytic methods: small businesses can use data analytic methods and resources to acquire insights, make informed choices, and drive growth in today's data-driven environment [32]. There are numerous methods available, ranging from preparation and exploratory research to advanced techniques such as clustering and predictive modelling.

Languages: Python and R are popular data analysis languages because of their extensive resources for machine learning, statistical analysis, and data visualisation [33]. Their adaptability enables unique analysis customised to the company's specific requirements.

Excel and Spreadsheets: Excel and equivalent spreadsheet software, while not as sophisticated as specialised data analysis tools, can be useful for basic analysis, visualisation, and even simple modelling.

Time Series Analysis: This type of analysis is critical for businesses that have time-stamped data (for example, sales information over months or years). Shifting medians, exponentially smoothed data, and ARIMA models are used to uncover trends, the seasons, and potential forecasting.

Data Management: Databases management platforms (DBMS) such as PostgreSQL, MySQL, and other relational or non-SQL databases (MongoDB, Cass) are necessary for effective data archiving and retrieval. They guarantee the accuracy, safety, and effectiveness of the data.

Factor Analysis: A statistical method called factor analysis is used to find concealed trends or latent variables in a dataset. By focusing on a smaller number of invisible influences that account for the observed relationships between variables, it seeks to simplify the complexity of the data. This makes it easier for academics to comprehend data and aids in their understanding of how variables interact. Factor loadings, which indicate the strength and trajectory of the connection between variables and factors, are calculated in factor analysis under the assumption that observable variables are influenced by these latent components. The factors are selected to best capture the variation in the original variables. This method is used to build composite indices, detect hidden dimensions, and increase measurement accuracy in a variety of domains, including psychology, economics, and the social sciences. However, careful examination of the framework's presumptions, size of the sample, and data quality is necessary for the accurate interpretation of the results.

Tools

Spreadsheet Software: One of the instruments for data analysis that is the easiest to use is spreadsheet software. Utilising tools such calculations, pivot tables, and charts, it enables users to organise, modify, and visualise data. Spreadsheet applications can be a smart place to start for small businesses for doing fundamental data analysis activities like keeping track of sales, expenses, and client information.

Statistical Analysis: SAS is a complete software package made for sophisticated statistical analysis. It provides a variety of tools for regression analysis, hypothesis testing, and data exploration. When working with complicated datasets that demand sophisticated statistical modelling, SAS is especially helpful.

Python: Python is a programming language with great flexibility that has become more well-known in the areas of machine learning and data analysis. Tools for data processing, numerical calculation, and data visualisation are available from frameworks like the pandas, NumPy, and

Matplotlib. Python is a great option for small businesses wishing to undertake analysis of data without a challenging learning curve due to its simplicity.

Tableau: Users can create dynamic and shareable dashboards using Tableau, a potent data visualisation tool. Due to its ability to link to numerous data sources, such as databases and spreadsheets, small businesses may visualise their data and derive valuable insights without requiring a high level of technological expertise.

SQL: For managing and querying data kept in databases, SQL is necessary. Databases are frequently used by small businesses to store data such as transaction histories and client information. Knowing SQL well enables firms to execute studies based on their particular requirements and extract particular data subsets.

Google Analytics: Google Analytics is a useful tool for small businesses with an online presence to monitor website traffic, user behaviour, and conversion rates. It offers insights about customer behaviour on the business' website, assisting with the user experience and marketing strategy optimisation.

Qlik View: Users can build interactive visualisations and dashboards using the business analytics and data visualisation platform QlikView. It is appropriate for small businesses with a variety of data sets because it can link to different data sources.

Excel: For Excel, a number of add-ins are available to increase its data analysis capabilities. Solver and other tools help with optimisation, while the Data Analysis Tool Pak provides further statistical features.

JMP: JMP is a statistics discovery tool created for exploratory analysis and dynamic data visualisation. Users may examine their data and spot trends and anomalies with the use of adaptive and graphical tools that are provided.

To develop solutions to problems related to the analysis of data that has some characteristics that mimic those of big data.

Critically evaluate the techniques and principles applied to determine if they are effective in smaller organisation contexts

The primary goal emphasises the crucial need for an extensive and critical review of the wide range of methodologies and ideas that form the foundation of extensive data analysis. These strategies, painstakingly created and perfected to deal with massive datasets' complexities and sheer vastness, have received intense attention, mainly within the broad world of large-scale organisations [34]. However, the smooth transfer of these strategies to the comparatively constricted terrain of smaller organisational units demands an in-depth examination of their effectiveness, relevance, and adaptation in this context.

The big data analytics market combines powerful tools and approaches, ranging from sophisticated machine learning algorithms to resilient distributed computing frameworks [35]. This evaluation endeavour connects these strategies' theoretical potency with their practicality inside smaller organisational settings [36]. The research seeks to provide insights into their potential alignment with the nuanced challenges and opportunities inherent in smaller organisational contexts by subjecting these techniques to a critical appraisal that delves into computational requirements, scalability, ease of implementation, and resource efficiency.

The study aims to thoroughly analyse the usability of extensive data analysis methodologies created for large organisations within smaller companies. While these strategies are well-known for managing large datasets, their applicability in limited environments requires careful consideration [37]. Notably, the generally praised MapReduce technique, recognised for processing large amounts of data in distributed systems, may have difficulties when scaled down. The research examines the benefits and limits of these methodologies in smaller contexts, considering issues such as computing needs, scalability, implementation complexity, and efficiency. This evaluation connects theoretical promise with actual practicality, offering insight into potential changes required for seamless integration [38]. The research intends to provide insights on aligning these strategies with the particular problems and opportunities in smaller organisational environments by critically examining these features.

Assess Applicability in Smaller Organisation Contexts:

Smaller businesses' operational dynamics differ significantly from those of more prominent entities, laying the stage for possible differences when using complicated significant data approaches. Smaller organisations frequently face unique resource restrictions, such as financial limits, limited IT infrastructures, and a scarcity of specialised staff committed to data analysis. As a result, a straight transplanting of complex significant data approaches may confront impracticality, inefficiency, or exorbitant prices in this environment. Furthermore, it needs a thorough examination of the projected advantages, considering the unique operational dynamics and limits typical in smaller organisations [39]. Through this comprehensive examination, the research intends to give insights into the potential of modifying current procedures and provide light on their smooth absorption inside smaller organisations. The results of this evaluation attempt to provide a comprehensive knowledge of the compatibility of these strategies with the practical complexities and constraints that characterise smaller organisational contexts.

The study aims to properly assess the appropriateness of the investigated significant data approaches within smaller organisational contexts. Because of the operational variations between small and big organisations, implementing complicated data approaches may meet disparities owing to restricted resources, confined IT setups, and a lack of specialised people. Transferring these approaches directly may encounter challenges such as cost inefficiency or impracticality [40]. This evaluation step necessitates thoroughly examining computational requirements, resource requirements, and potential integration roadblocks.

Develop Tailored Solutions:

The creative endeavour of building unique and precisely designed solutions based on the insights gained from the critical evaluation and applicability assessment. The primary goal here is not simply to transfer techniques from one context to another but to devise procedures that balance the theoretical robustness of big data techniques with the pragmatic needs of smaller organisations. The goal is to create an artistic equilibrium that embraces the potential of modern approaches while deftly negotiating the resource limits imposed by smaller organisations' size and scope. Scalability, resource efficiency, integration feasibility, and alignment with current organisational procedures are all factors to consider when developing customised solutions [41]. This phase may include a variety of measures, such as simplifying complex algorithms,

leveraging cloud computing capabilities to avoid infrastructure limits, or integrating modern approaches with more traditional statistical methodologies. The ultimate goal is to create intellectually sound and viable ways within smaller organisations' operational limits. The research aims to grow solutions that optimise both theoretical potency and practical practicality by cleverly connecting sophisticated approaches with the restrictions and obstacles faced by smaller organisations.

Simplifying algorithms, utilising cloud computing to overcome infrastructure restrictions, and combining sophisticated approaches with classic statistical methods are possible [42]. The objective is to develop ways that are theoretically sound and operationally possible within the operational frameworks of smaller organisations. The project attempts to nurture solutions that combine theoretical soundness with practical feasibility by adeptly connecting complex methodologies with the constraints of smaller businesses.

Empirical Validation:

Empirically validate the painstakingly constructed customised solutions within the complicated fabric of smaller organisational contexts. While theoretical frameworks give core knowledge, empirical validation proves the suggested solutions' real-world practicality and usefulness [43]. This validation procedure functions as a litmus test, revealing the concrete effects, possible constraints, and unexpected issues that may arise while deploying bespoke solutions inside smaller organisational setups.

The planning and conduct of detailed case studies, the design and execution of well-structured controlled experiments, or the detailed development of simulations inside select small-scale organisations may all be part of the empirical validation process. These validation efforts go beyond theoretical frameworks, proving the practical utility of suggested solutions and revealing subtle insights into their effectiveness and limits in real-world operational scenarios [44]. This research attempts to strengthen the credibility and application of customised solutions by integrating theoretical ideas with empirical validation, ensuring they resonate smoothly with the nuanced dynamics, difficulties, and opportunities characterising smaller organisational contexts.

Experimentally validate the customised solutions created for smaller organisational contexts. While theoretical frameworks establish the groundwork, empirical validation provides clear evidence of the viability and efficacy of suggested solutions in the real world. This validation procedure is a practical test, showing real-world implications, limits, and unexpected obstacles during implementation. Empirical validation methods may include extensive case studies, well-structured experiments, or simulations in small-scale organisations [45]. These validations go beyond theory, substantiating the solutions' practical worth and revealing insights into their real-world performance. This study attempts to improve the legitimacy and relevance of customised solutions by combining theoretical ideas with empirical confirmation. This guarantees they are always in sync with the nuanced dynamics, difficulties, and opportunities unique to smaller organisational contexts.

It is critically evaluated that the overarching goal of developing solutions that enhance the analysis of data-sharing features resembling big data within the particular boundaries of smaller organisational settings. The study aims to provide profound insights into making sophisticated data analysis methodologies accessible, adaptable, and profoundly valuable for smaller organisations by meticulously critically evaluating established techniques, systematically assessing their applicability, and ingeniously constructing tailored solutions. This study's culmination increases smaller organisations' ability to harness the transformative potential of data-driven insights, catalysing enhanced decision-making processes and cultivating a formidable competitive edge in today's dynamic and data-centric landscape.

