Project Submission Pro-Forma

Student name: **Javeria Dal**

Student ID: **5508978**

I wish the dissertation to be considered for the course (select one only):

☐ MSc in Cyber Security Engineering
☐ MSc in Cyber Security Management
☐ MSc in e-Business Management
☒ MSc in Engineering Business Management
☐ MSc in Healthcare Operational Management
☐ MSc in Innovation & Entrepreneurship
☐ MSc in Intelligent Manufacturing Systems
☐ MSc in International Trade, Strategy & Operations
☐ MSc in Programme & Project Management
☐ MSc in Supply Chain & Logistics Management
☐ MSc in Sustainable Automotive Engineering
☐ MSc in Smart, Connected and Autonomous Vehicles

I confirm that I have included in my dissertation:

☒ An abstract of the work completed
☒ A declaration of my contribution to the work and its suitability for the degree
☒ A table of contents
☒ A list of figures & tables (if applicable)
☒ A glossary of terms (where appropriate)
☒ A clear statement of my project objectives
☒ A full reference list (the Harvard referencing style is recommended for WMG)
☒ An appendix containing email confirmation of ethical approval or waiver

If receiving ethical approval, the ethical approval number for this research is: **WMG-R_2iydcYUpUFk0cOy**

☒ I consent to ongoing storage of this dissertation and potential access by third parties (e.g. for staff/ student training purposes)

Signed: Javeria Dal Date: 02/08/2024

New Product Forecasting using Machine Learning

by

Javeria Dal, MSc EngBM

Dissertation submitted in partial fulfilment for the Degree of Master of Science in Engineering Business Management

WMG
University of Warwick

Submitted September, 2024

Declaration

I have read and understood the rules on cheating, plagiarism and appropriate referencing as outlined in my handbook and I declare that the work contained in this assignment is my own, unless otherwise acknowledged.

No substantial part of the work submitted here has also been submitted by me in other assessments for this or previous degree courses, and I acknowledge that if this has been done an appropriate reduction in the mark I might otherwise have received will be made.

Project definition for my degree (as copied from http://www2.warwick.ac.uk/fac/sci/wmg/globalcontent/general/project/requirement/)

"The project should normally be related to the management of:

1. companies in the engineering sector,
2. the engineering function within a non-engineering company *or*
3. the supply chain within the engineering sector.

The project could address many different aspects such as operational, financial, human resource, technical or strategic management issues.

Where the project is of a technical nature, there must be clear evidence of business benefit from this technology.

If the focus of the project is outside the above industrial spectrum it MUST contain considerable comparative analysis of practices in the engineering sector."

This project relates to the domain of forecasting within supply chain. It addresses the key problem of forecasting demand for a novel product with scarce historical data. Forecasting accurate demand is critical for effective inventory management and closely relates to all business functions within supply chain. This research study investigates the role of machine learning models in improving the forecast accuracy and proposes a comprehensive framework to combat the challenges faced in new product forecasting.

Abstract

The purpose of this dissertation is to explore the challenges faced in new product forecasting and assess the impact of machine learning algorithms in improving forecast accuracy. Due to limited historical data available, forecasting for new products is usually conducted through qualitative methods. This research, based on the analysis of publicly available datasets, categorises new products into two dimensions, line extensions and diversification. A robust framework is developed, combining sales, customer reviews, social media and market research. For line extensions, M5 Forecast Accuracy time-series data was analysed to measure the impact of culture on demand. Linear Regression, ARIMA and Random Forest were performed to detect trends in events from 2011 to 2015 and predict demand for the next year. Linear Regression proved to be the most accurate model, yielding 96.08% accuracy when there was a logical trend observed, followed closely by Random Forest. To derive demand for diversification products, product differentiation and market analysis were used. Industry growth, target audience and competitor analysis were some of the key indicators of market analysis. Sentiment analysis was performed on Amazon's review dataset, incorporating customer preferences and product differentiation into the forecast. The findings from M5 data underline that machine learning algorithms can effectively improve forecast accuracy even when there is limited historical data available. Sentiment of users although an impersonal data source, can draw many insights from raw textual data. The main findings show that several data sources beyond traditional historical sales can be beneficial in demand sensing. The principal lesson learned from this project is that poor data quality can severely reduce the performance of forecasting models. Low aggregability of secondary data sources was also observed due to the different motivations of the datasets in consideration. Building upon the learnings, primary data sources are recommended over secondary data in new product forecasting. The premise of social media is discussed, suggesting directions for future research.

Acknowledgements

I would like to express my sincere gratitude to my supervisor, Mr. Dave Food, for his unwavering guidance and support. His expertise and insightful feedback greatly enriched the quality of this project and his unwavering encouragement was instrumental in its successful completion. I also appreciate the University of Warwick for providing the resources necessary to undertake this research.

Finally, I wish to convey my deepest appreciation to my parents and my sister for their steadfast support in all my academic endeavours. Their belief in my abilities and their unconditional love has been a source of great strength and motivation for me throughout this journey.

Table of Contents

List of tables

List of figures

List of definitions

Overfitting: When a model shows good performance on training data but fails to generalise when tested on new, unseen data. This happens when a model is fairly complex and over-explains the sample observations (Bilger & Manning, 2015).

Underfitting: When a model fails to detect patterns in the provided data points, leading to biased and inaccurate results (Kolluri et al., 2020).

Knowledge Discovery in Data (KDD): It is known as the practice of extracting important information, relationships and associations from large datasets (Pazzani, 2000).

Root Mean Squared Error (RMSE): The difference between actual and predicted values, used to measure the effectiveness of prediction models (Ćalasan et al., 2020).

Predictive Analytics: Artificial intelligence models that analyse historical data to identify patterns, make predictions and draw conclusions (LU et al., 2017).

Quantile Regression Forest (QRF): An extension of random forest that calculates the conditional distribution of dependent variables in a non-parametric way and provides better estimation than traditional regression models (Dang et al., 2022).

Bayesian Modelling: A probabilistic model that integrates past knowledge with observations to draw associations, model relationships and make predictions. It operates on the general theory of Bayes' theorem of hypothesis, evidence and cause and effect assumptions (Chen & Pollino, 2012).

Principal Component Analysis: A statistical technique widely used in data analysis for dimensionality reduction and simplifying complex, high-dimensional data. It reduces the number of variables while retaining the most important characteristics of the dataset.

Text Mining: An interdisciplinary field that extracts meaningful information and insights from unstructured textual data (Hassani et al., 2020).

Natural Language Processing: NLP is a branch of computer science that combines concepts from linguistics and machine learning, involving algorithms that interpret and process human language to analyse its syntax, semantics and sentiment.

General Adversarial Networks (GANs): Deep learning generative models that learn from high-dimensional data and output realistic time series, image, speech, audio and design patterns (Saxena & Cao, 2022).

Chapter 1: Introduction

1.1 Background

The survival of a new product is dependent on how accurately sales volume matches pre-defined expectations (Gartner & Thomas, 1993; Kahn, 2002). New product forecasting is the realistic estimate of attainable sales of a new product under a set of predetermined conditions. Forecast in its literal meaning is the prediction of an attribute in future. In this research project, the prediction of customer demand for a new product, ready to be launched, will be explored. Since the supply chain begins with customer demand it is critical to forecast demand accurately. Failure to do so would lead to erroneous decisions, shortage or excess in inventory, loss of customers and sometimes even bankruptcy (Ching-Chin et al., 2010; Lee et al., 2014; Sanders, 2017).

Vollmann et al. (2005, p. 2) depict demand planning as a gateway that connects the marketplace with the warehouses. Hence, organisations must determine required quantities in advance to plan for accurate inventory levels. Overstating demand leads to excess inventory, wasted resources and increased holding stock costs. Understating demand prompts decreased service levels, customer dissatisfaction, loss of reputation and distorted decision-making (Roettig, 2016, p. 246) and strains the business resources and service functions of the company (Gartner & Thomas, 1993).

New Product Forecasting refers to forecasting demand for a new or novel product. The biggest limitation of this type of forecasting is the lack of available historical data. Unlike traditional forecasting techniques like time series and regression, new product forecast has scarce sales data. This allows an organisation to leverage several data sources to optimise inventory levels of the new product (Lee et al., 2014).

As products' life cycles are getting shorter and customer preferences are changing rapidly, companies should fine-tune their forecasting models in accordance with shifting customer requirements (Tsafarakis et al., 2011). **Machine Learning**, a subfield of artificial intelligence, analyses large amounts of data in a relatively short time period compared to traditional forecasting methods. These models have the ability to adapt to changing trends and customer requirements quickly (Ching-Chin et al., 2010).

This research aims to explore the potential opportunities of combining machine learning methodologies with traditional forecasting techniques in new product forecasting. By integrating the two approaches, this research study investigates how using advanced computational algorithms alongside traditional forecasting methodologies can improve predictive accuracy, enhance decision-making processes and reveal new insights for optimising forecasting models across different domains and industries.

1.2 Primary Research Question

How can machine learning techniques be utilised to improve the forecast accuracy of new products when there is limited historical data available?

1.3 Research Objectives

1. Identify and validate key factors that influence new product forecasting and analyse their impact on demand.
2. Describe key challenges faced by forecasting models and recommend mitigation strategies.
3. Develop a mixed method approach to demand forecast of a novel product with limited sales data.
4. Discuss the feasibility of various data sources to enhance forecasting algorithms.
5. Evaluate the effectiveness of machine learning algorithms in improving forecast accuracy.
6. Identify limitations and risks associated with using machine learning algorithms for demand forecasting and mitigation plans to enhance the effectiveness of predictive models.

1.4 Significance

While the research in time series forecasting has been extensive, the forecasting of new products has experienced only modest exploration (Skenderi et al., 2024). A common practice is to adopt a judgmental, marketing or subjective approach (Fye et al., 2013; Smirnov & Sudakov, 2021). Feiler & Tong (2022) identified in their study

that managers tend to have inflated expectations of new products before their launch. This causes unrealistically high demand expectations for the product. Forecasting models provide a more realistic measure of future sales considering all internal and external variables. Forecasting is not a standalone decision. The empirical applications of forecasting include operations, supply chain, economics, finance, consumption, demographics, demand and many other fields (Valenzuela et al., 2023). It affects all interconnected operations of the supply chain from sales and marketing to downstream operational decisions. Therefore, a systematic approach should be adopted when forecasting demand for new products.

Classical forecasting methods assume historical sales data for a specific time period to be available which is not the case for new products (Smirnov & Sudakov, 2021). The advent of Industry 4.0 has enabled companies to gather massive amounts of real-time data (Choi et al., 2022). Sanders (2017) indicates that dealing with data intelligently could lower inventory costs, storage costs and wasted production. Data can reduce uncertainty in the supply chain and garner improved business performance (Shen & Chan, 2017).

Van Steenbergen & Mes (2020) argue that uncertainty is often overlooked in new product forecasting. Human beings are statistically naive and use noisy data to predict demand and thereby, undermine the uncertainty behind it (Feiler & Tong, 2022). One solution to this problem is to integrate production decisions with machine learning to reduce forecast uncertainty.

The main objective of this research project is to provide a robust forecasting framework for new products. This framework will enable companies to leverage diverse data sources and incorporate various machine learning techniques to reach accurate forecasts. The research seeks to pinpoint the most crucial factors evident in new product forecasting. Evaluating data-driven methodologies based on the accuracy of results and recommending the most effective approach for strategic decision-making will be the key milestones of this study.

1.5 Scope and Limitations

Big Data consists of a huge volume of data collected by companies and has many dimensions compared to traditional forecasting methods (Wu et al., 2014). There is a massive influx of internet data allowing them to gain deeper insights into customers' preferences (G. Wang et al., 2016a). The integration between traditional and data-driven methodologies can lead to the development of business models that extract valuable information from implicit data and provide key insights to transform business operations.

While reviewing the literature on best practices in new product forecasting, one of the biggest areas of improvement was the accuracy of the forecast. Lee et al. (2014) highlighted that analogical and subjective methods, common in new product forecasting, fail to predict accurate product estimates. Yamamura et al. (2022), argued that analytical methods when used in isolation do not provide accurate results for complex real-world problems, signifying a need to have a good mix between analytical and subjective approaches. Kim & Shin (2016) suggest combining historical data with other factors as forecasting models with a limited number of variables fail to capture fluctuations and market volatility. Building upon this motivation, this research proposes a mixed method framework of new product forecasting combining quantitative data such as historical time-series sales, online traffic and customer engagement and qualitative data such as product attributes, social media, market research, expert opinion and customer reviews to improve the accuracy of demand predictions of new products.

Although the scope of this research project is limited to retail and e-commerce products it can be extended and generalised towards other industries. Since customer preferences, demographics and culture are common factors among all industries, this framework can be fine-tuned for all product types depending on the forecast level, objectives and time horizons.

1.6 Structure

The study begins with a literature review that dives into theoretical concepts of new product demand forecasting and machine learning, factors that impact new product forecasting, a review of recent developments in forecasting techniques, challenges and limitations of current frameworks, the role of emerging technologies in demand forecasting, mitigation strategies and commercial proposition of the project.