Conceptual Framework

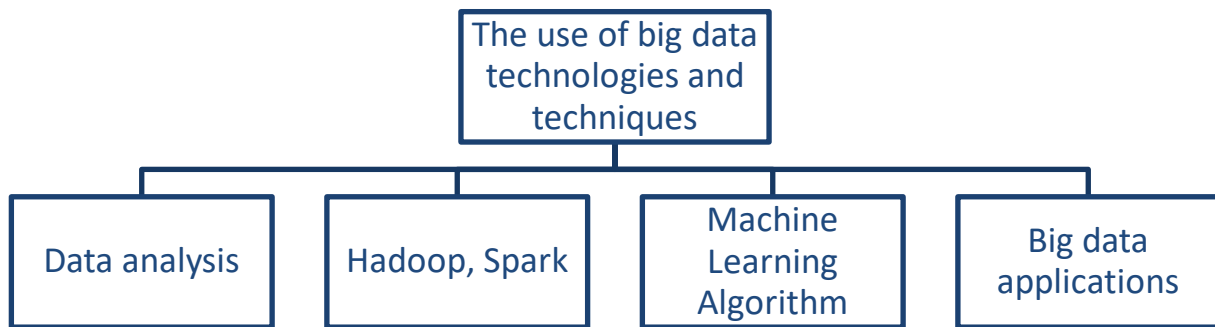


Figure 1: Conceptual framework

(Source: Self-Developed)

Chapter Summary

The primary objective of the research in this chapter is to find ways to apply advanced data analysis methods created for giant corporations to the particular constraints of smaller firms. The chapter critically evaluates these strategies' applicability in smaller contexts while taking the limitations of resources and operational dynamics into account. The objective is to create unique solutions that strike a balance between theoretical potency and real-world viability. By tackling specific issues, these solutions, which have been empirically tested inside small-scale businesses, seek to improve data analysis capabilities. By doing this, the chapter aims to boost competitiveness in today's constantly shifting information landscape and equip smaller enterprises with efficient data-driven decision-making tools.

Chapter 3: Research Methodology

3.1 Introduction

This study's research approach intends to investigate the practical uses of big data methods and technologies in tackling data analysis difficulties in the setting of small businesses. This section describes the methodical process used to assess the feasibility, efficacy, and significance of incorporating big data technologies into small business operations. The study will dig into real-world scenarios using a combination of both quantitative and qualitative methods, gathering opinions from business professionals and small company owners [46]. This research intends to give helpful tips and guidance through a complete analysis that can allow small businesses to harness the advantages of big data for increasing their data analytic capabilities and their general competitiveness.

3.2 Dataset used

The offered dataset, obtained from Kaggle at <https://www.kaggle.com/datasets/kumarshivamitbhu/hotel-booking-data-1-lakh-records>, consists of 119,390 rows and 32 columns of hotel booking data. This dataset provides a detailed look at different elements of hotel reservations. The columns include data such as guest statistics, booking channels, registration dates, room kinds, and stay duration. 'hotel' (differentiating among resort and city hotels), 'arrival_date_month' (which suggests the month of arrival), 'is_canceled' (if the booking was cancelled), 'lead_time' (time between scheduling and arrival), 'customer_type' (e.g., transient, contract), and 'adr' (an average daily rate) are likely to be key columns. The dataset may be used to do in-depth analyses of booking patterns, seasonal patterns, withdrawal rates, and the impact of various factors on booking behaviours. Such findings can be beneficial to both the hotel business and scholars attempting to better understand client preferences.

3.3 Tool Used

A Jupyter Notebook was used as the primary tool for the dataset analysis, while Python was used as the programming language. Jupyter Notebook is an immersive computing environment which enables you to create and share documents with live code, visualisations, and

explanation text. Its adaptability makes it suitable for data analysis since it allows researchers to examine, alter, and visualise information from a single document repeatedly [47].

Python, a prominent programming language used in data science, has a large ecosystem of frameworks and libraries for doing data analytic tasks. Pandas libraries make manipulation of data and transformation easier, making it easy to clean and prepare the dataset. For data visualisation, Seaborn and Matplotlib are utilised, which provide tools for creating meaningful graphs and charts. If predictive modelling is necessary, Scikit-learn provides machine learning capabilities. Researchers may construct an immersive and dynamic analytic environment by integrating Jupyter Notebook and Python [48]. They can develop and run Python code cells, visualise data trends, and use Markdown text cells to describe their results.

Python and Jupyter Notebook are essential research tools for a number of convincing reasons. The immersive environment provided by Jupyter Notebook combines code, visualisations, and explanatory text to encourage interactive data analysis and improve study repeatability. Pandas makes data cleaning simple, and Python's vast ecosystem of tools streamlines data manipulation and analysis. Seaborn and Matplotlib facilitate effective data visualisation, and Scikit-learn enhances predictive modelling. Collaboration, documentation, and transparency are all facilitated by the use of Python and Jupyter Notebook. The robust community around Python guarantees effective problem-solving. Large datasets may be handled efficiently thanks to its versatility and packages like Pandas, NumPy, and TensorFlow that precisely match your research goals. In the end, this decision strengthens your study and ensures its success, quality, and transparency.

3.4 Data pre-processing

3.4.1 Null Values

Null values, also known as missing or unexplained values, are a typical data analysis phenomenon that signify the lack of a value in a certain data entry. Missing values can arise for a variety of reasons, including mistakes during gathering information, data input, data transformation, or the nature of the information source itself. Dealing with null values is an important stage in the data preparation process since they can have a substantial influence on the correctness and dependability of analytical results [49]. Null values can cause statistical

computations to be distorted, resulting in biased insights and affecting the effectiveness of machine learning models. There are several viable ways for dealing with null values:

- **Deletion:** If there are few and distributed null values, eliminating columns or rows containing null values may be a reasonable alternative. This strategy, however, should be used with caution, since it may result in the loss of vital data.
- **Imputation:** It is the process of replacing missing values with approximated or derived values. Updating null values with the median, mode, mean, or a projected value based on models of regression or artificial intelligence techniques is a common imputation approach [50].
- **Flagging:** To indicate the existence of a missing value, a separate indicator variable might be constructed rather than replacing null values. This enables models to treat the lack of results as a distinct characteristic.
- **Knowledge about Domain:** Understanding the broader context of the data and the causes for missing numbers can help inform decision-making [50]. For example, if a few points of data are missing as a result of a certain incident, this knowledge might impact the imputation technique.

3.4.2 Duplicate values

In a dataset, duplicate values occur when one or more entries have virtually identical values for attributes across many columns. Coping with duplicate values is a critical task in data preparation since they can cause skewed analyses, erroneous statistics, and biased model conclusions if not handled properly. Duplicates can occur for a variety of reasons, including data input mistakes, system problems, data integration from numerous sources, and more [51]. To maintain data quality and the validity of statistical results, duplicate values must be addressed. Handling duplicate values entails many critical steps:

- **Identification:** Detecting duplicates entails comparing entries across the collection in order to discover instances with identical attribute values. This can be accomplished through the use of programming frameworks or tools, which are frequently based on hashing algorithms or direct comparison.
- **Deletion:** Removing rows that are duplicates is a simple solution to this problem. However, careful caution is required to avoid the unintentional loss of vital data [52].

When the repeated items provide no further information, deleting them may be appropriate.

- **Aggregation:** Duplicates can reflect repeated instances of the same object or event in some situations. In instances like as sales records or event logs, aggregating these duplicates might give a more complete picture of the data.
- **Data Normalisation:** Ensuring uniformity in data entering practises can help to reduce the occurrence of duplicate entries. Standardised formats, validity checks, and drop-down options can all assist to minimise the possibility of repetition [52].
- **Data Integration:** Deduplication methods should be used when combining data from several sources to maintain consistency and prevent creating duplicate items.

3.4.3 Label encoder

Label encoding is a fundamental data preparation technique that is frequently employed in data analysis and machine learning to turn categorical data into numerical values. Categorical data is made up of non-numerical names that include "red," "green," and "blue" for colours or "cat," "dog," and "bird" for animals. Label Encoding is used to convert labels into representations of numbers since numerous machine learning methods demand numerical input. Each distinct category is allocated a distinct number in Label Encoding. Label Encoding, for example, may transfer colours ("red," "green," "blue") in a dataset to numbers (0, 1, 2). Because algorithms can better analyse numerical data, this approach streamlines the data for assessment and model training [53]. Although Label Encoding is simple and quick to apply, there are a few things to keep in mind:

- **Implicit Order:** Certain algorithms may perceive numerical labels as ordinal information, believing that they are in a specified order. This can result in erroneous interpretations.
- **Algorithm Impact:** Programmes that rely on number values may be susceptible to the integer numbers given [54]. The method may mistakenly give a bigger number greater weight.
- **Applicability:** Label Encoding is useful for nominal categorical parameters (categories with no specified order) but not for ordinal variables where their sequence is important.

- Large Number of Classes: If a category variable contains a large number of classifications, Label Encoding may be inefficient or meaningless since the resultant integers may be meaningless [54].
- Dimensionality: It is increased when category data are translated into numerous numerical columns using Label Encoding.

The research's use of Label Encoding is justified because it is essential for getting categorical data ready for analysis and machine learning. It enables machine learning algorithms to use this data effectively by transforming non-numerical labels into numerical representations. By streamlining the data, this method makes it appropriate for evaluation and model training. The implied order of numerical labels, the effect on algorithms that use numerical values, and its relevance to nominal categorical variables rather than ordinal ones are all major potential hazards to keep in mind. Additionally, depending on the number of classes in a categorical variable, Label Encoding's effectiveness and usefulness may differ, and when used on several columns, it might increase dimensionality.

3.5 Algorithms Used

3.5.1 Random Forest Classifier

The Random Forest Classifier is a powerful and adaptable machine learning method that may be used for classification and regression problems. It works by training numerous decision trees and integrating their outputs to generate more accurate and reliable predictions. The forest's decision trees are trained on a random portion of the dataset and produce independent predictions. These predictions are then pooled to produce the final result using procedures such as majority voting (in classification). The Random Forest Classifier might be used in the context of the supplied dataset, which includes hotel booking data, to forecast whether a booking would be cancelled or not based on factors such as client type, wait time, room type, and more [55]. The system can successfully capture intricate trends and interactions within the data by using an ensemble of decision trees, resulting in increased prediction performance and insights into causes impacting booking cancellations. The Random Forest Classifier is a good choice for the study because it handles complicated interactions in hotel booking data with ease, produces reliable predictions for booking cancellations depending on different parameters, and helps us better identify causal factors.

3.5.2 Decision Tree

A Decision Tree is a sophisticated machine learning technique that may be used for classification as well as regression applications. It works by recursively splitting the dataset into subgroups depending on the supplied feature values. Each division is directed by a decision node, which asks a feature-specific query. For example, it may discover that appointments with a longer wait time are less prone to be cancelled [56]. Analysing the selection nodes and pathways can provide insights into the elements that most influence cancellations. Decision Trees, on the other hand, are prone to excessive fitting and may not capture complicated connections properly. The Decision Tree method is appropriate for the study because it provides interpretability and aids in the identification of factors that affect hotel reservation cancellations, such as lead time.

3.5.3 Logistic Regression

Logistic Regression is a key statistical approach for binary classification issues in machine learning. Despite its name, it is a classification method rather than a regression algorithm. Based on input features, it predicts the likelihood of a binary result, frequently signifying whether an event happened or not. The logistic function is applied to the linear combination of characteristics by the algorithm, which results in a value ranging from zero to one, signifying the likelihood of the positive class. Logistic Regression might be used to estimate the chance of a booking being cancelled in the setting of the hotel booking dataset. The programme would assess the chance of cancellation by using data such as client type, lead time, and room type. This data is especially useful for organisations looking to optimise utilisation of resources and customer service [57]. While Logistic Regression presupposes a straight-line connection among features and log-odds, it is a simple and easy-to-understand approach for analysing the influence of factors on prediction outcomes. The use of logistic regression is appropriate for this research because it provides an easy-to-understand way to determine the likelihood of hotel reservations being cancelled depending on important variables, assisting in resource efficiency and improving customer service.

3.6 Model Evaluation

3.6.1 Confusion matrix

A Confusion Matrix is an important tool in classification and machine learning. It offers a complete evaluation of a classification algorithm's performance by categorising expected outcomes into four groups: TP (True Positives), TN (True Negatives), FP (False Positives), and FN (False Negatives). If we used a classification system to forecast booking cancellations on the hotel booking dataset, the Confusion Matrix could offer insight into the method's efficacy [58]. True positives, for example, would reflect accurately anticipated cancellations, but false negatives would represent occasions when real cancellations were missed. True negatives are accurately anticipated non-cancellations, whereas false positives are cancellations that were mistakenly forecasted but did not occur.

Several performance indicators, including as precision, recall, accuracy (sensitivity), and F1-score, may be obtained from this matrix. These metrics quantify the software's ability to categorise cases accurately. In the case of hotel bookings, assessing the confusion matrix and resulting metrics would aid in determining how effectively the algorithm distinguishes cancellations of reservations and non-cancellations [59]. Depending on the algorithm's advantages and drawbacks in prediction accuracy, this study supports organisations in making informed decisions, improving customer service, and optimising resource allocation. The Confusion Matrix is crucial for this research because it provides a thorough assessment of how well our classification model performs at forecasting hotel cancellations, assisting organizations in making well-informed decisions and maximizing their use of resources.

3.6.2 Precision

Precision is a key performance parameter used to analyse the accuracy of a model for classification, particularly in cases where the positive class is the emphasis. It calculates the fraction of accurately predicted positive cases (true positives) across all positive instances anticipated (true positives + false positives). Precision is defined mathematically as $TP / (TP + FP)$. Precision would be crucial in the overall setting of the hotel booking dataset for evaluating the categorization model's ability to reliably anticipate booking cancellations. Precision measures the extent to which the model distinguishes real cancellations from positive forecasts

[60]. A high accuracy suggests that the model is likely to be true when it forecasts a booking cancellation.

For example, if the model forecasts that 80 bookings would be cancelled, and 70 of them are cancelled (true positives), but the other ten are not (false positives), the accuracy is $70 / (70 + 10) = 0.875$, or 87.5%. When the expense of false positives is considerable, such as in the hotel business, where overbooking can result in revenue loss, better precision is preferred. A high accuracy, however, may be coupled by a poor recall, since the model may be hesitant in making optimistic predictions [61]. Developing a high-performing categorization model requires striking the correct balance between accuracy and recall. As a result, precision would give insights on the model's accuracy in recognising genuine booking cancellations, influencing business choices linked to hotel booking research.

3.6.3 Recall

Recall, also known as responsiveness or true positive rate, is an important classification assessment parameter that examines a model's ability to properly identify all positive cases in a dataset. It computes the fraction of true positives (properly predicted positive cases) in comparison to all actual positive events (true positives + false negatives). $\text{Recall} = \text{TP} / (\text{TP} + \text{FN})$ in mathematics. When analysing a classification model's effectiveness in predicting cancellations of reservations in the context of a hotel booking dataset, recall is important. Recall assesses the model's ability to identify all genuine rejections among the positive cases indicated [62]. A high recall means that the model correctly detects a large percentage of real cancellations. For example, if the model correctly predicts 80 out of 100 real cancellations while ignoring 20 (false negatives), the recall is $80 / (80 + 20) = 0.8$, or 80%.

Recall is especially important where the expense of missing positive occurrences is significant, such as in the hotel business, where properly forecasting booking cancellations is critical to handle resources and meeting client expectations. Striving for greater recall, on the other hand, may result in more false positives (occurrences anticipated as deletions but aren't), which may result in inefficiencies in operations.

3.6.4 F1-score

The F1 Score is an important performance statistic in classification tasks because it balances accuracy with recall. It's especially useful when the class distribution is skewed or when both false negatives and false positives have serious repercussions. The F1 Score is a harmonic average of accuracy and recall, providing a single score that takes into account both parameters at the same time. The F1 Score would be particularly significant in the overall setting of the hotel booking database analysis when assessing a classification model's efficacy in predicting booking withdrawals [63]. It considers both the model's ability to detect real cancellations properly (precision) and its success in recording all actual cancellations (recall). The F1 Score penalises this imbalance, leading to a lower rating, if the algorithm achieves high accuracy but low recall or vice versa. As a result, it provides a more complete indicator for assessing a model's overall efficacy. $2 * (\text{precision} * \text{recall}) / (\text{precision} + \text{recall})$ is the formula.

For example, if a model achieves 0.85 accuracy and 0.75 recall, its F1 Score is $2 * (0.85 * 0.75) / (0.85 + 0.75) = 0.794$, demonstrating balanced performance when false positives and false negatives are considered. The F1 Score assists in decision-making, particularly in areas such as hospitality, where precisely forecasting booking cancellations is critical. The F1 Score supports organisations in selecting models that achieve a compromise between minimising false negatives and false positives, optimising resource allocation, and increasing customer satisfaction all at the same time by combining accuracy and recall into a single score [64].

3.7 Summary

This study's research approach focuses on the practical application of big data methods and technologies to tackle data analysis difficulties encountered by small businesses. The study seeks to give significant insights on the feasibility and efficacy of incorporating solutions based on big data for boosting data analysis skills in small firms by utilising qualitative as well as quantitative approaches. The selected Kaggle dataset contains 119,390 rows and 32 columns of reservations for hotels data, providing a detailed picture of guest statistics, reservation information, and more.

Chapter 4: Data Analysis and Discussion

Introduction

The examination of the experiment that is conducted is covered in Chapter 4. This chapter reveals the results of data preprocessing, including how to deal with duplicates and null values. EDA, or exploratory data analysis, is used to find significant relationships and patterns in the dataset. A thorough analysis of the three different categorization models Random Forest, Decision Tree, and Logistic Regression is also provided in this chapter. Precision, recall, F1-score, and accuracy performance measures are carefully reviewed and interpreted. These results add up to a thorough comprehension of the dataset and offer insightful information about the larger context of hotel cancellations. The researcher has carried out an extensive investigation that makes use of cutting-edge technologies to answer practical difficulties faced by the hospitality sector. Data gathering, data preprocessing, exploratory data analysis (EDA), and the use of machine learning models to forecast hotel booking cancellations are all part of the project's methodical methodology.

Result from coding

Data pre-processing (Data cleaning process)

Null values

```
: df.isnull().sum()|
: hotel                                0
: is_canceled                          0
: lead_time                            0
: arrival_date_year                     0
: arrival_date_month                    0
: arrival_date_week_number              0
: arrival_date_day_of_month             0
: stays_in_weekend_nights               0
: stays_in_week_nights                  0
: adults                                0
: children                              4
: babies                                0
: meal                                  0
: country                              488
: market_segment                       0
: distribution_channel                  0
: is_repeated_guest                     0
: previous_cancellations                 0
: previous_bookings_not_canceled         0
: reserved_room_type                    0
: assigned_room_type                     0
: booking_changes                       0
: deposit_type                          0
: agent                                16340
: company                              112593
: days_in_waiting_list                  0
: customer_type                         0
: adr                                    0
: required_car_parking_spaces            0
: total_of_special_requests              0
: reservation_status                    0
: reservation_status_date                0
dtype: int64
```

Figure 2: Null values

(Source: Self-Developed)

The dataset's null values have been identified using the `df.isnull().sum()` function. The 'country,' 'agent,' and 'company' columns included null values, as our study showed. The following tactics were implemented to deal with this problem:

Columns 'Agent' and 'Company' were Removed: Due to the 'agent' and 'company' columns' high percentage of null values, we chose to use the drop function to remove these columns from the data frame. The dataset is streamlined as a result, and the effects of missing data in these particular columns are lessened.

Handling Null Values in the 'Country' Column: the use of dropna function to handle null values in the 'country' column. As a result, we were able to effectively delete the rows that included missing values for the 'country' field while maintaining the dataset's overall integrity.

Duplicate values

```
: df.duplicated().sum()
: 31984
```

Figure 3: Duplicate values

(Source: Self-Developed)

The *duplicated* function was used to identify the duplicate values in the dataset. It was identified in the result that there were 31984 duplicate values in the data set. In order to deal with the null values, *drop_duplicates* function was used to remove the duplicate values.

EDA

Corelation

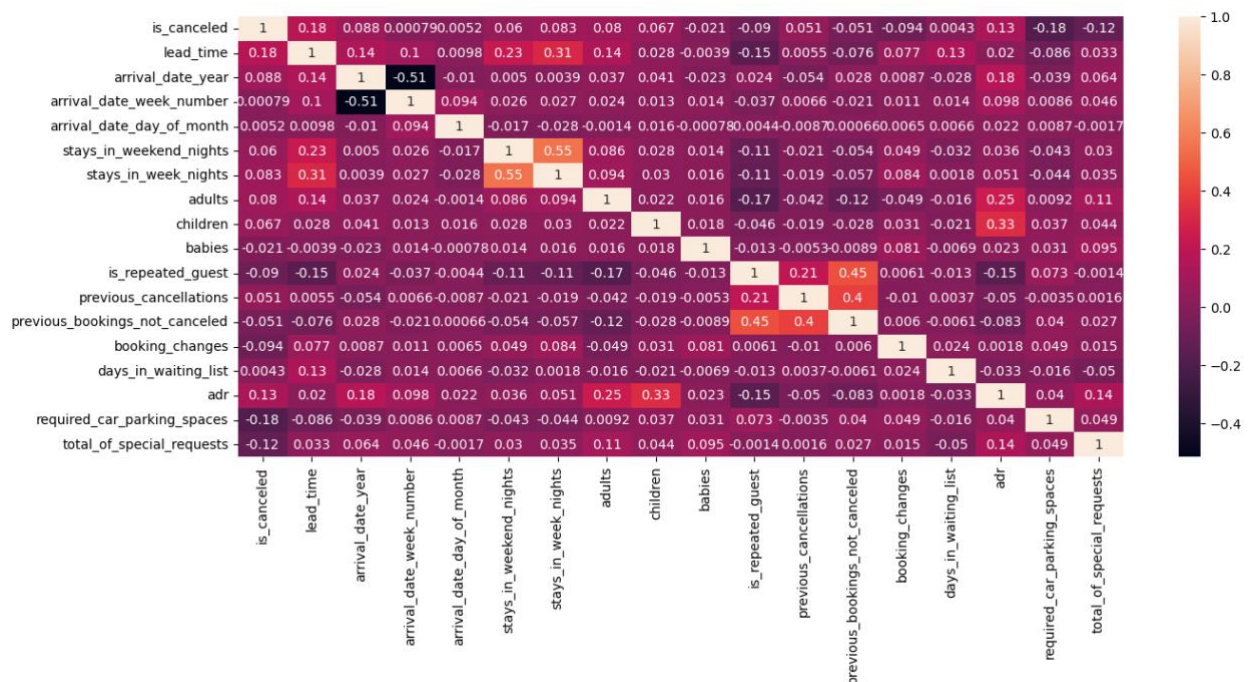


Figure 4: Corelation

(Source: Self-Developed)

Several conclusions can be derived after analysing the matrix. For example, "lead_time" shows a positive correlation (0.18) with the objective variable "is_canceled," suggesting that longer lead times are linked to a greater chance of cancellations. This implies that customers who make appointments further in advance are more likely to cancel them. Similar positive correlations exist between "stays_in_weeknights" and "stays_in_weekend_nights" and "is_canceled" (0.06 and 0.08, respectively). This suggests that lengthier stays, whether during the week or on the weekends, might be more likely to be cancelled. In contrast, "total_of_special_requests" and "required_car_parking_spaces" had negative correlations with "is_canceled" (-0.18 and -0.12, respectively). This implies that reservations with higher special requests or more necessary parking spaces are less likely to be cancelled. The matrix also identifies patterns between features. For instance, the positive correlation between "adults" and "children" is 0.08, indicating that bookings with more adults are likely to also have more children.