In methodology, research design and research methodology will be formulated. Participants, variables, measures and selection criteria for data sources will be established. Hypotheses are developed to guide the research design and achieve the research objectives. Data collection methods, data sources, data types and sampling methods will be defined. A research framework will be developed to validate the research objectives. Challenges in the data collection process and mitigation strategies will be highlighted in detail. There will be a discussion on ethical considerations of data collection processes and a refinement methodology plan will be outlined to improve the outcomes.

Results & Analysis presents the findings of the study and its significance with the research objectives. It follows a data analysis pipeline with data preprocessing, data modelling and results. The features will be analysed and prioritised based on their impact on the forecast model and evaluation metrics will be proposed. Exploratory Data Analysis (EDA) will be conducted to identify patterns and summarise the main characteristics of the data. Machine learning frameworks will be applied to the dataset to forecast demand for the selected product. Their effectiveness will be assessed based on the accuracy and forecast error.

Discussion is based on the interpretation of the findings and their significance in answering the research question. This section addresses the limitations, implications and applications of the study. The scalability, generalisability and commercial propositions of the research are discussed. The project concludes with a discussion on the future direction of the research in new product forecasting.

Chapter 2: Literature Review

2.1 Theoretical Concepts and Frameworks in Forecasting

2.1.1 Time Series Analysis

Time series analysis is a statistical technique used to analyse sequential data collected over time (Sanders, 2017). It includes measurements observed at specified time periods (hourly, daily, monthly, quarterly) and uses time as a variable to analyse trends based on the movement of data points (Kim & Shin, 2016). Time series analysis highlights repeated patterns and fluctuations in demand that occur at regular intervals. Moreover, it outlines noise or variation and reflects the uncertainty and randomness in data (Mills, 2019, p. 4).

Time series analysis is a benchmark for companies to predict future metrics based on sales data (Kahn, 2002; Sanders, 2017; Wang et al., 2016). The forecasting theory of time series encompasses the identification of past patterns to predict future values (Valenzuela et al., 2023, p. 157). It is an amalgamation of statistical methods that are distinct in their functionality and metrics. The selection of a statistical method is based on forecasting objectives, time horizon and product characteristics.

2.1.2 Autoregressive Integrated Moving Average (ARIMA)

ARIMA is among the most popular statistical methods used to measure the temporal behaviour of natural and social events throughout sequential time intervals. The ARIMA model forecasts future values based on past values. It does so by averaging all the observations from the previous periods of a time series and using a weighted sum to predict future demand (G. Wang et al., 2016a).

2.1.3 Random Forest

Random Forest is an algorithm that is widely used in classification and regression tasks (Y. Wang et al., 2018a). It constructs multiple decision trees to reach a singular, accurate prediction. The random forest quantifies the importance of variables while predicting the outcome, increasing the model's accuracy and interpretability (Cheng et al., 2020). The benefit of random forest lies in its ability to handle both categorical and continuous data (van Steenbergen & Mes, 2020).

2.1.4 Qualitative Methods

Expert judgement, executive opinion and market research are some of the most common qualitative forecasting methods in new product forecasting (Feiler & Tong, 2022; Fye et al., 2013; Gartner & Thomas, 1993). According to Gartner & Thomas (1993) founder's expertise is the most common decision-making strategy before product launch and directly relates to the accuracy of the forecast. Companies tend to trust the leader's experience and expertise more than complex forecasting methodologies. Although quantitative forecast methods have higher success rates and optimised performance than qualitative methods, in practice expert sourcing has the highest realisation rate (Fye et al., 2013).

Some of the frequently used methods for qualitative analysis include the Delphi method, expert panel and market surveys. This includes respondents such as marketing managers, practitioners, product owners and executives at the cross-section of industries (Kahn, 2002). Macroeconomic projections are important for companies to operate in non-stable and challenging environments such as COVID-19 and the Russia-Ukraine war. Usually, advanced data analytics when combined with expert knowledge is sufficient to derive market demand and consumer preferences (Afrin et al., 2018).

Analysis of customer behaviour is another popular forecasting technique that is mostly qualitative (Kahn, 2002). Yamamura et al. (2022) represent domain knowledge as the sum of a manager's experience, customer perceptions and market and competitor analysis. Customer Behaviour includes customer orders, buying patterns, popular trends, preferences and purchasing history (Gartner & Thomas, 1993; Hu et al., 2019; Meeran et al., 2017; Shah & Theodosoulaki, 2018). A significant disadvantage of customer research is the high cost that makes it unsuitable to be conducted under budget constraints (Gartner & Thomas, 1993; Kahn, 2006). However, in recent times market research is much easier and flexible to follow due to the availability of secondary data from government sources and industry journals.

2.2 Artificial Intelligence

Artificial Intelligence is the ability of a machine to do the right thing at the right time in the right manner (Russell & Norvig, 2021, p. 19). This is called the rationality of a computer program that considers intelligence as doing what is expected from it (Jung, 2022, p. 10). The right thing or right behaviour is defined by the objective. However, it should be noted that a perfect rational model that always takes the right action and generates the right output is unlikely to be achieved as the computational demand and complexity required tend to be very high (Russell & Norvig, 2021, p. 22).

2.2.1 Machine Learning

Machine Learning (ML) is a sub-field of Artificial Intelligence where computers learn from data and make predictions without being explicitly programmed (Jung, 2022). The human brain is incapable of foreseeing all potential outcomes at a given moment; therefore, an agent that learns from observations and enhances its performance is beneficial (Russell & Norvig, 2021, p. 69). ML algorithms enable computers to identify patterns, extract information and make data-driven decisions based on past experiences and observations. These algorithms are trained on a dataset and extract insights by identifying hidden associations and relationships between attributes. A machine learning algorithm learns in three ways:

- Supervised learning: In supervised learning, a model learns from labelled data with pre-defined output. These data points are labelled by human experts (Jung, 2022, p. 12). It creates a function that maps input features to output and makes predictions on unseen data called test data (Russell & Norvig, 2021, p. 671). Instance-based learning, probabilistic learning, regression, decision trees and support vector machines are a few examples of supervised learning (Jo, 2021).
- Unsupervised learning: In unsupervised learning, the model learns unlabeled data without having explicit feedback (Russell & Norvig, 2021, p. 671). The model's task is to discover hidden patterns within the data. **Clustering** is an example of unsupervised learning where the model makes clusters of data points, known as peer groups, based on discovered associations (Aggarwal, 2016, p. 45; Jo, 2021; Jung, 2022, p. 13).
- Reinforcement Learning: This is a reward-based system where the algorithm is given feedback (score) after each prediction (Russell & Norvig, 2021, p. 671). The score increases with each correct prediction and decreases with an incorrect one. The main goal of reinforcement models is to maximise cumulative scores over time.

2.2.2 Accuracy

Accuracy is a measure to assess the effectiveness of machine learning algorithms. It refers to the correctness or "rightness" of a model in making predictions on test data (Jung, 2022). Forecasting models based on only historical data or past observations often provide limited accuracy (Kim & Shin, 2016). There are many factors including sales that affect product demand and must be taken into consideration to get accurate insights. When comparing time series models, it is expected that their predictive ability is evaluated based on the forecast accuracy (Valenzuela et al., 2023). In applied modelling, error is another measure that is used in parallel with accuracy to reflect a model's performance (Y. Wang et al., 2018a). These include forecasting measures like errors like Mean Absolute Percentage Error (MAPE), Median Absolute Percentage Error (MdAPE) and Root Mean Square Percent Error (RMPSE) called relative errors which are calculated based on the deviation of predicted value from the actual value.

2.2.3 Bias

In machine learning, bias refers to the difference between the predicted values of the model and the actual observation of the target variable (Russell & Norvig, 2021, p. 672). It signifies a situation where a predictive value deviates from its true value. Bias is closely related to the **underfitting** of data which oversimplifies the dataset and makes the model unable to detect patterns and generate accurate predictions (Jo, 2021). Noise is commonly related to information, introducing bias in the data and leading to false conclusions and inaccurate forecasts (Feiler & Tong, 2022).

2.2.4 Variance

Variance refers to the change in the performance of a predictive model when there are random fluctuations in the results of the model that has been trained (Jung, 2022, p. 128). It is often a consequence of **overfitting**, when the algorithm is trained with insufficient training examples or when too much attention is given to particular data points causing it to perform badly on unseen data (Aggarwal, 2016; Jo, 2021R; Russell & Norvig, 2021; p. 673; Yamamura et al., 2022).

2.2.5 Predictive Analytics

Predictive analytics is the practice of extracting trends and meaningful insights from the dataset using statistical techniques, machine learning, and data mining. It typically includes the integration of artificial intelligence and traditional forecasting methods. Predictive analytics have been gaining a notable interest in the industry due to its computational capabilities and the increasing need for improved decision-making (McCarthy et al., 2019; G. Wang et al., 2016a). It enables organisations to predict consumer behaviour, identify new customers, improve operations, and mitigate supply chain risks (McCarthy et al., 2019). Furthermore, predictive analytics gives an organisation a competitive advantage by making data-driven decisions. Recommendation system is an example of predictive analytics, used by e-commerce giant Amazon and streaming sites like Netflix to elevate user experience and target customers accurately (Aggarwal, 2016; Amazon, 2019; Choy, 2021). Predictive analytics is the convention of making robust predictions about customer purchases and tracking consumer behaviour using explicit (i.e. sales) and implicit (i.e. reviews) forms of feedback (Aggarwal, 2016). This study will incorporate customer reviews into the solution to derive forecasts for new products.

2.2.6 Big Data Analytics

Big Data Analytics (BDA) involves dealing with large and complex datasets to uncover hidden patterns, trends, correlations and insights. It assists organisations in making informed business decisions (Ge et al., 2018; Kaleem et al., 2023; Ma et al., n.d.). Information sharing in real-time through big data increases transparency and accountability and leads to improved decision-making (G. Wang et al., 2016a). Since big data comes with significant volume and speed, traditional data analysis frameworks need to be revisited so they can handle the intrinsic complexities present in large datasets (Kaleem et al., 2023). The greatest advantage of big data is its ability to process data in real time to respond quickly to changing market conditions. Big data analytics can be studied in three dimensions:

- Descriptive
- Predictive
- Prescriptive

In the context of New Product Forecasting, the predictive approach will be considered.

2.2.7 Data Mining

Data mining involves discovering trends, patterns, and correlations in large datasets. This is achieved by applying machine learning algorithms to data acquired from various sources such as databases, social media, sensors, emails, and website traffic (Oracle, 2022). The main objective of data mining is to interpret complex data sets and predict future outcomes based on past observations. This process, also known as Knowledge Discovery in Data (KDD), involves converting raw (structured and unstructured) data into meaningful information (IBM, 2023). Data Mining is a technique used across various industries to generate patterns and provide recommendations.

2.3 New Product Introduction

Kahn (2006) classifies products into four categories highlighted in Figure 1.1 based on the novelty of the market and product technology. These categories are market penetration, market development, product development and diversification. New product development lies between the second and fourth quadrants referred to as line extensions and diversification products (See Figure 1.1).

Market Technology Matrix

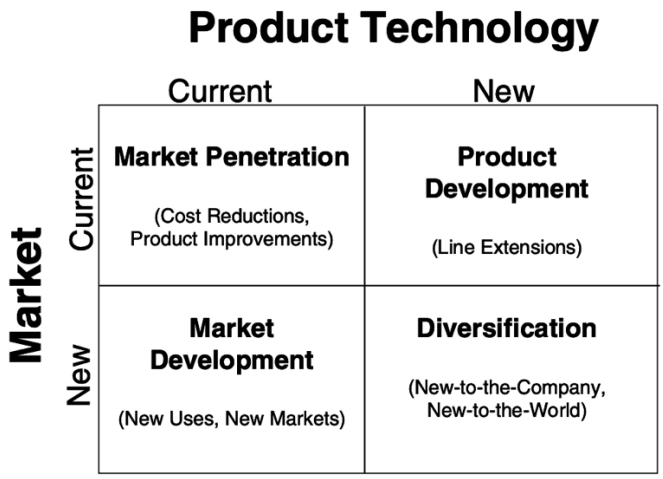


Figure 2.1: Market Technology Matrix, (Kahn, 2006)

The market-technology matrix assists organisations in developing growth strategies by analysing their existing products and target markets (Kahn, 2006). For this research project, the focus will be on product development (line extensions) and diversification. **Product Development** means adding new features to existing products or creating new products that address existing market needs. **Diversification** means introducing a product that is new to the company or new to the world. A company either creates a novel product in an existing market or creates a new market for a product never seen before.

According to (Kahn, 2006), new product planning becomes riskier as companies move from the upper left quadrant to the bottom right quadrant in Figure 2.1. Diversification is risky because the product and market are untested and unfamiliar to the organisation. Therefore, special care should be taken while driving forecasts for such products. This research study explores various datasets and develops a machine learning-based forecasting framework that aggregates customer data with product data to derive a demand forecast for a novel product.

Based on discussions with experts, it's suggested that using a machine learning-based solution for product diversification is costly and not viable for most organisations. On the other hand, line extensions are a more financially feasible option as there is already historical data available for forecasting. This research aims to investigate both categories of new products, product development and diversification, by projecting the forecast for two types of products. One is an improved version of an existing company product (line extension) and another is a new-to-the-company product which exists in the external market (diversification). This will allow us to explore the potential of the forecasting framework in both areas.