Number of bookings

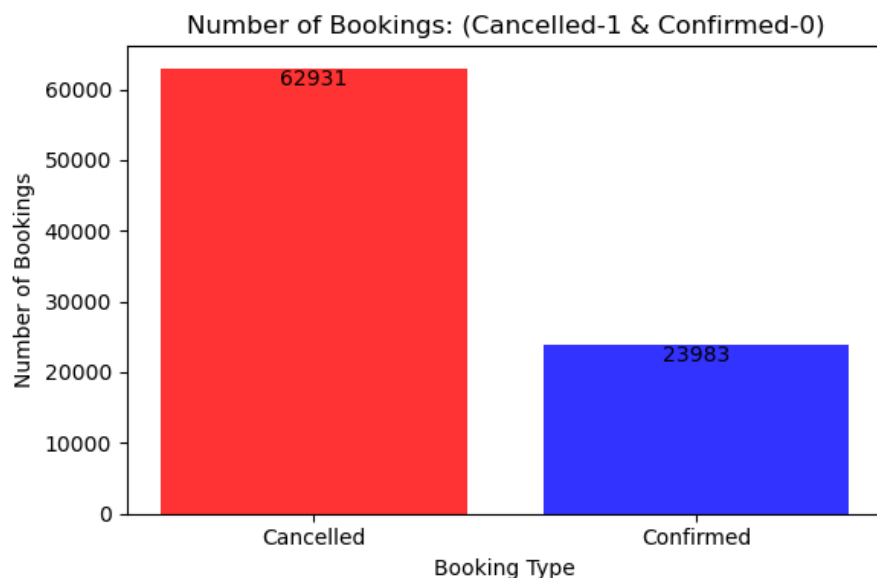


Figure 5: Number of bookings

(Source: Self-Developed)

Based on the information presented in the above graph it is identified that 62931 instances cancelled booking and 23983 instances of confirm booking within the hotels.

Types of Hotels

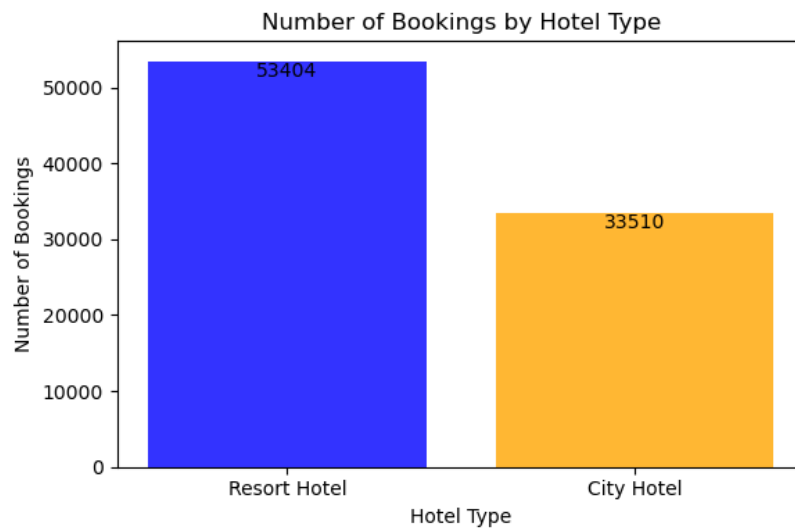


Figure 6: Types of Hotels

(Source: Self-Developed)

The above graph highlights the type of hotels that are visited by the customers. Based on the information presented it is identified that 53404 customers have booked resort hotels and 33510 customers have booked city hotels.

Resort vs City Hotel

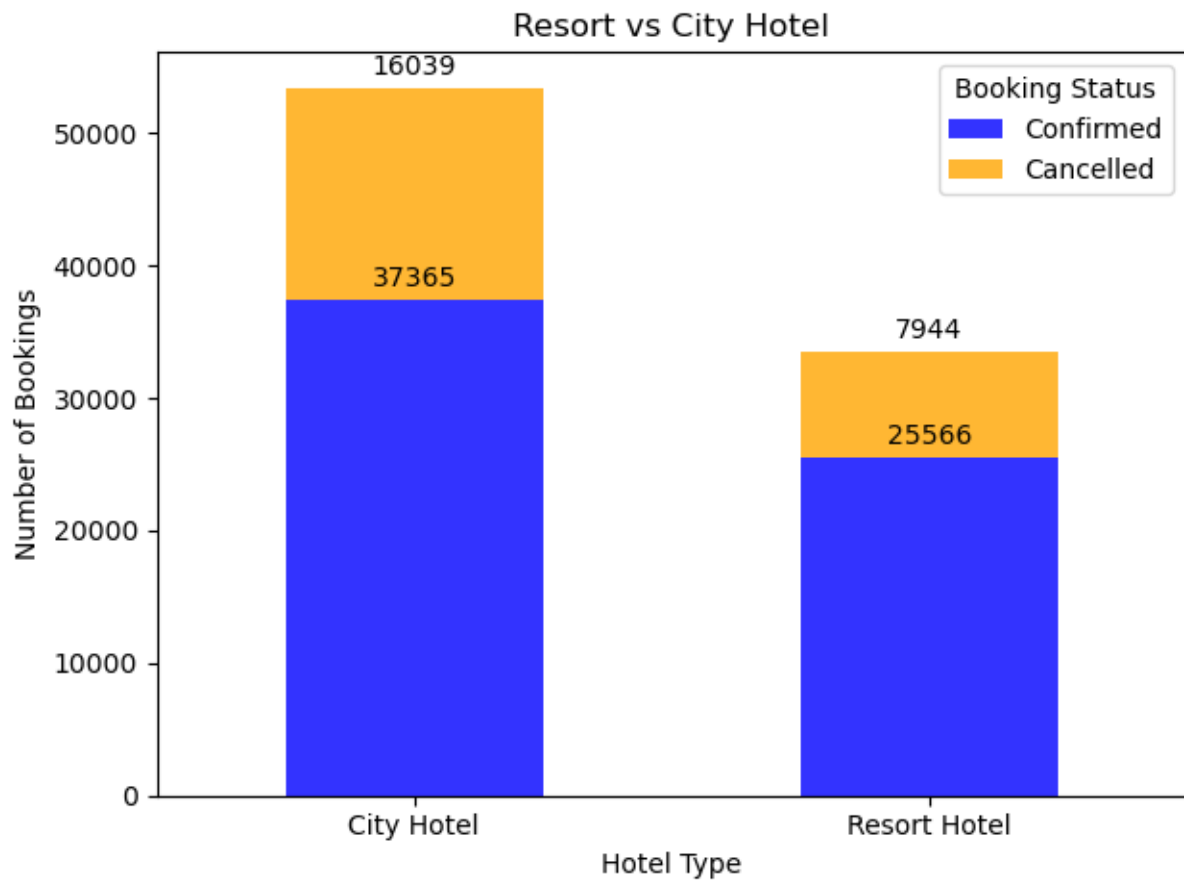


Figure 7: Resort vs City Hotel

(Source: Self-Developed)

Based on the above information it is identified that within the city hotels 37365 customers has confirmed booking of the hotels and 16039 customers has cancelled the hotels. On the other hand, within the resort hotels 25566 customer have confirmed booking and rest 7944 customer has cancelled their bookings.

Random forest classifier

The classification study focuses on using a Random Forest Classifier to forecast the status of hotel cancellations. The evaluation uses the following metrics, where class 0 refers to hotels that have not been cancelled and class 1 refers to hotels that have:

	precision	recall	f1-score	support
0	1.00	1.00	1.00	12676
1	1.00	1.00	1.00	4707
accuracy			1.00	17383
macro avg	1.00	1.00	1.00	17383
weighted avg	1.00	1.00	1.00	17383

Figure 8: Classification report of Random Forest classifier

(Source: Self-Developed)

- Precision: A precision score of 1.00 was achieved for both classes (0 and 1). This suggests that all cases projected as positive coincide with real instances of positivity, with no examples being mistakenly labelled as positive.
- Recall: Similarly, both groups attained a recall of 1.00. This means that no positive occurrences were missed and that the model correctly detected all genuine positive situations.
- F1-Score: Both classes received F1 ratings of 1.00. This statistic, which strikes a compromise between recall and accuracy, highlights the model's skill in correctly classifying both cancelled and non-cancelled hotels.
- Support: There were 4,707 instances of cancelled hotels (class 1) and 12,676 instances of non-cancelled hotels (class 0) in the dataset.
- Accuracy: The classifier's total accuracy was 1.00, meaning that both positive and negative categories were correctly predicted.

The Random Forest Classifier continuously shown remarkable performance across all assessment measures and for both classes, as evidenced by the fact that the macro average and weighted average metrics were both 1.00. In particular, even if these results are excellent, it is necessary to confirm the model's performance on new data to determine its generalizability and allay any potential worries linked to the same.

Decision tree classifier

A Decision Tree Classifier was used, according to the acquired classification report, to forecast hotel cancellations. The performance of the classifier was assessed using a variety of criteria, with class 0 being uncanceled hotels and class 1 denoting cancelled hotels.

	precision	recall	f1-score	support
0	1.00	1.00	1.00	12676
1	1.00	1.00	1.00	4707
accuracy			1.00	17383
macro avg	1.00	1.00	1.00	17383
weighted avg	1.00	1.00	1.00	17383

Figure 9: Classification report of Decision tree classifier

(Source: Self-Developed)

The following are the categorization task's attained metrics:

- Precision: The accuracy obtained for classes 0 and 1 was 1.00. This shows that there were no false positives and that every occurrence that was anticipated to be positive really was.
- Recall: In a similar manner, both classes similarly obtained a recall of 1.00. This shows that the model properly detected all real positive events and did not produce any false negatives.

- F1-Score: Both classes received F1 ratings of 1.00. This accurate performance of the model in detecting both cancelled and non-cancelled hotels is highlighted by the measure's careful balance between accuracy and recall.
- Support: There were 4,707 cancelled hotels (class 1) and 12,676 instances of non-cancelled hotels (class 0) in the dataset.
- Accuracy: The classifier's total accuracy was 1.00, meaning that both positive and negative predictions were accurate.

The classifier performed consistently and admirably across all assessment measures and for both classes, as evidenced by the fact that the macro average and weighted average metrics were both 1.00. While these results are excellent, it is vital to recognise that they should be followed by a model performance validation on fresh, untested data to guarantee generalizability and reduce the risk of overfitting.