2.4 Factors in New Product Forecasting

Research Objective 1: Identify and validate key factors that influence new product forecasting and analyse their impact on demand.

Several factors impact forecast accuracy, for example, demographics, shelf life, geography, customer preferences, product type, technology, and time horizon (Goetze, 2011; Kahn, 2006; Meeran et al., 2017; Rawlinson, 2017; Sanders, 2017). As there is no universally acceptable measure of forecast accuracy (Fye et al., 2013), evaluating their impact is a complex task. Understanding such factors and their influence on forecasts can improve the accuracy of predictions. According to Son & Han (2011), the long-term success of novel and innovative products depends on the technology readiness of customers, that is, how prepared the consumers are to use that technology.

Kahn (2006, p. 5) describes **time horizon** as one of the initial steps in new product forecasting which dictates whether the forecast is for the short term or long term and how often the forecast needs to be updated (weekly, monthly, quarterly). According to Sanders (2017, p. 5), the long-term forecast is used for strategic planning such as expansions and the short-term forecast is used to make tactical decisions. Fye et al. (2013) investigated the factors

that affect forecast accuracy and deduced that short and medium-term forecast projections were two times more successful than long-term forecasts. In this research focus will be on short-term, seasonal forecasts for a novel product.

Another important factor in forecast prediction is the organisational **leader's expertise**. Gartner & Thomas (1993a) investigated the uncertainty between forecast and actual demand and highlighted that a founder's experience directly relates to forecast accuracy. Another factor critical to forecast accuracy that was identified was the number of resources employed in the decision-making.

Understanding **customer behaviour** and underlying buying patterns provides the best success opportunities that arise from customers (Rawlinson, 2017, p. 28). Gartner & Thomas (1993) elaborate on the relationship between proximity of the customer and demand forecast, that is, the nearer one is to the customer the more accurate the forecast. Hence, customers' buying patterns and identification of **seasonal trends** are found to be directly linked to new product forecasting.

Goetze (2011) analysed group conformity and face consumption in the adoption of new products among Chinese customers. The author outlined that **culture** plays an important role in customers' buying behaviour. People belonging to different cultures have different attitudes towards new products. Thus, it can be inferred that culture and **demographics** are critical factors in product sales and they should be treated strategically due to the increased globalisation of the consumer market.

Traditional values have a significant impact on the receptiveness and willingness of customers to purchase new products. When forecasting demand for different demographics, companies should consider local culture and the purchasing willingness of different consumer groups. This complicates forecasting as it is ineffective to generalise a forecast for all consumer groups across a geographical area.

Customer preferences are labile and change over time. According to Meeran et al. (2017), there are certain psychological factors involved in customer's subjective perception of the product which adds uncertainty and randomness to the forecast. Therefore, recurrent modifications should be made to forecasts with respect to changing customer preferences to get improved performance and accuracy.

2.5 New Product Forecasting Models & Techniques

According to Lee et al. (2014), three types of models have been developed for pre-launch forecasting:

- Bayesian approach incorporates additional data to the pre-launch forecast limit.
- Subjective approach derives forecast parameters based on managerial opinions and external factors.
- Analogical approach assumes the products are analogous and new product will have a diffusion pattern similar to earlier products. It derives similarities between them and calculates a weighted sum of parameters. The analogical approach incorporates machine learning and statistical algorithms to remove subjectivity from the forecast.

Another approach that could be implemented to forecast demand is the **New Product Life-Cycle analysis** (see Figure 2.2). Hu et al. (2019) utilised Product Lifecycle (PLC) and order data of similar products to fit a PLC curve to forecast demand for new products. The authors derived clusters to divide the dataset into categories and came up with a representative cluster to generate a PLC curve. Afrin et al. (2018) developed a life-cycle demand prediction model for automobiles based on design specifications and historical data.

Expert opinion and **executive judgment** are the most common forecasting methods in organisations (Feiler & Tong, 2022; Fye et al., 2013; Gartner & Thomas, 1993; Hu et al., 2019; Kahn, 2002; Yamamura et al., 2022). Experts have knowledge and expertise in a specific area or domain. They assess information and guide decisions based on their experiences.

Kahn (2002) conducted exploratory research to analyse cross-industry perspectives on preferred pre-launch forecasting practices and identified customer market research, executive opinion and sales force composite as the most popular forecasting techniques. The research suggested that managers prefer less sophisticated techniques such as expert opinion over statistical and quantitative forecasting methods.

Diffusion theory helps forecasters predict the adoption and penetration rate of a novel product in a new market by making educated guesses based on similar products (Kahn, 2002; Lee et al., 2014; van Steenbergen & Mes, 2020). Albeit excellent in forecasting demand for new-to-the-world/new-to-the-company products, this theory is not suitable for incremental, SKU-level products (van Steenbergen & Mes, 2020; Yamamura et al., 2022). Incremental products are often referred to as line extensions (Figure 2.1) when an organisation introduces a product similar to an existing product to improve business.

Van Steenbergen & Mes (2020) introduced a novel model called 'demand forecast'. This model integrates random forest, K-means, and Quantile Regression Forest (QRF) that create prediction intervals to aid in making inventory management decisions. Overfitting is reduced by leveraging the randomness of random forests during feature selection. **Random forests** are particularly valuable for analysing both categorical and continuous data. **Linear regression** is another modelling technique widely used in demand prediction. Smirnov & Sudakov (2021) developed a linear regression model with gradient boosting for predicting the demand for new products in an online store.

Yamamura et al. (2022) assert that analytical methods do not provide accurate results for complex real-world problems. According to Yamamura et al. (2022), linear statistical tools are not enough to capture non-linear relations in data. For this purpose, the authors developed a **neural network** to predict demand for company-specific products by using expert domain knowledge. A neural network replicates the working of a human brain to make predictions (Aggarwal, 2016, p. 87). The domain knowledge was the sum of the manager's experience, customers' perceptions, technologies and market competitors. The network provided 82% accurate demand forecasts. Skenderi et al. (2024) extracted exogenous knowledge from Google Trends as a time series and combined it with multimodel information such as colour, fabric, description, release date and more. The solution was an Artificial Neural Network (ANN) model with the encoder-decoder sequence where the encoder learns representation from Google trends and the decoder analyses metadata and produces a forecast of the sales based on encoder output.

Ching-Chin et al. (2010) developed a standard procedure for new product sales forecasting and determined that New Product Forecasting is related to the company's competitive strategy as the sales mark the financial returns on an organisation's investment. Therefore, managers must derive precise forecasts for new products.

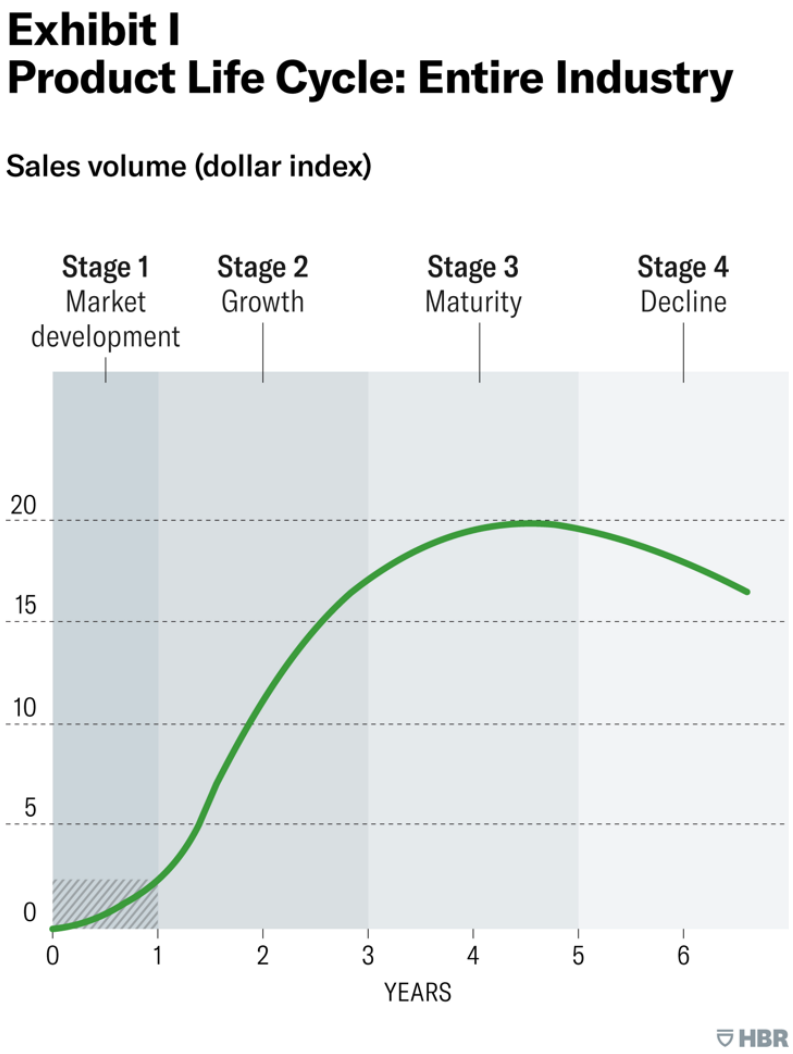


Figure 2.2: Product Lifecycle, (Levitt, 1965)

2.6 Challenges in New Product Forecasting

Research Objective 2: Describe key challenges faced by forecasting models and recommend mitigation strategies.

1. **Lack of historical data**
   Due to limited sales data, new product forecasting models lack a baseline for predicting demand. According to Kahn (2006), the least risky challenge is cost reduction and product improvement followed by market penetration due to the availability of existing internal and external data. Diversification through new products on the other hand is the riskiest gamble a company can make amid a lack of data and highly volatile external factors.

   With scarce sales data, extracting information from impersonal data sources such as managerial judgement and secondary research is typically adopted (Gartner & Thomas, 1993). These methods do not require direct contact with the customer. Usually, the information collected from indirect sources has noise interpolated with it. Making predictions from information diluted with such noise leads to incorrect predictions (Feiler & Tong, 2022).

2. **Marketing Uncertainty**
   Forecasting in the supply chain is a challenge for companies as it deals with uncertainty and increasing variability in demand (Packowski, 2014). This challenge accumulates when a product is novel for the

company and peaks when a company is targeting a brand-new market with a new-to-world product (See Figure 2.1).

This uncertainty results from increasing competition, changing market landscape and technological advancements. Market research although an appropriate response to mitigate this challenge, is both time-consuming and expensive in practice (Kahn, 2006). Therefore, for organisations that are dealing with budget constraints, market uncertainty is a threat constantly looming when challenging new customer spaces.

3. **Data Quality and Relevance**
   Using data from past products poses a challenge of relevance as the past products may not lie in the same category as the new product in consideration. Data quality refers to the variation in the forecasting indicators present in the data (Ching-Chin et al., 2010). A high number of variations leads to outliers, introduces subjectivity and bias, obscures results and presents inaccurate conclusions.

   Moreover, complicated information may result in noise and cause overconfidence in the forecast of selected products (Feiler & Tong, 2022). The mitigation strategy of using human judgment and interpretations of multiple experts can only be helpful if the data is convoluted but primarily correct. However, if the positive forecast error exists in the information itself then this solution would be less effective (ibid). In essence, data quality is one of the major challenges faced when forecasting demand for new products in diversification.

4. **Changing customer requirements**
   Understanding customer behaviour is necessary to forecast demand for a product. Due to the lack of information present in new forecasting models and shifting customer preferences, businesses find it challenging to derive accurate forecasts during the early phases of a product's life cycle (Packowski, 2014).

   The product lifecycle in Figure 2.2 represents the life of a product across four stages. The introduction (market development) stage is critical for companies as customers provide explicit feedback about the product. For products with shorter lifecycles for example tech gadgets, the variation in customer feedback has the biggest impact on the purchase probability (Meeran et al., 2017; Packowski, 2014). Customer requirements change over time and this change must be reflected in demand management to accelerate profit (Matthias et al., 2017; Vollmann et al., 2005, p. 50).

5. **Product Differentiation**
   For new products in the diversification quadrant (Figure 2.1) that will enter an existing market, a firm needs to identify the gap between competitors' products and customer requirements to develop an appropriate brand identity (Kahn, 2002, 2006). This is often done through product differentiation. This identification of competitive advantage is necessary for a company to penetrate a new market and failure to do so would result in suboptimal results and inflated demand forecasts.

   Kato (2012) investigated the effect of differentiating products in the Japanese retail industry and indicated that product differentiation has a positive impact on the productivity and market revenue of a business. Internally a product faces many challenges from the initiation to the execution phase. According to Afrin et al. (2018), design and market aspects of new product development are often treated separately. This approach has been criticised as differentiation found in the design of products is proven to be an important source of data exploitation for accurate forecasts.

6. **Forecasting methods**
   In addition to the lack of data and incorrect information, another challenge is to assess and select a forecasting method. As mentioned by Wu et al. (2014), data is only useful when correct methods are

selected to extract insights from it. The accuracy of forecasting results depends on the right methodology used to derive those numbers.

Kahn (2002) found one manager using diffusion models for product improvement when it was not a suitable statistical model in that scenario and produced inaccurate results. The use of incorrect forecasting methods would only lead to overstocks, stockout, increased costs, supply chain disruptions and lost revenue (Sanders, 2017). Thus, the choice of forecasting model is a critical success factor in new product forecasting.

7. **Data Integration & Complexity**

Variety, one of the attributes of big data, refers to distinct formats in which data is stored (Ge et al., 2018; G. Wang et al., 2016a; P. Zhao & Cao, 2020). When data is consolidated from multiple sources such as market research, customer reviews, sales, economic indicators and more, converging this information into one cohesive dataset can be time-consuming and challenging.

Data can be structured, semi-structured and unstructured. Different sources incorporate different formats of data. A considerable amount of big data is either semi-structured or unstructured (Sakr, 2016, p. 2). Discrepancies in attributes, data collection methodologies and formats lead to misalignment of data. Data Transformation, one of the most critical stages in the data analysis timeline, refers to the transformation of multiple formats into one consistent structure (Ge et al., 2018). This becomes challenging when dealing with diverse data sources present in new product forecasting. Integration of large datasets, used in new product forecasting, demands extensive storage and significant processing power. The discussion on big data, its dimensions and challenges will be explored in detail in subsequent sections.

2.7 Role of Emerging Technology in Forecasting

2.7.1 Artificial Intelligence

As mentioned previously, one of the biggest challenges faced by businesses today is forecasting demand amidst constantly changing customer requirements and market trends (Matthias et al., 2017; Packowski, 2014; Vollmann et al., 2005). A major advantage of AI models is their capability to rapidly adapt to these changes through continuous learning.

In recent years, there has been a shift from traditional forecast models to data-driven algorithm-based methods (Kuo & Kusiak, 2019). Time series, Regression, Ensemble, Reinforcement Learning, Deep Learning and Bayesian models are some examples of AI models used in forecasting. These methods use machine learning to provide insights to develop a decision-making engine. Time series and regression models are well-established statistical methods extensively used to predict demand (Kahn, 2006; Sanders, 2017).

There has been a growing use of Neural Networks that has been documented in forecasting literature (Scher & Messori, 2019; Vrbka, 2021; Yamamura et al., 2022; Zhang et al., 2020). Artificial Neural Networks (ANNs) replicate the processing of a human brain by taking a predefined set of inputs and generating output (Jung, 2022). Neurons, the most basic component of ANNs, are layered interconnected nodes with weights attached to them (Aggarwal, 2016). Unlike other ML models, ANNs do not require a significant amount of data to make predictions (Yamamura et al., 2022). These programs can find complex patterns and relationships in data, making them suitable for forecasting products with limited historical data.

2.7.2 Internet of Things

Enabled by Industry 4.0 significant amount of real-time data is being collected by companies today that can lead to improved strategic decisions (Choi et al., 2022). IoT provides benefits of cost reduction and improved efficiency (Hodge et al., 2015; Zheng et al., 2016). These devices in a retail environment collect customer insights and transfer them to AI models. The models analyse this information to predict demand patterns and help businesses manage inventory efficiently.

Sensors can observe and share real-time data through a channel of interconnected devices (Gao et al., 2019; Hodge et al., 2015; Laranjeiro et al., 2019). They share information through a communication mechanism that provides reliable data coverage for transmission (Hodge et al., 2015). This mechanism not only allows for seamless integration of data points to feed into the model but also boosts improved accuracy and supply chain optimisation.

However, IoT devices in highly complex environments have various constraints. The sensors must be placed carefully and there needs to be a communication channel for the device to transmit data (Hodge et al., 2015). Wireless communication poses security concerns when it shares sensitive information over a network (Srinivas et al., 2021). Hence, privacy measures must be in place to protect the transmission of sensitive data.

2.7.3 Big Data

The commencement of the World Wide Web (WWW) and recent advancements in computing technologies have facilitated the existence of large data sets, a phenomenon known as big data (Russell & Norvig, 2021, p. 44). Big data encompasses a large volume of information and is often characterised by multiple dimensions (Wu et al., 2014).

Big data is the ability of a machine to process large data and is represented by five major dimensions including volume, velocity and variety (G. Wang et al., 2016a; P. Zhao & Cao, 2020), extended by Ge et al. (2018), adding two new dimensions; value and veracity. The dimensions of big data can be assigned to the following data pipeline stages:

Data Processing Stages	**Big Data Dimensions**
Storage	Volume, Velocity, Variety
Processing	Velocity, Veracity
Analysis	Variety, Veracity
Visualisation	Value

Table 2.1: Data Pipeline & Relevant Big Data Dimensions (G. Wang et al.).

The **volume** characteristic of big data represents large streams of data coming from emails, webpages, sensors, satellite images, clickstreams, blogs, search indices, documents, demographic information and more (Gandomi & Haider, 2015; P. Zhao & Cao, 2020; Zheng et al., 2016). This attribute is concerned with the storage, capacity and management of large datasets. The **velocity** of big data refers to the tremendous speed of data collection (Kaleem et al., 2023). It includes real-time data from Global Positioning System (GPS) receivers, streaming platforms, social media, satellite images, smart cards and customer interactions (Gao et al., 2019; Laranjeiro et al., 2019; Ma et al., n.d.).

The **variety** attribute of big data refers to heterogeneous data formats and structures. The majority of big data is semi-structured and unstructured (Sakr, 2016). Structured data include relational and non-relational databases that are highly organised and easy to process. Integration, standardisation, transformation and normalisation are related to the variety of big data. Unstructured data refers to raw qualitative data that is unorganised and must be pre-processed before it can be used for analysis. **Veracity** means quality of data to derive accurate results and **value** refers to the ability of data to contribute to decision making. Together these dimensions influence the scalability, processing capabilities, computing speed, and quality of a solution and shape the design and implementation of the data pipeline.

Big Data Analytics (BDA) provides insights into buying patterns, customer preferences and business trends and offers solutions for cost-cutting and making informed business decisions. Kuo & Kusiak (2019) indicate the moving shift in recent years from analytics-based forecasting to data-driven forecasting. Big data processing contains data from multiple, heterogeneous and autonomous sources having complex relationships and associations between their attributes (Kaleem et al., 2023). As companies are collecting large amounts of data from customers (Wu et al., 2014), this shift could lead to deeper insights into **customers' purchasing** behaviour and preferences and improved forecast **accuracy** (Shen & Chan, 2017). BDA is particularly beneficial when it comes to handling uncertainty and variability in demand and gives an organisation its competitive advantage.

In the information age, companies are analysing the text reviews left by customers on online portals and websites to improve predictive engines. This process is called customer profiling, a sub-category of data mining, where

customers are grouped based on their history (clustering) and this information is fed into the forecast engine to make user-specific predictions (Aggarwal, 2016).

2.8 Mitigation Strategies

Model Overfitting & Improvements

Analytical models fail to provide accurate results for complex real-world problems (Yamamura et al., 2022). New product development is often based on subjective methods such as expert judgement, surveys and extrapolations with competitor products (discussed on page 8). The results are often inaccurate and highly biased. Although machine learning algorithms help reduce such bias, overfitting is a common problem found in machine learning models (Feiler & Tong, 2022). Increasing sample size, decreasing the total number of variables, feature engineering, iterative refinement, regularisation of input and cross-validation are a few techniques used to combat overfitting and inaccuracy in forecasts.

Using Personal Data sources

In recent times, conducting interviews with individual customers and focus groups, online surveys and product demonstrations have become cost-effective (Gartner & Thomas, 1993). This leads to a customer-centric forecast that is in accordance with the changing customer behaviour leading to improved accuracy (Feiler & Tong, 2022; Goetze, 2011; Tsafarakis et al., 2011).

Conjoint Analysis

Simulating market behaviour using conjoint data is an approach frequently adopted in new product design and sales optimisation (Pullman et al., 2002; Sawhney et al., 2017). Conjoint analysis is a statistical method used to analyse customer preferences in market research (Tsafarakis et al., 2011). Typically used in feature combination and attribute trade-off analysis, the simulation helps managers reduce uncertainties in their new product design when targeting a specific market. Its effectiveness depends on the accuracy of market share prediction.

Using a multitude of data sources

Uncertainty is often overlooked in the demand forecasting of new products (van Steenbergen & Mes, 2020). One way to overcome the weakness of low accuracy is to design a decision support system that combines expert knowledge with sales information. Afrin et al. (2018) proposed a data-driven framework which combined design specifications, expert knowledge and product differentiation to produce accurate demand for new products. Similarly, Yamamura et al. (2022), introduced expert domain knowledge and leveraging customers' perceptions to remove subjectiveness and inaccuracy from prediction results. Lee et al. (2014) combined the Bayesian approach, subjective approach and analytical approach to reduce subjectivity from the results.

Update Forecast Data

For time-series datasets to provide accurate results the data must be updated regularly (Valenzuela et al., 2023). Meeran et al. (2017) emphasise the need for firms to update the data with constantly changing customer preferences to improve forecast accuracy. The methods employed to update forecast data are discussed in detail in the discussion chapter.

Second Forecasting

This strategy is usually implemented when there is bias involved in the inferences derived by human experts. Inflated expectations quickly lead to overconfidence in the forecast which can be mitigated by using a second forecasting entity (Feiler & Tong, 2022). For example, if first person's forecast is too extreme, it will be normalised by averaging it with the second person's forecast. This second entity's interpretation moderates the overconfidence in initial forecasts.

Risk Assessment

New products either lie on the diversification quadrant or line extension quadrant depending on the novelty of the market and technology (Figure 2.1). These products often pose a high risk of failure (Ching-Chin et al., 2010;

Kahn, 2006). According to Meeran et al. (2017), risk assessment should be prioritised when establishing future demand as it is unwise to base inventory decisions solely on point forecasts. There are several risks involved in product forecasting. Some of them are market uncertainty, changing customer preferences, demand fluctuations, regulatory changes, technological advancements, supply chain disruptions and inaccurate demand forecasts (Kahn, 2006; Matthias et al., 2017; Meeran et al., 2017; Packowski, 2014; Tsafarakis et al., 2011; Valenzuela et al., 2023). Thus, understanding the risks and having contingency plans is critical for inventory planning and effective decision-making.

2.9 Commercial Proposition of Project

The availability of big data combined with improved computational power of machine learning algorithms, has given Artificial Intelligence great commercial attractiveness (Russell & Norvig, 2021, p. 44). The ability to process information quickly, improved scalability, automation, personalisation, optimisation, concrete insights and improved accuracy has made AI a cornerstone to decision-making across various industries.

Due to increasing technological advancements, some of the emerging trends in new product forecasting include the integration of artificial intelligence into traditional forecasting methods. Predictive analytics have been gaining a fair amount of interest due to their computational capabilities and the increasing need for improved decision-making (McCarthy et al., 2019). Industry 5.0 will further validate the use of advanced simulations and "what-if" scenarios set forth by organisations to predict outcomes accurately. Moreover, e-commerce and retail companies are now leveraging data mining techniques to personalise recommendations for users according to their preferences (Aggarwal, 2016). These recommendation systems are powered by big data and ML algorithms to anticipate demand and manage supply chain operations effectively.

Forecasting models must anticipate the changes in dynamic environments due to the shifting business landscape and the sensitivity of customer preferences (Matthias et al., 2017). Factors like consumer behaviour, changing market conditions, geopolitical events and economic fluctuations all need to be taken into account while deriving demand for a product. Since demand management and forecasting are the cornerstone of all operational activities and a precursor to business growth (Robert Jacobs & ‘Ted’ Weston, 2007; Roettig, 2016; Vollmann et al., 2005), future research is expected to be dedicated towards AI-driven supply chain optimisation and inventory planning.

This research project aims to forecast demand for new products with limited historical data by using machine learning and traditional statistical forecasting models. The New Product Forecasting framework will allow companies to use a multitude of data sources, qualitative and quantitative, to derive accurate forecasts. Another purpose of this research is to identify factors that are most critical in new product forecasting. This study investigates and evaluates machine learning models based on accuracy and suggests the most effective methodology for new product forecasting.

The proposed forecasting framework delves deep into the practicability and scalability of machine learning in the supply chain realm and aims to solve the ever-increasing problem of inaccurate forecasting of new products. As much as subjective qualitative methods like expert judgment and manager’s foresight are important, these are highly influenced by subjectivity, noise and bias (Feiler & Tong, 2022). This study will expand upon the advantages of integrating traditional forecasting methods with machine learning and therefore, alleviating the weaknesses of the former. Different impersonal data sources, usually overlooked in traditional forecasting, will be incorporated into the solution. The capabilities and limitations of big data sources are explored, and suggestions are made on their applicability in new product forecasting. This research study investigates the potential impact and use of social media reviews and textual data in demand forecasting.

While the research objectives are targeted at a specific category of products namely electronics, the proposed solution can be applied to a wide range of industries. Machine learning algorithms are flexible and can be adjusted to changes easily. The New Product Forecasting framework considers various factors when generating forecasts, making it adaptable to different domains and product types.

Chapter 3: Methodology

3.1 Introduction

This study aims to research the role and impact of machine learning models in forecasting demand for new products. All the factors critical to new product forecasting have been identified in the literature review. The methodology section sets forth a suitable approach to validate these factors and constructs a logical framework to achieve the research objective. This section is divided into two parts; Methodology and Data Collection. The methodology section covers the research design and methods used to answer the primary research question. It presents a philosophical paradigm with epistemology, ontology, reasoning, implications and alignment of the selected paradigm with research objectives.