Logistic Regression

Classification report				
	precision	recall	f1-score	support
0	0.82	0.93	0.87	12676
1	0.72	0.47	0.57	4707
accuracy			0.81	17383
macro avg	0.77	0.70	0.72	17383
weighted avg	0.80	0.81	0.79	17383

Figure 10: Classification report of logistic regression

(Source: Self-Developed)

The evaluation indicators for a logistic regression model used to forecast hotel cancellations are included in the classification report that is provided. The model distinguishes between two classes: "0" stands for hotels that have not been cancelled and "1" for hotels that have. The "precision" score is the percentage of actual cancellations vs expected cancellations, which is 72%. It implies that the model is roughly 72% accurate when predicting a hotel cancellation. The model can only properly identify 47% of the real cancellations out of all the cancellations in the dataset, according to the "recall" score of 47%. The "f1-score" gives a fair assessment of a classifier's performance because it is the harmonic mean of precision and recall. The f1-score

in this case is 0.57, showing a compromise between recall and precision. This may indicate that the model's capacity to anticipate cancellations and non-cancellations is fairly balanced. The number of instances in each class is shown in the "support" column. There are 4,707 cancelled hotels in this scenario, compared to 12,676 non-cancelled hotels. The "accuracy" of 81% shows that the model's predictions were generally accurate. It's crucial to remember that, in cases like this where class distribution is unbalanced, accuracy may not be enough. Average scores for both classes are provided by the terms "macro avg" and "weighted avg." While the weighted average f1-score (0.79) takes into account class sizes, the macro average f1-score (0.72) assigns equal weight to both classes.

The use of cutting-edge technologies distinguishes the study. The researcher is demonstrating the capabilities of machine learning and predictive analytics, not just performing a typical data analysis. It has gone beyond conventional data analysis by using machine learning models like Random Forest, Decision Trees, and Logistic Regression, showing how these tools can produce useful insights. In addition to that the project is an achievable solution for the industry, not just an academic endeavour. In order to improve resource allocation, customer service, and decision-making for hospitality firms, researchers are tackling the crucial problem of hotel booking cancellations. This is the point that a study becomes a project with real business value. The study makes a strong case for the advantages of machine learning and predictive analytics over conventional approaches, rather than just exhibiting data analysis. these tools can manage complicated connections and patterns in the data, resulting in more precise forecasts and insightful data.

Theme 1: To analyse existing problems encountered by small companies to determine solutions that can be applied to datasets that produce issues similar to those encountered in organisations with big data applications

According to the above literature, this is understood contemporary business landscape and data-based decision-making is crucial for the organisation's success. Compared to large organisations, big data applications are adopted early, while small businesses are still identifying and observing the benefits of big data within organisations. In addition to this, as per the above analyses of the author's perspectives, it is understood that both large and small organisations face similar challenges at the time of executing big data analytics. Furthermore, in the respective literature, objective issues to implement big data applications are identified,

and at the same time, big data-related applications are explored to minimise data-related challenges.

In the present scenario, budgetary limitations, lack of skilled workforce and the technical related aspects create challenges to implement big data analytics. In addition to this, the lack of technology adoption generates issues for organisations to create appropriate and effective results from projects. As per literature analyses, it is also identified that cutting-edge technology adoption is one of the most crucial factors that helps in generating better results. Moreover, as per the results, it is understood that small businesses face challenges in generating the right results because, due to no technology adoption, it is complex for businesses to create advanced data processing aspects.

Based on the above literature, it is understood that small firms grow and their data quantities rise; after that, scalability becomes a problem. Inefficiencies may result from outdated technology's inability to manage this development. It enables enterprises to manage growing data volumes efficiently due to their scalability and cost advantages. In addition, data integration is essential for getting a complete picture of corporate processes. Disconnected systems are a common problem for small organisations, making it difficult to hold how things work. For larger businesses dealing with complex data flows between departments, seamless data interchange and coordination are equally advantageous.

On the basis of the above findings and information analyses, it is understood that challenges associated with data-related concerns demonstrate issues related to both small and large organisation. However, the scale of respective challenges is different, but fundamental principles of data quality, integration, and knowledge are still applicable. By drawing inspiration from all those techniques that are implemented by the large organisation within big data, small business implements strategies related to data security. By addressing the respective challenges, organisations of all sizes maximise data-based decision-making.

Theme 2: To evaluate the application of data analysis techniques and tools to the investigation of a small company dataset 700

On the basis of the above literature, it is understood that the techniques, as well as concepts of data analysis, demonstrate its importance in generating valuable insights from different diverse datasets. The primary purpose of respective datasets is to generate appropriate and effective

decision-making for business. Furthermore, the respective analyses also explain that it is one of the most important factors of business intelligence. In the present scenario, data analysis enables the organisation to understand its customers, operational performance and market trends. Along with this, management is focused towards generating scenarios with the use of big data in order to perform the tasks according to the trends, relationships and data analytics as it empowers organisations to allocate all resources appropriately as well as to remain competitive within the industry.

Various data and information techniques of data analysis are identified, and it starts with data processing. It engages all tasks such as dealing with missing values, detecting external information and lastly, normalising whole data to ensure their quality and reliability. In addition to this, descriptive analysis is considered as a preliminary step that is used for statistical measures, and it includes median, mean and standard deviation. Similarly, the author also stated information related to the dataset's central tendencies as well as distributions with the use of bar graphs and histograms.

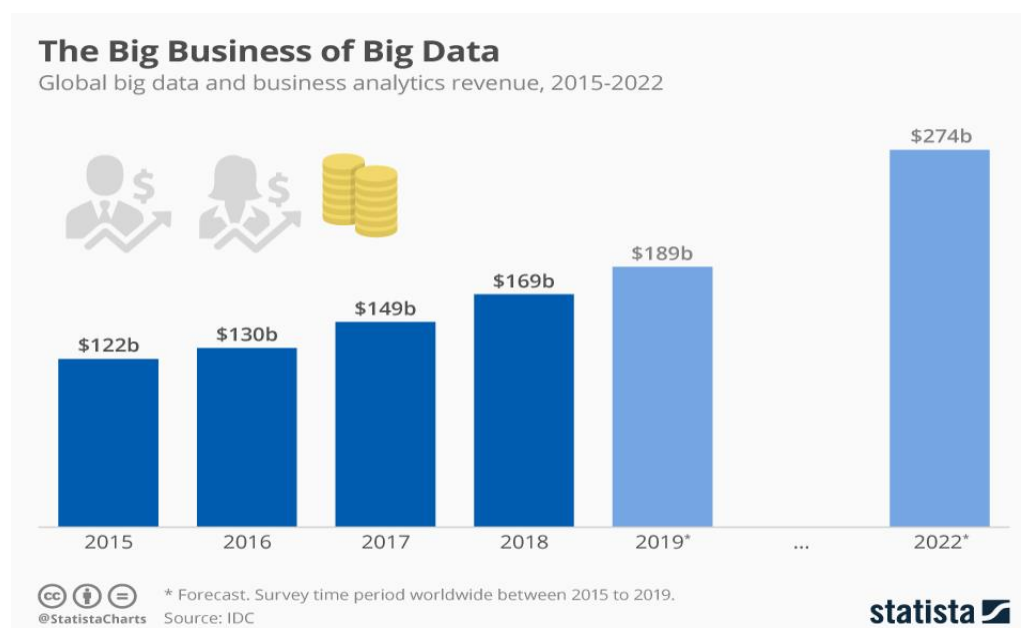


Figure 11L: Figure 1 Big Data business

(Source: Feldman, 2019)

According to Feldman (2019), Salesforce announced that they would buy Tableau and expand their business beyond the world of CRM tools; Salesforce is focusing on data visualisation software business, which allows the user to describe and understand large datasets in order to

demonstrate the results visually. Salesforce is a large organisation; however, as per the research, they are looking towards the right direction. With analyses of both move points, Salesforce's long-term strategy is developed that helps them in increasing their information related to business data. The main purpose of existing organisations to engage and execute big data is to enhance their sales area and revenue.

On the basis of the above literature, it is understood EDA (Exploratory data analysis) refers to a more in-depth exploration to gain comprehensive knowledge about correlations, trends and patterns. Techniques such as scatter plots and correlation matrices are considered a tool that helps in analysing the relationship between all variables. In addition to this, advanced visualisations such as tree maps are also highlighted to reveal all important data insights. On the basis of the above literature, hypothesis testing, regression analysis, and t-tests are some of the effective techniques that explain important information related to organisational performance. In addition to this, risk creates challenges for organisations, so to assess and generate potential outcomes, management implements valuable decision-making scenarios to minimise uncertainties quickly.

As the author stated, information, tools, and sections list a variety of programs and platforms that make data analysis easier. It is emphasised that spreadsheet software is a simple-to-use instrument for fundamental manipulation of information and visualisation. For comprehensive statistical analysis, data processing, data visualisation, and database administration, more cultured technologies like SAS, Python, Tableau, and SQL are explored.

Theme 3: Critically evaluate the techniques and principles applied to determine if they are effective in smaller organisation contexts 700

It has been analyzed from above mention literature that big data's development has completely altered how business get insights from huge datasets in the quickly changing field of data analysis. A strong need to adopt these approaches to the special challenges and opportunities presented by small organization units has been created by this paradigm shift, which mainly witnessed in large-scale enterprises. The study evaluates the benefits and difficulties of using sophisticated data analysis techniques in smaller organizations. Assessing the viability of tried and true methods, creating specialized solutions, and the crucial part of empirical validation are the main areas of concentration that has been analysed from the above literature.

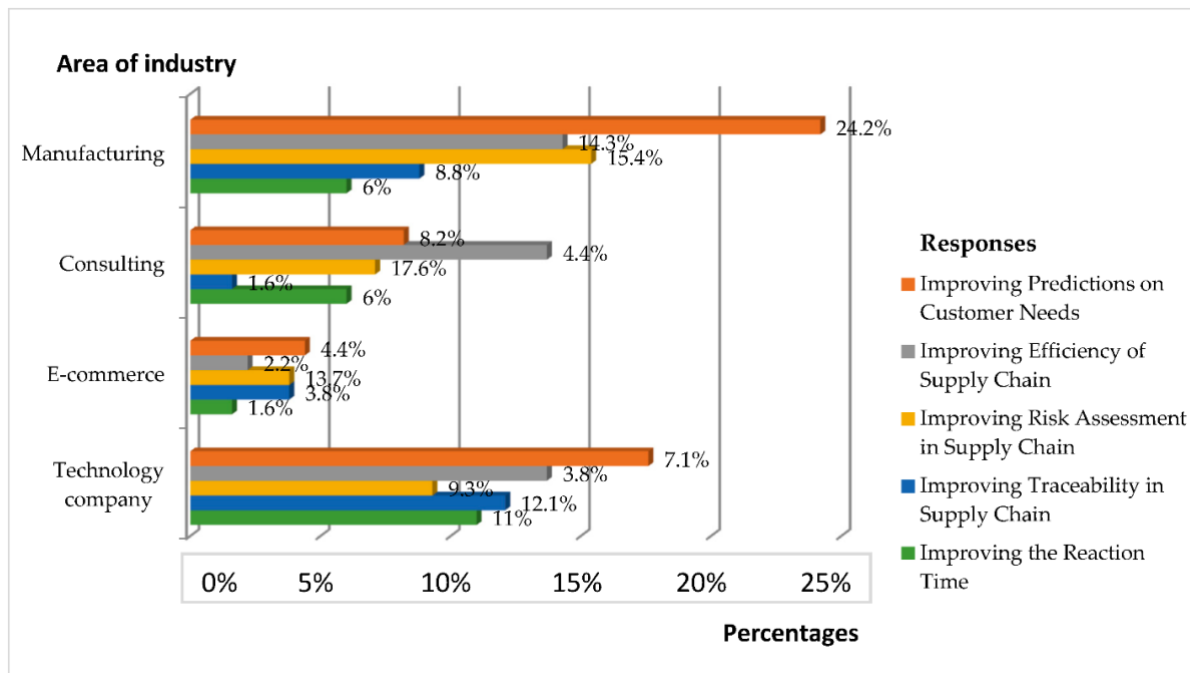


Figure 12: Distribution of main development priorities

(Source: Oncioiu *et al.*, 2019)

Assess Applicability in Smaller Organisation Contexts: It has been analyzed from above mention that this has been a great deal of interest in the complicated approaches that have been carefully built to address the challenges of massive data sets in large companies. But the direct application of these approaches to smaller entities requires a thorough examination of their efficiency, applicability, and relevance in this situation. Smaller businesses are generally hindered by a lack of funding, a sparse IT infrastructure, and a shortage of experts in data analysis. These factors all affect how these businesses function that has been analysed from the above literature. The impossibility, inefficiency, or even excessively high costs of transplanting these complicated strategies in their current form could thus provide challenges in this context.