Data collection covers methods, scope, description, sampling techniques, validity and contribution of selected data sets. Discussions with experts to formulate and align the research methodology with research objectives are covered. Limitations and challenges observed during the data collection are discussed and mitigation strategies are developed. Potential risks are identified and refinement methodologies are formulated.

Research Objectives

This study builds upon the limitation of historical data in new product forecasting and incorporates machine learning algorithms to improve demand sensing. Several factors are considered and evaluated based on their impact on demand. To evaluate the impact of machine learning models on forecasting accuracy, a methodology must be devised to guide the data collection and analysis methods which will become a cornerstone to answer the research question.

3.2 Research Design

3.2.1 Research Paradigm: Pragmatism

The research paradigm selected for this project is pragmatism. Pragmatism represents an evaluation of ideas, methods, tools and proposals in the context of problem-solving, actions, usefulness and improvements (Saunders et al., 2019, p. 130). The main idea of pragmatism is to assess the implications of previously held concepts and their applications in practice to solve real-life problems (Tashakkori & Teddlie, 2016). Since this research is guided by current practices in new product forecasting, pragmatism is well suited to evaluate the impact and usefulness of forecasting models. Building upon current knowledge, this study proposes a robust machine-learning framework with improved accuracy, which aligns with the aspect of extending knowledge from previous experiences in pragmatism.

Epistemology (knowledge)

The pragmatist view of knowledge considers the relationship between actions and consequences and the implications of that knowledge in the real world (Tashakkori & Teddlie, 2016). The forecasting methods are specific to a situation or context and may vary depending on factors like geography, weather, market conditions, consumer preferences and more. These factors can be validated by evaluating their impact on demand forecasts and assessing the accuracy of results.

Rationale

The epistemology defined above provides forecasts based on data collected from sources like historical sales, user reviews, market surveys, social media and industry reports. Accuracy, reliability and error percentages are the measures used to evaluate the impact of data sources and analytical models. The result is a robust solution which generates accurate forecasts for quantitative data using statistical methods such as time series, regression analysis and qualitative data using text mining and sentiment analysis. By using an epistemological position of pragmatism, the use of machine learning and data-driven forecasts is investigated and a new approach is proposed.

Ontology (reality)

The ontology of this research study presents changes in reality through decisions and actions. In today's volatile marketplace, organisations must think strategically and deal with demand changes swiftly (Packowski, 2014).

Strategic decision-making has become data-driven and actions are now guided by historical information (sales, products, customers) and the results obtained from the forecasting solutions.

Rationale

Demographic information, customer behaviour and market trends all exist independently of subjective interpretations and should be incorporated into forecasting decisions to obtain accurate demand predictions (Fye et al., 2013; Goetze, 2011; Tsafarakis et al., 2011; Yamamura et al., 2022). Such factors reflect reality and directly impact forecast projections (Goetze, 2011; Matthias et al., 2017; Packowski, 2014). The ontology of forecasting decisions guided by the knowledge of these factors and assessment of their impact aligns with the research objectives. While considering the impact factors, this study seeks to establish the ground for the introduction of a machine learning solution in new product forecasting unlike traditional statistical methods and subjective approaches that are commonly adopted during pre-launch stages (Afrin et al., 2018; Fye et al., 2013; Kahn, 2006).

3.3 Philosophical Reasoning

Deductive reasoning means testing and confirmation of a general theory, where the logical conclusion is derived from a set of informed premises, whereas inductive reasoning means the formation of a theory where the conclusion is judged based on logical arguments (Saunders et al., 2019, p. 152). In this project, both inductive and deductive reasoning are necessary to accomplish research objectives.

Deductive Reasoning

Hypothesis 1: Culture has a profound effect on the demand and supply of a product.

Upon reviewing the literature it has been confirmed that culture has a great influence on customer's buying behavior (Goetze, 2011). People belonging to different cultures have different perceptions towards new products. There are certain psychological factors involved in customers' subjective perception of the product which adds uncertainty or randomness to the forecast (Meeran et al., 2017). Eroglu et al. (2023) analysed the effects of culture on firm-level inventory and demonstrated that in cultures with high uncertainty avoidance, managers tend to have increased inventory levels whereas cultures that promote long-term planning are more strategic and less reactive. Cultural events, practices and traditions shape consumer behaviour. National holidays, religious events, sports, Independence Day and many more are deeply embedded in one's culture and therefore, can be analysed to assess their impact on demand.

Hypothesis 2: Integrating customer reviews improves the prediction accuracy of forecasting models.

Online reviews have become a source of valuable knowledge, preferences and opinions. As much as text reviews are beneficial for customers to make informed decisions during product selection, they are equally valuable for firms to improve their products and quality (Huang et al., 2021; Zhao et al., 2022). Text reviews are found to be superior to product ratings because they are rich in sentiment (Huang et al., 2021). The emotional depth of text reviews on existing products enables companies to gain insights into the improvements required by customers. Building on this motivation this study computes the degree of sentiment in user reviews and discusses its potential to derive forecasts. Although user reviews cannot be used on their own for demand forecasts, their benefits in an aggregated framework should not be overlooked.

Hypothesis 3: Insights from market and competitor analysis increase the accuracy, reliability and relevance of demand prediction in new product forecasting.

Factors like industry growth projection, market competition and demographics assist in identifying market risks, economic shifts and emerging trends. Recognising macroeconomic projections like inflation, interest rate, market growth, consumer confidence, monetary predictions and industrial and operational projections is important for companies to operate in volatile and challenging environments (Valenzuela et al., 2023).

Analysis of seasonal patterns, market dynamics and demographics is critical for data-driven decision-making (Afrin et al., 2018; Goetze, 2011; Matthias et al., 2017; Sanders, 2017). This way forecasts can be customised to geographical locations and dedicated to specific demographics. This segmentation combines market trends,

financial metrics and external risks and ensures the forecasting results are grounded in real-world market conditions.

Inductive Reasoning

Hypothesis 4: Artificial Intelligence can uncover undiscovered relationships and patterns in the dataset.

One of the benefits of artificial intelligence that has been well documented in research is its ability to uncover relationships and patterns that traditional algorithms fail to detect easily (Aggarwal, 2016; Russell & Norvig, 2021; Saunders et al., 2019). Machine learning algorithms can uncover complex associations often overlooked by traditional statistical models (Kahn, 2006). This includes economic indicators, risk indicators, demand increases, cultural impact, purchase patterns, recession, natural disasters, and several other contextual and dynamic factors. This study seeks to find evidence of AI drawing patterns in specific sales instances and draw conclusions on its effectiveness.

3.4 Research Design Methods

Mixed Methods Approach

The mixed methods approach is a combination of qualitative and quantitative techniques to guide the collection of numerical and non-numerical data (Buchanan & Bryman, 2009; Tashakkori & Teddlie, 2016). This research aims to study the potential of triangulation (mixed methods) to achieve improved accuracy in new product forecasting. The solution will leverage the benefits of quantitative and qualitative data and different analysis techniques to evaluate diverse data sources and forecasting models.

A forecasting framework is formulated that integrates diverse data sources to forecast the demand for new products. The data sources are categorised into two types; quantitative and qualitative. Quantitative data includes sales data, market research, consumer demographics and financial metrics. Qualitative data constitutes textual reviews, social media (Twitter/WeChat) and YouTube video reviews.

Justification/Rationale

The choice of a mixed-method approach should be guided by the procedures that answer the research question (Buchanan & Bryman, 2009; Tashakkori & Teddlie, 2016)**.** Due to the absence of significant historical data in this study, many diverse data sources have been covered to propose a solution. The biggest driver for forecasting decisions in statistical methods is historical sales (Ching-Chin et al., 2010; Lee et al., 2014; van Steenbergen & Mes, 2020). When companies predict sales of a new product, either diversification or line extensions, there is a noticeable absence of sales data available to derive accurate demand forecasts (Kahn, 2006). Many companies in this situation conform to marketing data, expert advice and human judgment (Feiler & Tong, 2022; Lee et al., 2014).

This study adopts a different approach and proposes a solution that converges insights from sources that are often implicit, impersonal and contain a lot of noise (Feiler & Tong, 2022). It provides a more nuanced approach to new product forecasting compared to traditional frameworks. A parallel mixed method design including quantitative and qualitative data sources is constructed, producing a robust framework. This framework is implemented and the results are assessed for interpretation. The effectiveness of mixed-method research is measured by calculating the accuracy of predictions derived from machine learning models.

3.5 Alignment with Research Objectives

After validating key factors that influence new product forecasting and the challenges faced by forecasting models, the remaining research objectives are the development of a mixed-method forecasting framework and the evaluation of machine learning models when implementing that framework. The data collection for this research study is guided by the objective of addressing the applications, results, accuracy, computational power and scalability of machine learning algorithms. Utilising diverse data sources enables the researchers to identify the plausible risks of combining internal and external data in forecasting methodologies. The risks and barriers associated with data collection and mitigation strategies followed by refinement methodologies are explored in detail.

Research Objective 3: Develop a mixed-method approach to the demand forecast of a novel product with limited sales data.

New Product Forecasting Framework

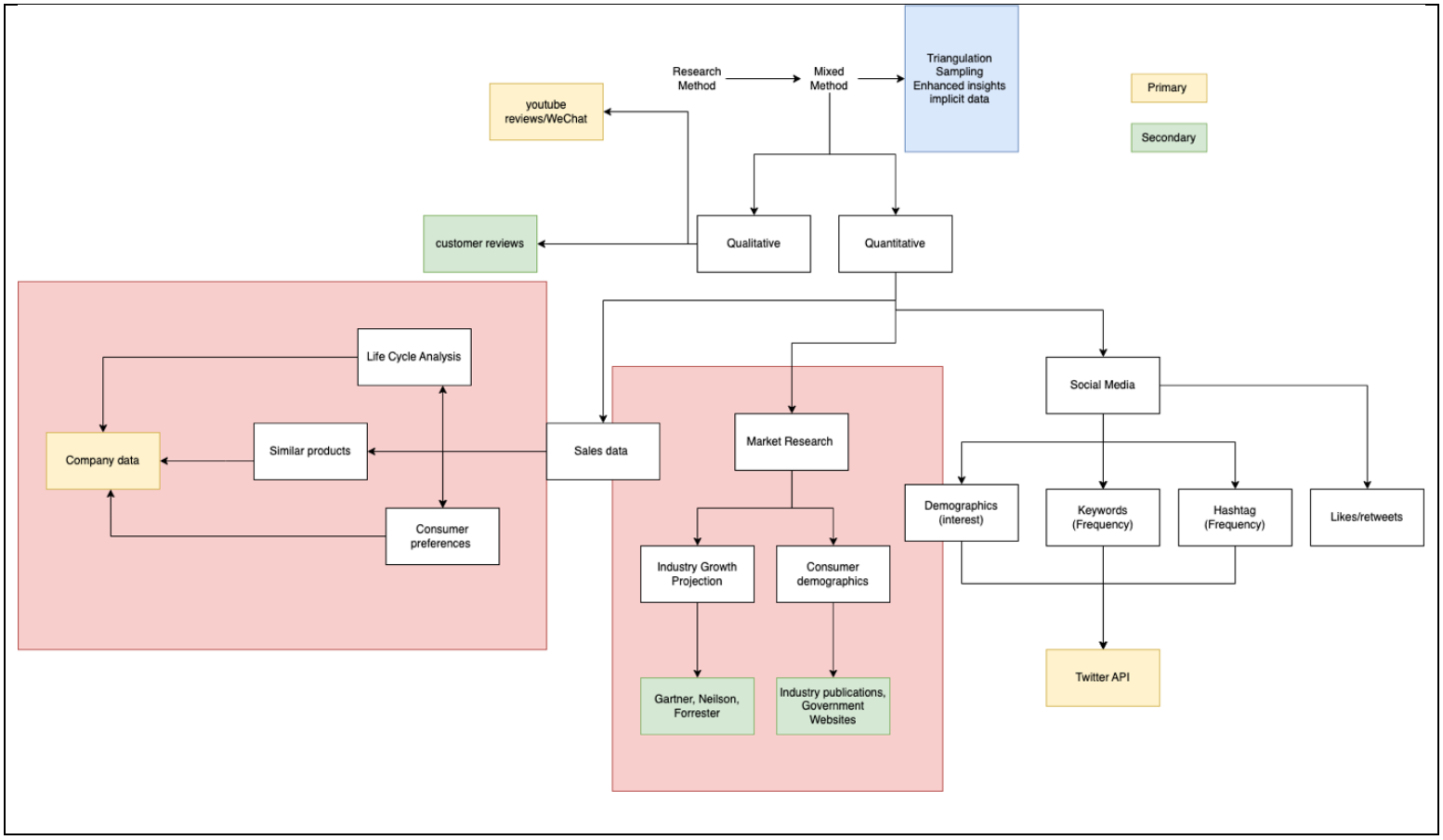


Figure 3.1: New Product Forecasting Framework with diverse data sources (created by author)

3.6 Data Collection

Research Objective 4: Discuss the feasibility of various data sources to enhance forecasting algorithms.

3.6.1 Quantitative Data Collection

M5 Sales Forecasting Dataset from Walmart

Type: Secondary

Source: The dataset is publicly accessible and can be found at M5 Forecasting - Accuracy | Kaggle

Industry: Retail

Description
The M5 dataset is the time series data provided by Walmart in collaboration with Kaggle in 2020. It is 5th dataset in a series of forecasting competitions hosted by Walmart to predict future sales. The dataset has numerical calendar sales about multiple product categories and store information. The stores are spread around three states namely California (CA), Texas (TX) and Wisconsin (WI) and represent a time series hierarchical unit sales (Makridakis et al., 2022). (Note: See Figure 4.2 on page 28 for full hierarchy).

Scope

The dataset contains the following information:

1. Sales: Contains product information such as product ID, store and department.
2. Calendar.csv: Represents dates when products were sold. This includes special occasions like Christmas, holidays, weekdays and other special events that can affect sales of a product and create anomalies in data.
3. Prices: Sales price of products at different stores.
4. Item: Products are divided into HOUSEHOLD, HOBBIES and FOOD categories.

Sample Size

Dataset	Number of Records
Calender.csv	1970 (29/01/2011 – 19/06/2016)
sales_train_evaluation.csv	30,490
sales_train_validation.csv	30,490
sample_submission.csv	60,980
sell_prices.csv	30,490

Table 3.1: M5 dataset files with the total number of records

Justification for Selection

- Discussions with Experts
  To approach data collection for this study several experts were consulted due to the sheer complexity of data necessary to meet research objectives. One of these experts was Dr. Sven F. Crone, lecturer for Forecasting for Logistics and Supply Chain Management at the University of Lancaster. During the discussion, it was established that organisations generally do not invest in machine learning during the pre-launch stage. Line extensions as opposed to diversification (see Figure 2.1) are the most commercial quadrant for firms to invest their R&D resources.

  For the time frame restrictions, the experts advised to utilise the secondary dataset (M5) for quantitative sales data in the pre-launch stage of product development. Secondary datasets are open source with less ethical risk and no copyright issues. The M5 dataset allowed the same quality as the primary dataset because it was collected by Walmart for internal use and later publicised to improve forecasting solutions.

- Ethical Considerations
  Privacy Regulations
  The sales data has sensitive information about customers which must be adhered to GDPR (General Data Protection Regulation) to provide individuals right to protection of their data (GDPR, 2013). The M5 dataset and its predecessors have been anonymised and removed from any identifiable information and are a suitable choice for this study.

  Non-Disclosure Agreement (NDA)
  An NDA must be signed to regulate the sharing of sensitive information. In case any clause is breached, the researcher would be held liable for a hefty fee. To steer clear of this issue, utilising a carefully curated secondary dataset was found to be an excellent choice.

Barriers and Challenges

- **Data Availability**
  The M5 dataset is publicly available but there is a requirement for a Kaggle account, user verification and accepting terms and conditions before the data can be accessed.
- **Incomplete Data**
  Although the items are categorised in their respective categories there are no subcategories, for example, HOBBIES could be divided into art, sports, books, music and more. The lack of granulation restricts the creation of complex user-profiles and subsequently accurate targeted forecasts.

- **Data Integration and Computational Resources**
  To forecast demand with respect to each department and special occasions, the files (See Table 3.1) need to be integrated into one dataset. This can be challenging as each file contains different columns with distinct formats and thousands of entries, requiring significant processing power.
- **Missing Data**
  Although the M5 dataset is complete, sales data commonly has empty cells and blank rows. The missing data should either be removed or replaced with appropriate values during data preprocessing to avoid biased results and inaccuracies. Handling missing data for large datasets is often a time-consuming and memory-intensive process which can be a challenge.

Market Research

Type: Secondary

Industry: Consumer Electronics
Description

Harvard Business Review

HBR is a management magazine managed and published under the umbrella of Harvard Business Publishing, a non-profit organisation and an affiliate of Harvard Business School. It bridges the gap between academia and industry offering insights into consumer trends and interests (harvard business review, 2016). The publications discover findings about challenges faced by the companies when managing demand for new products. It also advises supply chain managers on the current developments in academia that can be helpful to them in their forecasting endeavours.

Gartner

Gartner is an American consultancy firm that conducts technological research in the industry and shares this information over private and public channels. It acts as an advisory for organisations to make critical decisions (Gartner, 2023). Industry growth projection is an essential factor that is considered when driving demand for products. Products, especially in their early phase of the lifecycle, are vulnerable to external factors (Kahn, 2006; Levitt, 1965; Sondhi, 2008) and determining their growth projection is critical to predicting their future demand.

Neilson

Neilson is an American data analytics firm that measures audience behaviour and interest (Nielson, 2024). It includes retail measurement, digital audience measurement, ad campaigns, consumer panels and financial reports. This provides companies with knowledge about the global consumer trends in diverse markets. For companies doing line extensions (see Figure 2.1) the data is usually driven by previously deployed in-line products and received customer feedback. On the contrary for new products, the company is either treading in a new market or creating a new market on its own (Kahn, 2002, 2006). In this scenario, interest measurement data and comprehensive insights on market trends would support companies in their strategic decision-making process.

Government Websites

The data collected by government agencies is reliable and comprehensive. These agencies collect data for a wider population over greater geographical areas. For example, consumer expenditure surveys, employment statistics, regional data and annual business reports help companies determine their target audience based on demographic information.

Justification

As mentioned previously, forecasting for novel products is the riskiest for a company (Kahn, 2006). During pre-launch sales data or the lack of it presents a major risk when forecasting demand in the earliest phase of the product lifecycle (Levitt, 1965). In this case, external data such as industry journals and publications can be leveraged to make informed forecasting decisions.

Scope

The forecasting framework in Figure 3.1 is an amalgamation of several sources of internal and external data. For external data, industry journals are used to analyse market trends, and public interest and infer market growth projections before new product launches. Demographic information is important for customer segmentation so that businesses invest their marketing resources to target potential customers. Moreover, the information provided by industry publications and consultancies is a reliable source as they have access to the private data of organisations and form a bigger picture of external factors.

Barriers and Challenges

- **Relevancy**
  Data collected by firms could be lagging due to the extended process of data collection, analysis and distribution. In some industries like retail and high-tech gadgets, trends tend to change quickly leading to the data provided obsolete or irrelevant.
- **Accessibility**
  Many journals have limited free accessibility for example, HBR provides only three free article reads without a subscription. Other online data platforms like Statista have a subscription fee to access relevant statistics.
- **Data Fragmentation**
  To gain comprehensive insights into the industry, data from many external sources should be interpolated together. This creates a challenge as each data source has data stored in different formats with varying structures and definitions.
- **Granularity**
  Demographic information collected by the government and local agencies is generalised over geographical location in broader categories. This generalisation is not suitable for companies who are seeking to target one specific area or group of consumers.

3.6.2 Qualitative Data Collection

Amazon Review Dataset

Type: Secondary

Source: Publicly available dataset can be accessed from https://amazon-reviews-2023.github.io

Industry: E-commerce

Description
The Amazon Review Dataset (2023) is a large dataset collected by e-commerce giant Amazon. It contains product reviews posted by users and is widely used in machine learning and sentiment analysis. It provides qualitative insights into customer reviews of Amazon products spanned across a period of 27 years (1996-2023).

Sample Size

Reviews	**571.54M**
Users	**54.51M**
Items	**48.19M**
Time Span	**May'96 - Sep'23**

Table 3.2: Amazon Reviews Dataset Size in millions, (Amazon, 2023)

The **items** dataset includes the following features:

Field	Type	Explanation
main_category	str	Main category (i.e., domain) of the product.
title	str	Name of the product.
average_rating	float	Rating of the product shown on the product page.
rating_number	int	Number of ratings in the product.
features	list	Bullet-point format features of the product.
description	list	Description of the product.
price	float	Price in US dollars (at time of crawling).
images	list	Images of the product. Each image has different sizes (thumb, large, hi_res). The "variant" field shows the position of image.
videos	list	Videos of the product including title and url.
store	str	Store name of the product.
categories	list	Hierarchical categories of the product.
details	dict	Product details, including materials, brand, sizes, etc.
parent_asin	str	Parent ID of the product.
bought_together	list	Recommended bundles from the websites.

Figure 3.2: Amazon Reviews, Item Attributes and their Description (Amazon, 2023)

The **user** data includes the following features:

Field	Type	Explanation
rating	float	Rating of the product (from 1.0 to 5.0).
title	str	Title of the user review.
text	str	Text body of the user review.
images	list	Images that users post after they have received the product. Each image has different sizes (small, medium, large), represented by the small_image_url, medium_image_url, and large_image_url respectively.
asin	str	ID of the product.
parent_asin	str	Parent ID of the product. Note: Products with different colors, styles, sizes usually belong to the same parent ID. The "asin" in previous Amazon datasets is actually parent ID. **Please use parent ID to find product meta.**
user_id	str	ID of the reviewer
timestamp	int	Time of the review (unix time)
verified_purchase	bool	User purchase verification
helpful_vote	int	Helpful votes of the review

Figure 3.3, Amazon Reviews, User Attributes and their Description (Amazon, 2023)

Note: The snapshot of the review dataset is not provided in the official documentation, so the spreadsheet of reviews and their attributes is available in Appendix B, Table B.1.

Justification

- **Relevancy to Research**

One of the project objectives is to utilise machine learning techniques in product forecasting. The Amazon dataset is commonly used in natural language processing, a subset of machine learning, to determine the sentiment of users. Since line extensions are extended from previously deployed products (Kahn, 2006), it is necessary for organisations to incorporate user sentiment to determine demand for their next-in-line products. Therefore, text mining will be performed on the review dataset to categorise reviews based on specific products and derive the expected degree of satisfaction from the results.

- **Anonymity**

The dataset has been publicly available for many years and has been reviewed for any privacy concerns. All reviews are anonymised with no Personally Identifiable Information (PII), ensuring the privacy of the users.

- **Consumer Behaviour and Trend Analysis**

Customer feedback is important when determining demand for the next period as positive customer interest is directly proportional to higher sales (Meeran et al., 2017; Packowski, 2014). Text reviews are rich in sentiment and provide explicit insights into customers' preferences and behaviour. Furthermore, these sentiments can be mapped on time frames to track trend changes over time.

- **Diverse Product Categories**

One of the limitations of the M5 dataset as explained above is the lack of granularity (subcategories) in the product dataset. In the Amazon reviews dataset, products are divided into a diverse range of categories allowing for additional depth in the data.

Barriers and challenges

- **Incomplete reviews**

Some reviews may involve spam words or lack the keywords necessary to determine sentiment. For a text to transform into sentiment leading to meaningful insight, it is important to have a semantically correct review.

- **Bias**

The dataset does not represent all customers since many customers do not post their reviews online. Consequently, it can introduce bias in the dataset as it represents a certain group of customers and not the total population of buyers.

- **Inconsistencies in reviews**

Most often, reviews have errors and slang inside the text which is difficult for a pre-defined language model to parse. Thus, even though the reviews may represent strong sentiment they won't be captured in the algorithm due to the intricacies and constantly evolving nature of slang that cannot be easily translated into computer programs.

Risk Mitigation Strategies

1. **High-power processor**
   To mitigate the challenge of required power to run the machine learning model, a virtual machine with 64 GB Random Access Memory (RAM), a 13th Gen Intel Core i7 Processer and a dedicated graphics card were installed. This would allow for the models to run smoothly.

2. **Missing Data**
   Data with missing values will be approached using the following strategies:

   - Replacing the missing value as null.
   - Replacing the blank cell with the most occurring value (mode).
   - If the filled cell values are similar in range then the blank cell will be replaced by the median of all cells.

- If the number of missing values is minimal compared to the size of the dataset, or the missing attribute is product review then these rows would be simply removed.

If removing the row or replacing it with a new value would skew results then missing data can be replaced by machine learning algorithms like K-Nearest Neighbors (KNN) or classification models that predict the missing value based on similarities between other attributes.

3. **Relevancy**
   To ensure data is relevant and appropriate for new product forecasting the latest 2023 version of the dataset is used for analysis. There are 2013, 2014, and 2018 versions with smaller sizes but due to the requirement of novel products and latest reviews, any datasets before 2020 will be discarded so that the information can be classified as pre-launch data for diversification and line extensions.

4. **Bias and Noise**
   The dataset will undergo a thorough preprocessing stage to remove any noise. Outliers, such as peak seasons and holidays, will be highlighted during data analysis. Randomness will be introduced by using shuffle split and random sampling in the dataset to remove bias before training the machine learning model.

5. **Granularity**
   Aggregation can be applied to remove the generalisation from the collected data. Any similarities between the three data sources mentioned previously would allow for the integration of one or two sources. This can lead to demographic data becoming targeted and more specific towards the products in consideration.

3.7 Alternatives for Data Collection

For initial research methodology, Michael Mortenson who is an Associate Professor at Warwick Business School and a leading industry expert in Business Analytics and Data Science, was consulted. In the worst-case scenario of missing primary and secondary sales data, alternative data collection methods were suggested. Scikit-Learn and NumPy are Python libraries consisting of functions that produce random data and follow a Gaussian distribution. This distribution represents a bell curve. TimeGPT was another alternative that was suggested. It is a generative transformer model adopted from General Adversarial Networks (GANs) that creates synthetic time series data in real-time (Nixtla, 2023).