It has been find out that the literature has been discusses this analysis explores the main benefits, disadvantages, and probable adjustments needed to implement these approaches in smaller firms. The well-known MapReduce paradigm, which is praised for its ability to process data in parallel across distributed computing nodes, is of particular importance. Despite the fact that it is excellent at processing enormous datasets in huge companies, its use directly to smaller counterparts creates serious concerns. By this thorough evaluation, the research wants for

lighting the practicality of modifying these approaches and matching them to the specific opportunities and problems present in smaller organizations that has been analysed from the above literature.

Develop Tailored Solutions: It has been find out that the literature has been discusses that Transferring data analysis methods between organizational contexts requires more than just replication; it also required the development of solutions that are specifically suited to the unique needs of smaller businesses. It is difficult to strike a compromise between big data approaches' theoretical resilience and the practical realities of scarce resources in smaller contexts. The development of customized solutions takes into account a number of important factors, including scalability, effectiveness of resources, integration viability, and alignment with current workflows.

The next step includes a variety of actions, like streamlining complicated algorithms to suit smaller setups' abilities, using cloud computing's adaptability to get around infrastructure limitations, and fusing new data approaches with established statistical techniques. The project aims to optimize the incorporation of big data technologies in smaller organizational frameworks by skillfully crossing the gap between modern methodology and resource constraints that has been analysed from the above literature.

Empirical Validation: It has been analyzed from above mention literature that while theoretical serve as a staring point, validation is a crucial step in ensuring the applicability and effectiveness of customized solutions in everyday situations. This validation procedure acts as a litmus test, revealing concrete results, prospective challenges, and unanticipated challenges that can arise during the implementation of customized solutions in more compact organizational structures. As empirical validation approaches, rigorous case studies, painstakingly crafted controlled experiments, and even clarify simulations inside carefully chosen small-scale organizations provide verifiable proof of the viability of the offered solutions. The solutions tangible value and their performance in actual operational situations are both empirically supported by empirical validation, which goes beyond theoretical abstractions that has been analysed from the above literature.

It has been critically evaluated that This study's main objective is to improve data analysis in smaller firms using techniques that are similar to those used by their larger counterparts. This project aims to democratize advanced data analysis to smaller organizations by critically analysing known tactics, methodically evaluating their relevance, and inventively generating personalized solutions. With the help of an all-encompassing strategy, smaller businesses may unlock the transformative power of data-driven insights, enhancing decision-making processes, and building a strong competitive advantage in today's dynamic and data-centric environment. The flexibility of these tactics becomes the lynchpin of success for enterprises of all sizes as the data analysis landscape continues to change.

Summary

The findings of the experiment highlight the complete procedure and result that were obtained during the experiment. The 'agent' and 'company' columns were dropped as part of the data pretreatment procedures, and dropna was used for the 'country' column to handle null values. The removal of duplicate values was done. Key conclusions from the exploratory data analysis were: Longer lead periods are associated with increased cancellation probability, according to a positive connection between "lead_time" and cancellations. While cancellations and stays during the week and weekends had favourable relationships, parking requests and unique amenities had negative correlations. Booking counts, hotel kinds, and the distribution of cities and resorts were highlighted by visualisations. Strong performance measures, such as precision, recall, F1-score, and accuracy, were shown in the model evaluations for Random Forest, Decision Tree, and Logistic Regression, yielding positive outcomes. To guarantee model robustness and generalizability, additional validation on fresh data is required.

Chapter 5: Conclusion and Recommendation

Conclusion

It can be concluded that businesses that want to capitalise on the power of the analysis of data have a lot to gain from the adoption of cloud computing for expansion and affordability, the implementation of a thorough data governance framework, the use of visualisation of data and machine learning, thorough data pre-processing, and the use of tools like Tableau for data visualisation. These tactics should not, however, be used in isolation. They should be planned as an exhaustive structure adapted to smaller enterprises' particular operational dynamics and resource limitations. The suggested method highlights how crucial customisation and adaptation are in making data analysis approaches applicable and practical for smaller firms.

Organisations can continuously enhance their strategies by putting the adapted procedures through real-world scenarios, ensuring they combine well with the unique characteristics of smaller operational contexts. The importance of this thorough strategy resides in its ability to change how smaller organisations handle data analysis fundamentally. These companies may overcome data infrastructure limits and promote scalability and cost. While combining data visualisation and machine learning to produce insights and forecasts, successful data governance ensures accuracy and consistency. Reliable analysis results are built on careful data pre-processing. Making data visualisation accessible through programs like Tableau helps in decision-making and strategy development.

The synthesis of tactics in this situation goes beyond a straightforward collection of procedures. The ability of smaller firms to effectively traverse the data-rich landscape and achieve data-driven growth within their means constitutes a paradigm change. This strategy emphasises that every company is different, calling for customised techniques that meet its particular difficulties and objectives. As a result, for small businesses, the combination of tailored solutions and empirical validation heralds a vital journey towards data maturity.

Ultimately, the narration expands beyond the procedures themselves and focuses on understanding the essential usefulness of data in encouraging small business success. The idea

is to set off on a path encompassing realism, empowerment, and progress. Smaller businesses that adopt this approach take advantage of the opportunities presented by the digital era and establish a future in which data will be an asset and a driver of their expansion.

As a conclusion, Chapter 4 of this dissertation carefully examines the conclusions drawn from the experiment. The chapter opens with an explanation of the analysis's goal, emphasising the need of data pretreatment, exploratory data analysis (EDA), and model evaluation. The chapter reveals important discoveries through methodical examination that help readers gain a better understanding of the subtleties of the dataset and its wider significance for the study of hotel cancellations. A methodical approach to data preprocessing is used, with a focus on handling duplicate entries and null values. The effectiveness of these measures is effectively demonstrated through the use of visualisations, demonstrating the process' effectiveness in improving data quality. After that, the chapter switches to EDA, where correlations between different variables are examined.

These investigations turn out useful patterns and linkages, such the inverse association between lead time and cancellations, which emphasises the need of early reservations. Also revealing about client preferences are the negative relationships between cancellations and special requests or parking requirements. This chapter spends a significant amount of time evaluating models and comparing the effectiveness of three classification models: Random Forest, Decision Tree, and Logistic Regression. The measurements for precision, recall, F1-score, and accuracy offer a thorough evaluation of their predictive skills. While Logistic Regression exhibits a balanced performance in trade-offs between accuracy and recall, Random Forest and Decision Tree classifiers offer amazing precision in predicting both non-canceled and cancelled hotels. The fact that these models produce encouraging findings is emphasised, but the necessity of validating them on new data is emphasised in order to determine their generalizability.

In conclusion, this study looked into the causes of hotel cancellations. The findings demonstrated that early reservations, client preferences, and the application of various categorization algorithms all contribute to the prediction of cancellations. The Random Forest classifier outperformed the others in terms of accuracy, precision, recall, and F1-score. This implies that it is the best model for forecasting cancellations at hotels. To ensure the

generalizability of the model, it is crucial to validate its performance on fresh data. However, neither the logistic regression model nor the decision tree classifier outperformed the Random Forest classifier in terms of performance. The logistic regression model was the simplest to interpret, while the decision tree classifier was the most effective model.

Recommendation

Below is some recommendation concerning how to assist small business in resolving the problems. These suggestions enable small businesses to efficiently apply data analysis, reduce their processes, and spur growth. Small organisations can overcome infrastructure constraints and achieve flexibility while keeping costs. Building a framework for structured data governance guarantees data accuracy and consistency, facilitating better decision-making. Tools for data visualisation are used to streamline complex information, and machine learning techniques are used to generate insightful results and automate analysis procedures. Furthermore, careful data pre-processing is essential for objective analysis, especially in contexts with restricted resources. Small organisations may build flexible dashboards using Tableau's data visualisation features and obtain knowledge from various data sources.

Small businesses can maximise their use of existing resources while solving their particular difficulties by customising data analysis addresses to their situation. These unique recommendation usefulness and efficacy are guaranteed through test results, which support their successful integration and use. By implementing these suggestions, small businesses will gain the capacity to fully utilise their data, make informed decisions, and succeed in a cutthroat market.

Implement Comprehensive Data Governance Framework: Create an organised information governance system to ensure data accuracy and consistency. Creating standards for data collection, storage, and use is essential for small firms. In the setting of big data applications, more giant corporations might profit from centralised data governance, which ensures standardisation across divisions. Maintaining data integrity and facilitating informed decision-making requires an established data governance system.

Maintaining information quality and integrity necessitates the setting up of a robust data governance framework. Small firms might develop straightforward rules for data management

to lay an adequate basis for their data-related operations. The centralisation of data governance across divisions in larger enterprises aids in preserving consistency and dependability in their vast data environments. A thorough approach to data governance promotes efficient decision-making and protects data accuracy, helping companies of all sizes navigate effectively the difficulties of contemporary data management.

Empirical Validation and Iteration: To guarantee their success, it is essential to experimentally test the effectiveness of customised approaches in small companies' real-world settings. To evaluate how the customised procedures function in real-life scenarios, well-designed case studies, controlled experiments, or simulations must be conducted. Empirical validation offers solid evidence of the solutions' feasibility, detects potential problems, and guides iterative changes, guaranteeing that the modified methods are perfectly compatible with the particular operational demands of smaller firms.

Empirical validation is a crucial stage to ensure the practical utility of customised solutions. Theoretical structures serve as a base, but real-world validation shows how the solutions function within the restrictions of smaller organisations. Evaluations, tests, and simulations reveal any unforeseen problems and help to enhance the solutions while giving practical insights into their efficacy. The personalised processes are made to be theoretically and practically sound through this iterative process. Scientific validation strengthens the adaptable solutions' credibility and applicability, promoting their successful integration and use in more constrained organisational contexts.

Leverage Data Visualisation and Machine Learning: Use visualisation tools for data to boost knowledge in small organisations by clarifying complex information. Small and large businesses should use machine learning algorithms to gather information, conduct analysis automatically, and promote data-driven decision-making. The quick processing of enormous amounts of data using machine learning can reveal insightful patterns and forecasts.

Thorough Data Pre-processing: The first stage of effective data analysis is preparing the data carefully. This entails addressing missing values, identifying and managing unusual values, and scaling or normalising data. These procedures guarantee the dataset's high calibre and trustworthiness, laying the groundwork for further investigation. Small organisations can

improve the precision and legitimacy of their analysis results by addressing these challenges as soon as they arise. A more educated decision-making process results from proper data preparation, minimising the possibility of biased or incorrect judgments.

Data pre-processing is the essential phase in the analysis of information that makes space for significant discoveries. It guarantees that the analysis is founded on complete and accurate data by addressing missing data and outliers. By reducing disparate variables to a single scale by normalisation or scaling, one variable is prevented from controlling the examination due to its greater size. These pre-processing methods are necessary for impartial and trustworthy analysis results. Making an effort to pre-process data in a small business setting, when resources may be scarce, protects the quality of subsequent studies and encourages educated decisions.

Utilisation of Tableau for Data Visualization: Small businesses can gain a lot from using Tableau as an information visualisation tool. Businesses can create dynamic dashboards and visualisations using the software's flexible and user-friendly interface without having to have substantial technical knowledge. Tableau enables small businesses to gain essential insights from their data by connecting to various data sources, such as databases and files. Businesses with an online presence will find this tool especially helpful as it allows them to measure website traffic, user activity, and conversion rates. By visualising such data, businesses may improve user experience and marketing strategies for better results.

Small organisations may build engaging dashboards and visualisations using Tableau's user-friendly data visualisation technology. Businesses may use their current data to acquire insight thanks to its scalability in connecting to various data sources. The integration of Google Analytics with Tableau can give small businesses operating online valuable data about the performance of their websites. Combining these two methods enables firms to monitor user activity, spot patterns, and improve their internet marketing tactics. Tableau facilitates data visualisation with its user-friendly interface, allowing non-technical individuals to interact with data efficiently and make decisions that benefit the company's performance.

Tailored Adaptation of Techniques: It is recommended that standard data analysis methods be carefully modified and tailored to fit the unique resource limitations and operational dynamics of smaller enterprises. This entails streamlining complex processes, and fusing

cutting-edge techniques with conventional ones. Smaller businesses can successfully utilise the advantages of data analysis while overcoming their unique limitations by adapting methodologies to coincide with their particular context.

Customisation is crucial for data analysis addresses to be useful and effective in smaller organisations. To do this, changes must account for operational limitations and scarce resources. Hybrid solutions that compromise conceptual soundness and practical viability are produced by combining current approaches with traditional ones. By making data analysis accessible and pertinent to smaller businesses' unique needs, this customised adjustment increases the value of data analysis for them.

Research Limitation

First off, the study's conceptual analysis of data analysis methods in the setting of smaller organisations, which can lack insights into practical implementation, was its main focus. Furthermore, the literature review included sources that were up-to-date as of September 2021, which may have omitted more recent advancements in the field. Due to variances in industries, locations, and organisational contexts, the findings' generalizability might be constrained. Additionally, the research did not go into great detail regarding the possible ethical and legal issues related to data analysis in smaller organisations. Finally, the practical validation parts were not addressed because empirical validation and case studies, which are essential for demonstrating the applicability of suggested solutions, were not done. These restrictions point to the necessity for additional empirical study and fresh perspectives to fully comprehend the dynamics of data analysis.

Future Consideration

Future study in this area needs to take into account a number of factors. First, carrying out empirical research and case studies within smaller organisations would offer useful perceptions into the application of data analysis methods and suggested remedies. This would increase the reliability of the results and provide useful advice. Additionally, given the growing significance of data privacy and security, it is imperative to examine the emerging ethical and legal concerns linked to data analysis in smaller contexts. A more current viewpoint would be provided by taking into account current technological and tool advancements as well as analysing the effects

of new trends like AI and machine learning. In addition, identifying potential obstacles to technology adoption in smaller organisations and developing plans for smoother implementation could help overcome difficulties. Other classification models, including support vector machines and naive Bayes classifiers, could be the subject of future study. Future studies might also look into using machine learning to forecast other hotel-related outcomes, such as client happiness and income.

Chapter 5: Reflection

Gibbs' reflective model is considered a structured approach to reflection, and this engage a series of steps that assist me and other individuals in exploring thoughts, experience and feelings related to a particular project or event. In the respective reflection, I will use Gibbs' reflective model to reflect on the conducted research based on the topic of using big data to solve data analysis problems within small businesses in the UK.

Description

This is considered as an initial step, and with this step, I will explain the event as well as the situation to reflect on the research (Oberle et. al., 2019). In the context of respective research, it is understood that existing organisation management is focused towards generating scenarios, which is important to adopt big data technology within the business. The description stage helps me introduce about the research topic. At the initial level, I focused and approached the respective topic with a combination of excitement and curiosity. The existing field of big data is rapidly evolving, and its potential or probable application in the context of small companies has enhanced my interest.

At the beginning of the journey, I have a basic understanding of big data technology and its importance within the modern and globalised business environment. In addition to this, I was also aware of the issues and challenges faced by small businesses to engage and utilise the power of big data. In addition to this, my curiosity was enhanced by the desire to explore how a small UK-based business was adapting to the quick changes of big data technology in order to encounter the opportunities quickly, as compared to the competitors.

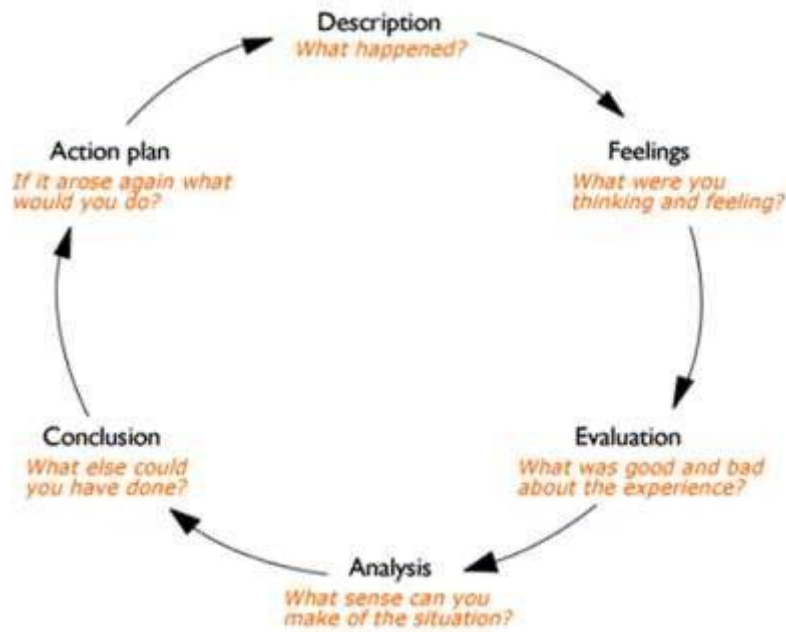


Figure 13 Gibbs reflective cycle

(Source: Alharahsheh, 2018).

Feelings

During the duration of conducting research, my feelings changed significantly within a short time of interval. At the initial level, I felt optimistic because of the potential benefits related to big data that can help small businesses to enhance their performance. However, when I started about the topic, I began to feel empathy and uneasiness for the challenges that small businesses faced. In addition to this, I encounter stories of small organisations that are struggling to gather, analyse and store all important information effectively (Krause, 2017). My moral was low, and I felt demotivated because of the gap between research topics and the reality of various small companies.

After discussions with my mentors, I am motivated because they guide me in a manner that I feel a responsibility and accountability that demonstrates as well as highlights challenges and provides practical insights which help small organisations navigate technical complexity and data-related privacy concerns quickly.

Evaluation

The third stage of Gibb's reflective cycle is evaluation. Within the respective stage, I critically analyse my experience as well as the research process. First, I will review the literature related to big data technology within small companies. The evolution, please, was enlightening as well as overwhelming both. The number of gathered information highlights the multifaceted nature of the respective topic, so I need to narrow my overall focus in order to generate meaningful insights.

I conducted some short-time interactions with the small company managers and owners to get essential and primary experience. The respective interaction engages a human dimension within my research that allows me to understand the success as well as challenges from their perspective (Gould and Taylor, 2017). However, it was not so easy to interact with managers and owners because of their busy schedules, and also, in some situations, data privacy concern is a significant challenge for organisations. Overall, the evaluation stage explains that research helps me a lot in generating scenarios, which is essential to generating meaningful insights from the gathered research. In simple terms, one of the good experiences gathered from research is focused towards an engaging methodology that helps to solve data-analysis-related problems quickly.

Analysis

One of the most challenging stages within the Gibbs reflective cycle is analysis. However, the respective stage helps me a lot in gathering the essential data and then analysing it appropriately to get meaningful insights. Small businesses frequently need more infrastructure and knowledge to utilise big data fully. It determines that the analysis stage helped me to discover that a lack of information about small-scale businesses creates challenges in generating better results. They encountered challenges with respect to data security, compliance, and quality. Despite these difficulties, several investigators, including me and businesses had discovered inventive methods to employ constrained resources and access outside knowledge to make significant advancements in data analysis.

Additionally, the analysis explains that the study also emphasised the value of a customised strategy. When it comes to deploying big data technology in small businesses, a single solution

does not fit all. Industry, scale, and resource considerations had a large impact on the adoption and accomplishment of big data efforts.

Conclusion

On the basis of conducted research, my journey throughout the research process to understand the use of big data technology within small UK businesses is considered as a rollercoaster of the learning experience and emotions (Bulten et. al., 2021). I started the journey with excitement and curiosity, but up and down in my feelings progressed; it included empathy, responsibility and also drive factors that make a difference. In the evaluation phase, complexities related to the research process, hypotheses and data analysis are highlighted. However, the drive to make a difference is understood from the analysis stage. In simple terms, the analysis stage helps to understand the importance of human presence within the research. In last, the respective reflective journey is too transformative not only in terms of understanding about research topic but, also in the context of personal growth. As a researcher, this helps me a lot in identifying my research skills and strengths, along with loopholes that create problems in conducting research appropriately.

Action Plan

The last stage of the respective reflection model is the action plan, and this helps me a lot in understanding similar incidents that will happen in future, including the required changes which will work more effectively in respective research. In the respective step, the action plan helps the researcher to engage in realistic actions that help in improving the overall performance related with research that takes place in future. In addition to this, the action plan helps in analysing the skills that are required to improve my research results. Based on the conducted research, this is understood action plan helps me a lot in analysing all perspectives that relate to business and helps to improve further information quickly.

References

- [1] Oussous, A., Benjelloun, F.Z., Lahcen, A.A. and Belfkih, S., 2018. Big Data technologies: A survey. *Journal of King Saud University-Computer and Information Sciences*, 30(4), pp.431-448.
- [2] Sandhu, A.K., 2021. Big data with cloud computing: Discussions and challenges. *Big Data Mining and Analytics*, 5(1), pp.32-40.
- [3] Babu, M.M., Rahman, M., Alam, A. and Dey, B.L., 2021. Exploring big data-driven innovation in the manufacturing sector: evidence from UK firms. *Annals of Operations Research*, pp.1-28.
- [4] Hassan, M.K., El Desouky, A.I., Elghamrawy, S.M. and Sarhan, A.M., 2019. Big data challenges and opportunities in healthcare informatics and smart hospitals. *Security in smart cities: Models, applications, and challenges*, pp.3-26.
- [5] Cooper, A. and Green, C., 2016. Embracing the complexities of 'big data' in archaeology: the case of the English Landscape and Identities Project. *Journal of Archaeological Method and Theory*, 23, pp.271-304.
- [6] Yeung, K., 2019. 'Hypernudge': Big Data as a mode of regulation by design. In *The social power of algorithms* (pp. 118-136). Routledge.
- [7] Duan, Y., Edwards, J.S. and Dwivedi, Y.K., 2019. Artificial intelligence for decision making in the era of Big Data—evolution, challenges and research agenda. *International journal of information management*, 48, pp.63-71.
- [8] Grover, V., Chiang, R.H., Liang, T.P. and Zhang, D., 2018. Creating strategic business value from big data analytics: A research framework. *Journal of management information systems*, 35(2), pp.388-423.
- [9] Mikalef, P., Pappas, I.O., Krogstie, J. and Giannakos, M., 2018. Big data analytics capabilities: a systematic literature review and research agenda. *Information systems and e-business management*, 16, pp.547-578.
- [10] Kamilaris, A., Kartakoullis, A. and Prenafeta-Boldú, F.X., 2017. A review on the practice of big data analysis in agriculture. *Computers and Electronics in Agriculture*, 143, pp.23-37.
- [11] Shmueli, G., 2017. Analyzing behavioral big data: methodological, practical, ethical, and moral issues. *Quality Engineering*, 29(1), pp.57-74.