Generative time-series data does not capture the complex nuances found in real-world data. It is randomly generated data observations that do not reflect the relationships and trends in historical data. Due to missing historical context and complexity, this type of data is not suitable for critical business decisions or in this context, an exploratory research study.

3.8 Refinement Methodologies

1. Feature Selection
   In the context of the volume of big data, the datasets it represents are exceptionally large and complex making traditional processing systems insufficient to process them. A machine would require a significant amount of computing power to perform operations on the dataset. Feature selection can be applied to reduce the size of the data. Through feature selection techniques like Principal Component Analysis (PCA), only the features that are relevant, correlated and important to the research objectives would be selected, discarding the rest. Features that improve the performance of the forecasting model will be identified and considered in the final dataset.
2. Data Sampling
   Creating a smaller subset by applying filters on relevant attributes reduces the size of the total subset and makes it manageable for the system to process. This technique improves computational efficiency and simplifies analysis while preserving the important characteristics of the original data set.

3. Cross-validation
   This technique divides the dataset into training and testing subsets to assess the performance of a model. By splitting data into subsets, the model ensures there is no overfitting of data points and the estimates are reliable for unseen data.
4. Iterative Refinement
   It refers to a process that improves the performance of a forecasting model progressively through multiple iterations. Performance metrics like Mean Absolute Error (MAE) and Mean Squared Error (MSE) are used to re-assess the model results. The forecasting models are revisited and fine-tuned after each iteration. Based on the performance feedback the values are adjusted to ensure continuous improvement of the model's accuracy.

Chapter 4: Results & Analysis

4.1 Introduction

This chapter will address the following research objectives by highlighting the assumptions, justification of analysis methodologies, results, confidence factor, accuracy and user bias.

- **Research Objective 5: Evaluate the effectiveness of machine learning algorithms in improving forecast accuracy.**

- **Research Objective 6: Identify limitations and risks associated with using machine learning algorithms for demand forecasting and mitigation plans to enhance the effectiveness of predictive models.**

Given that all collected datasets differ in structure and the type of data they represent, different analytical approaches are required to analyse each dataset. The discussion follows a common data analysis pipeline, with the data collection already discussed in the methodology.

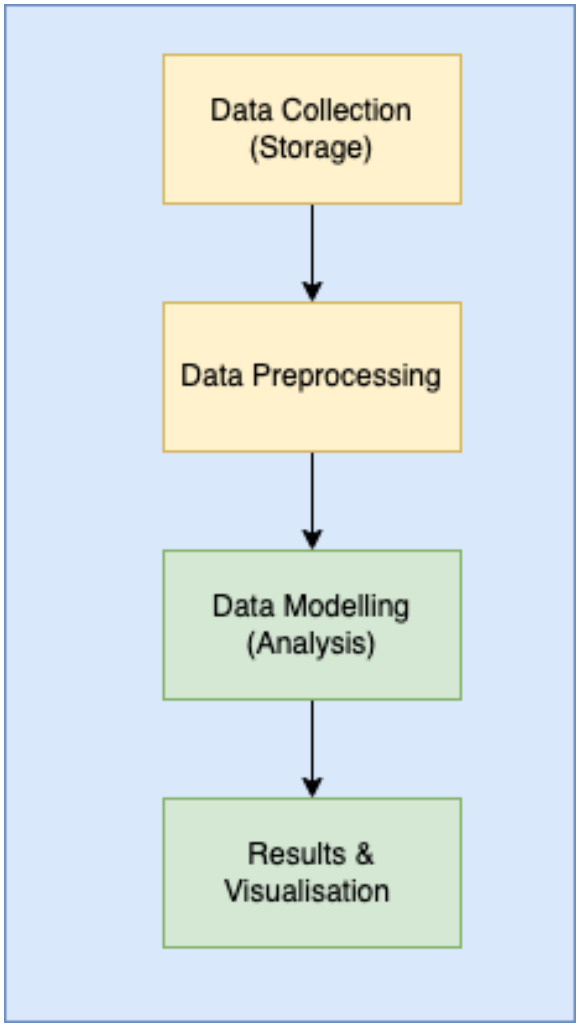


Figure 4.1: Data Analysis Pipeline Stages

To assess the role of machine learning in new product forecasting, three distinct datasets from different sources were selected. The qualitative and quantitative attributes represented by the datasets were the motivation for this selection. These attributes can be combined to represent a mixed-method framework for new product forecasting.

4.2 M5 Forecasting Accuracy Dataset

The Walmart dataset is structured in the following hierarchy:

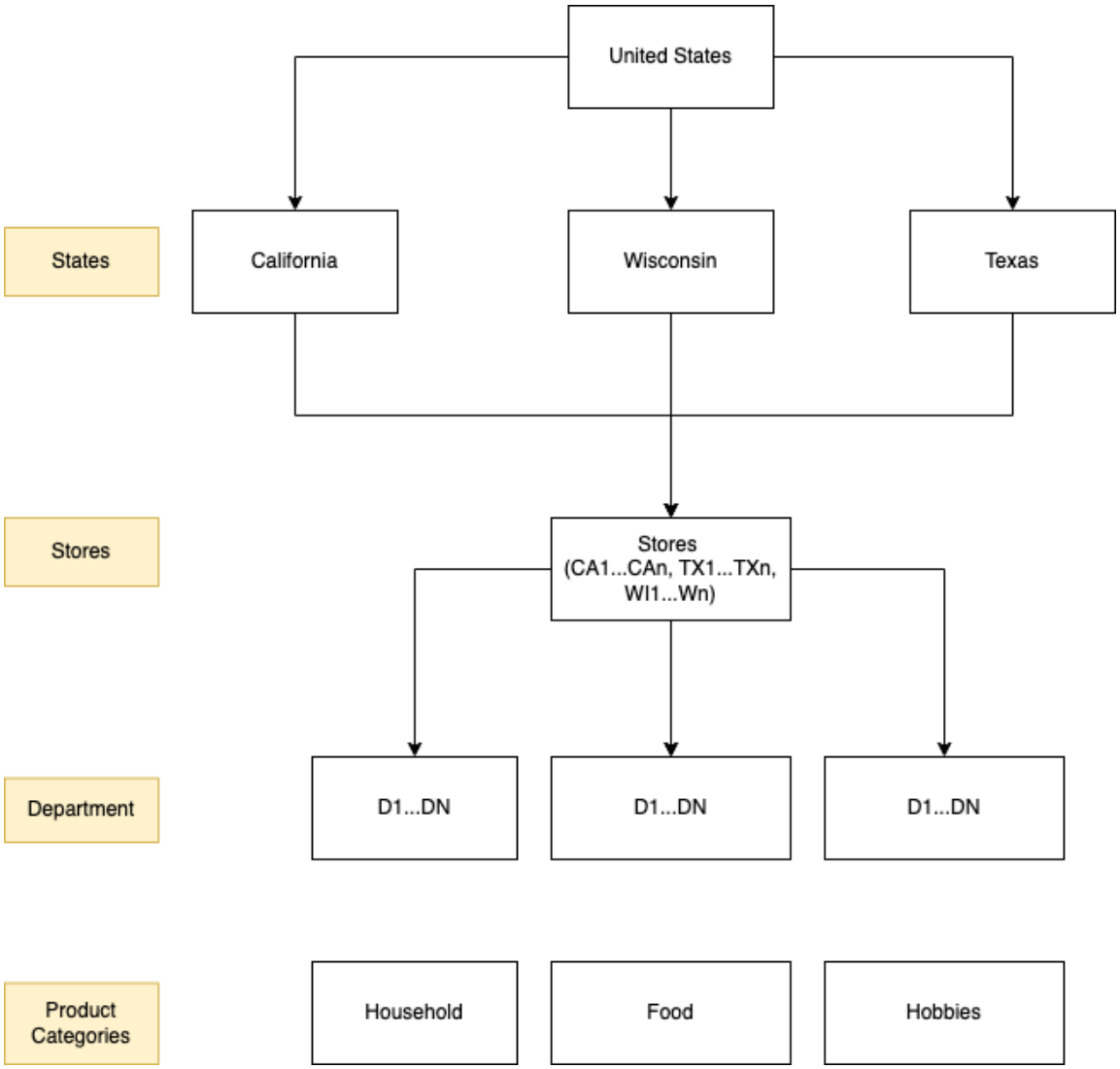


Figure 4.2: M5 dataset hierarchy represented by primary attributes in all files

Assumptions

1. Product Category Selection
   As this study gravitates towards a specific type of product mainly electronics, the household category was selected. This selection was based on the assumption that the food category is not suitable due to its minimal shelf life (Galic et al., 2009) and hobbies have little in common with electronics. As household items generally have a longer shelf life and product life span compared to other categories, the dataset has been restricted to household products only.
2. Granularity
   The forecasting results in this study are limited to nationwide sales of a product, representing combined product sales from all states, departments and stores. This is because the forecasting accuracy was found to be gradually decreasing at the lower levels of the hierarchy. Therefore, a more balanced approach to forecasting at a countrywide scale is considered in the research.

4.2.1 Data Preprocessing

Following are the files and columns of the M5 dataset;

Files	Attributes
calendar.csv	date, wm_yr_wk, weekday,wday, month, year, d, event_name_1, event_type_1, event_name_2, event_type_2, snap_CA, snap_TX, _snap_WI
sales_train_validation.csv	id, item_id, dept_id, cat_id, store_id, state_id, d_1, d_2, d_3... d1913
sample_submission.csv	Not Applicable for this study
sell_prices.csv	Store_id, item_id, wm_yr_wk , sell_price
sales_train_evaluation.csv	Not Applicable for this study

Table 4.1: M5 dataset files and their respective attributes

Missing data: All the rows with empty columns that were identified, were either removed or replaced by the average numerical value of other neighbouring rows.

Duplicates: Repeating rows were removed to reduce bias and prevent overfitting in the model.

Incorrect Entries: Although no incorrect data entries were found, the dataset was analysed thoroughly to validate that all entries conformed to the expected format, range and data types.

Feature Selection

The files in the M5 dataset encompass many records, occupying a large amount of memory (Table 3.1). Using the whole file increases the complexity, storage and computational power required by the algorithm. Thus, features were selected based on their impact and relevancy to the research objectives.

For example, the dataset shown in Appendix A Table A.2 was reduced from 1947 columns to 2 columns after applying feature selection (see Figure 4.3 for the resulting data frame).

	item_id	sales
565	HOUSEHOLD_1_001	678
566	HOUSEHOLD_1_002	540
567	HOUSEHOLD_1_003	1122
568	HOUSEHOLD_1_004	2639
569	HOUSEHOLD_1_005	2322
...	...	...
29048	HOUSEHOLD_2_512	609
29049	HOUSEHOLD_2_513	608
29050	HOUSEHOLD_2_514	209
29051	HOUSEHOLD_2_515	120
29052	HOUSEHOLD_2_516	247

10470 rows × 2 columns

Figure 4.3: Total number of household products in the M5 dataset

Product Subset

The dataset was first filtered to provide rows only with household item sales (Figure 4.3). Then it was further segmented based on the popularity of a single item. It is assumed that the household item with the most sales records is the most popular in the household category and therefore should be considered for the analysis.

	item_id	sales
0	**HOUSEHOLD_1_118**	**44018**
1	HOUSEHOLD_1_459	37392
2	HOUSEHOLD_1_334	37226
3	HOUSEHOLD_1_303	35226
4	HOUSEHOLD_1_521	30484
5	HOUSEHOLD_1_459	30291
6	HOUSEHOLD_1_465	29999
7	HOUSEHOLD_1_351	28882
8	HOUSEHOLD_1_019	28669
9	HOUSEHOLD_1_110	28545

Figure 4.4: Most sold household product making up 1.64% of all household sales throughout the recorded period.

Sequel command INNER JOIN was applied to merge the attributes **date**, **events** and **department** from calendar.csv and **cat_id** and **dept_id from** sales_validation.csv for HOUSEHOLD_1_118 to determine its time series data. However, this command failed to execute because it was computationally intensive for the program to run on thousands of rows. To solve this problem, all the departmental sales for HOUSEHOLD_1_118 in sales_validation.csv were combined with sales data from calendar.csv using a logic called LEFT JOIN. The calendar.csv has been transformed from 14 to 1956 columns (See Appendix Figure A.1) depicting all department unit sales for HOUSEHOLD_1_118 on all calendar days. These columns were added together in a new column "*aggregate_department sales_countrywide_for_HOUSEHOLD_1_118*" representing total department sales for the product across all states on each calendar day.

date	year	aggregate department sales countrywide for HOUSEHOLD_1_118
2012-01-01	2012	11
2013-01-01	2013	0
2014-01-01	2014	11
2015-01-01	2015	4
2016-01-01	2016	1

Figure 4.5: filter "event_name_1 = "NewYear providing aggregated department sales for HOUSEHOLD_1_118.