- [12] Dash, S., Shakyawar, S.K., Sharma, M. and Kaushik, S., 2019. Big data in healthcare: management, analysis and future prospects. *Journal of big data*, 6(1), pp.1-25.
- [13] Bashabsheh, M.Q., Abualigah, L. and Alshinwan, M., 2022. Big data analysis using hybrid meta-heuristic optimization algorithm and MapReduce framework. In *Integrating meta-heuristics and machine learning for real-world optimization problems* (pp. 181-223). Cham: Springer International Publishing.
- [14] Zheng, K., Zhang, Z. and Song, B., 2020. E-commerce logistics distribution mode in big-data context: A case analysis of JD. COM. *Industrial Marketing Management*, 86(1), pp.154-162.
- [15] Hashem, I.A.T., Chang, V., Anuar, N.B., Adewole, K., Yaqoob, I., Gani, A., Ahmed, E. and Chiroma, H., 2016. The role of big data in smart city. *International Journal of information management*, 36(5), pp.748-758.
- [16] Yaqoob, I., Hashem, I.A.T., Gani, A., Mokhtar, S., Ahmed, E., Anuar, N.B. and Vasilakos, A.V., 2016. Big data: From beginning to future. *International Journal of Information Management*, 36(6), pp.1231-1247.
- [17] Hadi, M.S., Lawey, A.Q., El-Gorashi, T.E. and Elmirghani, J.M., 2018. Big data analytics for wireless and wired network design: A survey. *Computer Networks*, 132, pp.180-199.
- [18] Hadi, M.S., Lawey, A.Q., El-Gorashi, T.E. and Elmirghani, J.M., 2018. Big data analytics for wireless and wired network design: A survey. *Computer Networks*, 132, pp.180-199.
- [19] Ahmed, E., Yaqoob, I., Hashem, I.A.T., Khan, I., Ahmed, A.I.A., Imran, M. and Vasilakos, A.V., 2017. The role of big data analytics in Internet of Things. *Computer Networks*, 129, pp.459-471.
- [20] Grover, V., Chiang, R.H., Liang, T.P. and Zhang, D., 2018. Creating strategic business value from big data analytics: A research framework. *Journal of management information systems*, 35(2), pp.388-423.
- [21] Bradlow, E.T., Gangwar, M., Kopalle, P. and Voleti, S., 2017. The role of big data and predictive analytics in retailing. *Journal of retailing*, 93(1), pp.79-95.
- [22] Zhang, Y., Ren, S., Liu, Y. and Si, S., 2017. A big data analytics architecture for cleaner manufacturing and maintenance processes of complex products. *Journal of cleaner production*, 142, pp.626-641.

- [23] Bardsley, M., Steventon, A. and Fothergill, G., 2019. Untapped potential: Investing in health and care data analytics. London: Health Foundation.
- [24] Trkman, P., McCormack, K., De Oliveira, M.P.V. and Ladeira, M.B., 2010. The impact of business analytics on supply chain performance. *Decision Support Systems*, 49(3), pp.318-327.
- [25] Vindigni, G., Janssen, M.A. and Jager, W., 2002. Organic food consumption: A multi-theoretical framework of consumer decision making. *British Food Journal*, 104(8), pp.624-642.
- [26] Williams, S., 2016. Business intelligence strategy and big data analytics: a general management perspective. Morgan Kaufmann.
- [27] Peersman, G., 2014. Overview: Data collection and analysis methods in impact evaluation. UNICEF Office of Research-Innocenti.
- [28] Birks, H.J.B., 2012. Overview of numerical methods in palaeolimnology. *Tracking environmental change using Lake sediments: Data handling and numerical techniques*, pp.19-92.
- [29] Kabacoff, R., 2022. *R in action: data analysis and graphics with R and Tidyverse*. Simon and Schuster.
- [30] Abdelghany, M. and Tahar, S., 2021. Cause-consequence diagram reliability analysis using formal techniques with application to electrical power networks. *IEEE Access*, 9, pp.23929-23943.
- [31] Halkjelsvik, T. and Jørgensen, M., 2018. *Time Predictions: Understanding and Avoiding Unrealism in Project Planning and Everyday Life* (p. 110). Springer Nature.
- [32] Tao, F., Qi, Q., Liu, A. and Kusiak, A., 2018. Data-driven smart manufacturing. *Journal of Manufacturing Systems*, 48, pp.157-169.
- [33] Jadoul, Y., Thompson, B. and De Boer, B., 2018. Introducing parselmouth: A python interface to praat. *Journal of Phonetics*, 71, pp.1-15.
- [34] Kumar, S. and Singh, M., 2018. Big data analytics for healthcare industry: impact, applications, and tools. *Big data mining and analytics*, 2(1), pp.48-57.
- [35] Kim, S.J., Marsch, L.A., Hancock, J.T. and Das, A.K., 2017. Scaling up research on drug abuse and addiction through social media big data. *Journal of medical Internet research*, 19(10), p.e353.
- [36] Vieira, A.A., Dias, L.M., Santos, M.Y., Pereira, G.A. and Oliveira, J.A., 2020. On the use of simulation as a Big Data semantic validator for supply chain management. *Simulation Modelling Practice and Theory*, 98, p.101985.

- [37] Chen, M., Mao, S. and Liu, Y., 2014. Big data: A survey. *Mobile networks and applications*, 19, pp.171-209.
- [38] L'heureux, A., Grolinger, K., Elyamany, H.F. and Capretz, M.A., 2017. Machine learning with big data: Challenges and approaches. *Ieee Access*, 5, pp.7776-7797.
L'heureux, A., Grolinger, K., Elyamany, H.F. and Capretz, M.A., 2017. Machine learning with big data: Challenges and approaches. *Ieee Access*, 5, pp.7776-7797.
- [39] Lee, I., 2017. Big data: Dimensions, evolution, impacts, and challenges. *Business horizons*, 60(3), pp.293-303.
- [40] Mikalef, P., Boura, M., Lekakos, G. and Krogstie, J., 2019. Big data analytics and firm performance: Findings from a mixed-method approach. *Journal of Business Research*, 98, pp.261-276.
- [41] Marr, B., 2016. Big data in practice: how 45 successful companies used big data analytics to deliver extraordinary results. John Wiley & Sons.
- [42] Urbinati, A., Bogers, M., Chiesa, V. and Frattini, F., 2019. Creating and capturing value from Big Data: A multiple-case study analysis of provider companies. *Technovation*, 84, pp.21-36.
- [43] Caputo, F., Keller, B., Möhring, M., Carrubbo, L. and Schmidt, R., 2023. Advancing beyond technicism when managing big data in companies' decision-making. *Journal of Knowledge Management*.
- [44] Schwertner, K., 2017. Digital transformation of business. *Trakia Journal of Sciences*, 15(1), pp.388-393.
- [45] Xu, L.D. and Duan, L., 2019. Big data for cyber physical systems in industry 4.0: a survey. *Enterprise Information Systems*, 13(2), pp.148-169.
- [46] Patel, Mimansha, and Nitin Patel. "Exploring research methodology." *International Journal of Research and Review* 6.3 (2019): 48-55.
- [47] Grotov, Konstantin, et al. "A large-scale comparison of Python code in Jupyter notebooks and scripts." *Proceedings of the 19th International Conference on Mining Software Repositories*. 2022.
- [48] Quaranta, Luigi, Fabio Calefato, and Filippo Lanubile. "Kgtorrent: A dataset of python jupyter notebooks from kaggle." *2021 IEEE/ACM 18th International Conference on Mining Software Repositories (MSR)*. IEEE, 2021.
- [49] Namiot, Dmitry. "Time Series Databases." *DAMDID/RCDL* 1536 (2015): 132-137.

- [50] Nguyen, Cattram D., John B. Carlin, and Katherine J. Lee. "Model checking in multiple imputation: an overview and case study." *Emerging themes in epidemiology* 14.1 (2017): 1-12.
- [51] Taleb, Ikbali, Rachida Dssouli, and Mohamed Adel Serhani. "Big data pre-processing: A quality framework." *2015 IEEE international congress on big data*. IEEE, 2015.
- [52] Maharana, Kiran, Surajit Mondal, and Bhushankumar Nemade. "A review: Data pre-processing and data augmentation techniques." *Global Transitions Proceedings* 3.1 (2022): 91-99.
- [53] Marcheggiani, Diego, and Ivan Titov. "Encoding sentences with graph convolutional networks for semantic role labeling." *arXiv preprint arXiv:1703.04826* (2017).
- [54] Joshi, Prateek. *Artificial intelligence with python*. Packt Publishing Ltd, 2017.
- [55] Petkovic, Dragutin, et al. "Improving the explainability of Random Forest classifier—user centered approach." *Pacific symposium on biocomputing 2018: proceedings of the pacific symposium*. 2018.
- [56] Song, Yan-Yan, and L. U. Ying. "Decision tree methods: applications for classification and prediction." *Shanghai archives of psychiatry* 27.2 (2015): 130.
- [57] Das, Abhik. "Logistic regression." *Encyclopedia of Quality of Life and Well-Being Research*. Cham: Springer International Publishing, 2021. 1-2.
- [58] Luque, Amalia, et al. "The impact of class imbalance in classification performance metrics based on the binary confusion matrix." *Pattern Recognition* 91 (2019): 216-231.
- [59] Ohsaki, Miho, et al. "Confusion-matrix-based kernel logistic regression for imbalanced data classification." *IEEE Transactions on Knowledge and Data Engineering* 29.9 (2017): 1806-1819.
- [60] Bzdok, Danilo, and Andreas Meyer-Lindenberg. "Machine learning for precision psychiatry: opportunities and challenges." *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging* 3.3 (2018): 223-230.
- [61] Kynkäänniemi, Tuomas, et al. "Improved precision and recall metric for assessing generative models." *Advances in Neural Information Processing Systems* 32 (2019).
- [62] Sajjadi, Mehdi SM, et al. "Assessing generative models via precision and recall." *Advances in neural information processing systems* 31 (2018).

- [63] Jia, Robin, and Percy Liang. "Adversarial examples for evaluating reading comprehension systems." *arXiv preprint arXiv:1707.07328* (2017).
- [64] Chicco, Davide, and Giuseppe Jurman. "The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation." *BMC genomics* 21.1 (2020): 1-13.