4.2.2 Exploratory Data Analysis

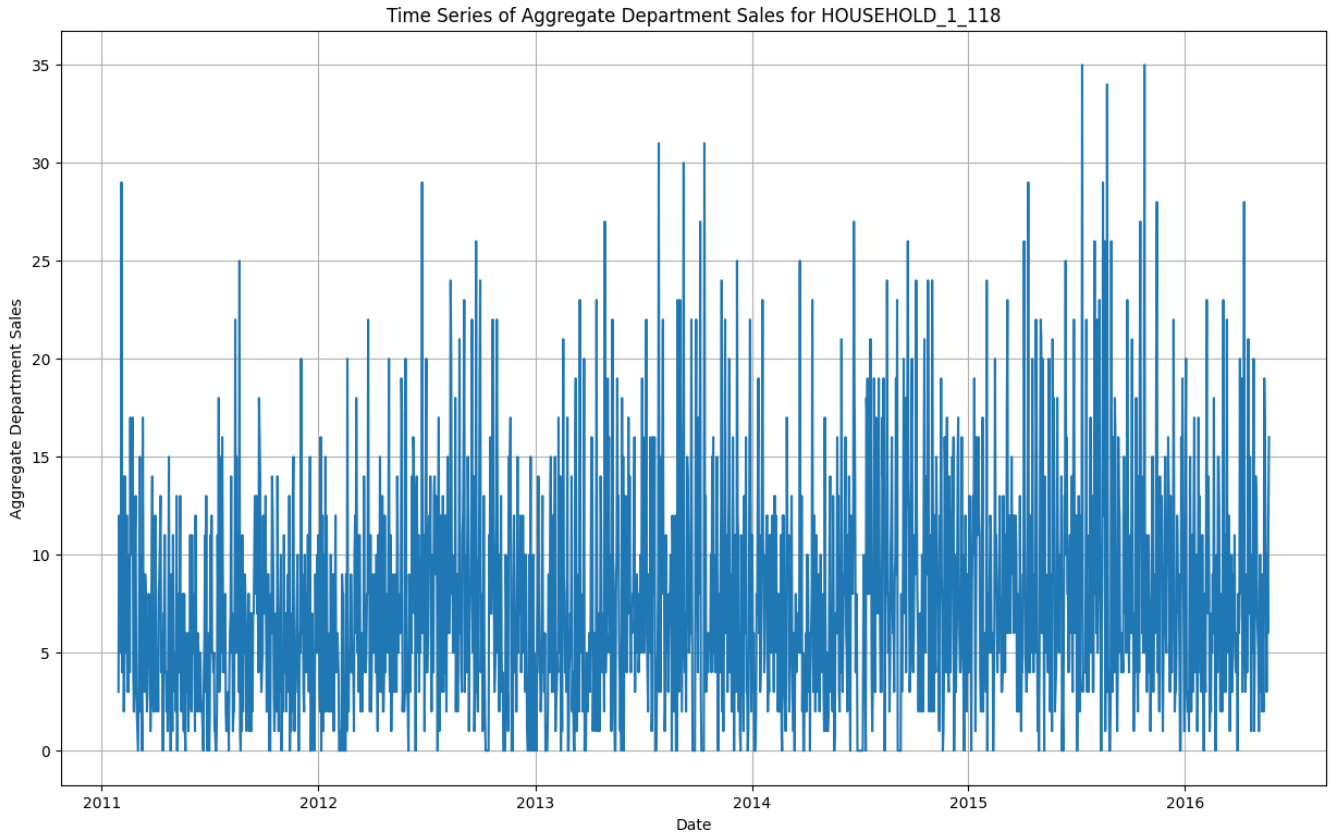


Figure 4.6: Nationwide sales (3 states) of HOUSEHOLD_1_118 across the years

The newly added column of aggregated sales is mapped onto a time series plot representing the total sales for HOUSEHOLD_1_118 in the United States throughout the years. A noticeable shift in the demand can be observed during the closing dates and opening dates in Figure 4.7. These events are assumed to be either Christmas or New Year, resulting in outliers or anomalies in the dataset compared to average calendar days.

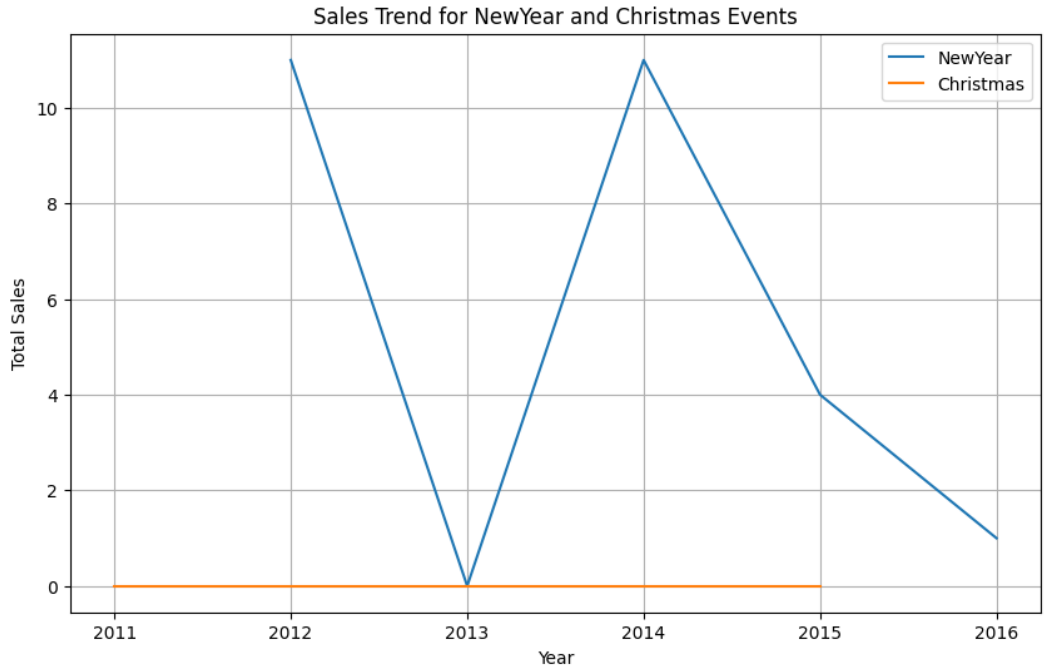


Figure 4.7: Trend Analysis of HOUSEHOLD_1_118 unit sales on New Year and Christmas in line graph. The specific values used to generate this graph can be found in Appendix A, Figure A.2.

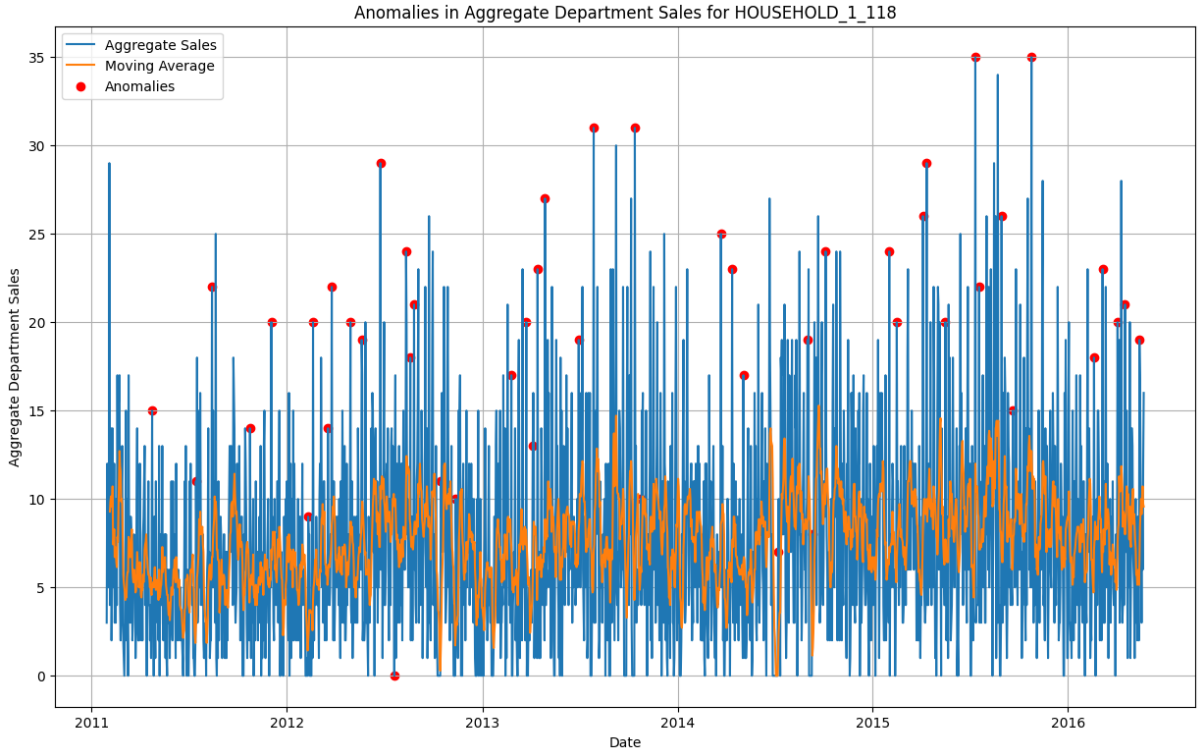


Figure 4.8: Aggregate department sales for HOUSEHOLD_1_118 with anomalies and moving average.

Moving average and standard deviation are calculated for each calendar day and anomalies are identified with a threshold value of 2 in Figure 4.8.

Next, the event names are extracted from the dataset and mapped onto the visible spikes in Figure 4.9. Several events with significant decline overlap making them unreadable. The majority of the peak days are not highlighted meaning they are not associated or labelled with any events. The missing entries could be due to oversight during data collection or the store observed an unexpected demand on an average calendar day.

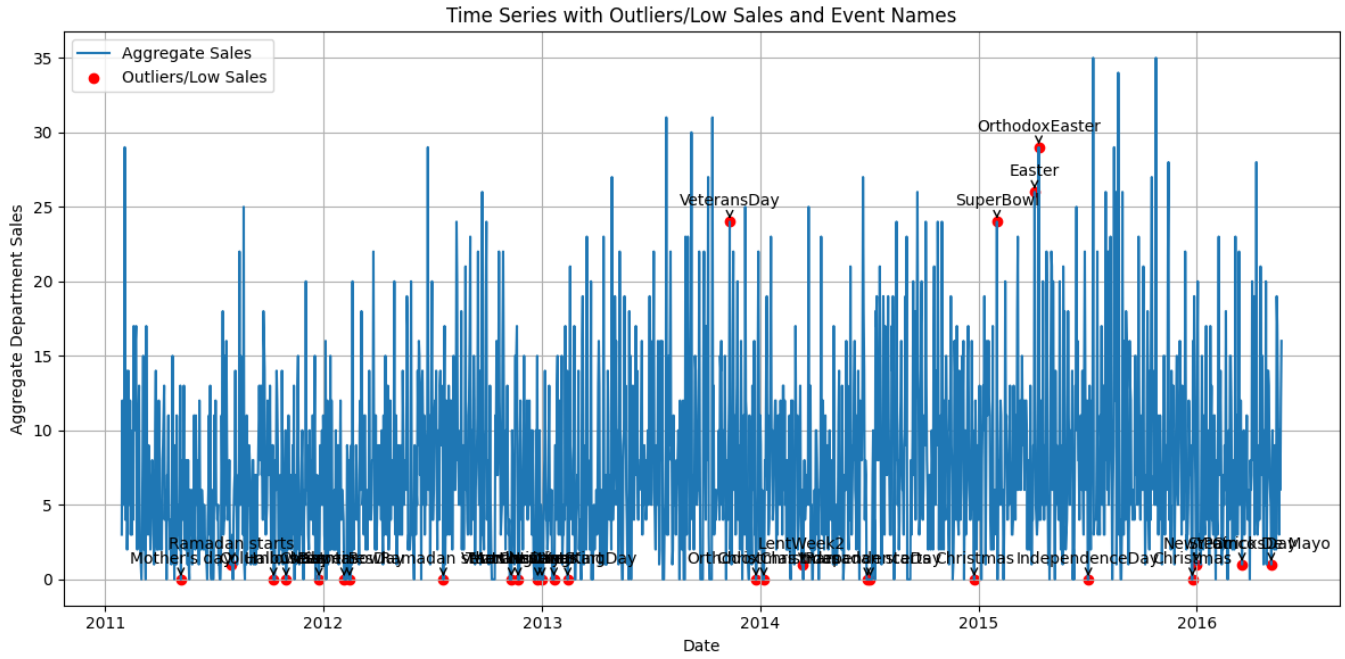


Figure 4.9: Aggregate department sales for HOUSEHOLD_1_118 with anomalies labelled as the value of "event_name_1" attribute.

4.2.3 Data Modelling

Hypothesis 01: Culture has a profound effect on the demand and supply of a product.

One of the factors identified in the data that greatly affects the demand for a product is the national, international and religious event ceremonies (see Figure A.3 in Appendix A). For example, the demand increases for the product on Super Bowl, one of the most watched sporting events in the U.S. and a sudden drop is observed on national holidays such as New Year and Independence Day recognised as federal holidays in the United States. Amongst many recorded events like Christmas, Independence Day, Easter and many more, two events show a repeated pattern of decline in demand throughout the years. These events are **Christmas** and **New Year** where the stores are usually either closed or receive very few visitors (Figure 4.7).

For better visibility, the time series data of 2016 in Figure 4.10 can be used to highlight the impact of cultural events on product demand.

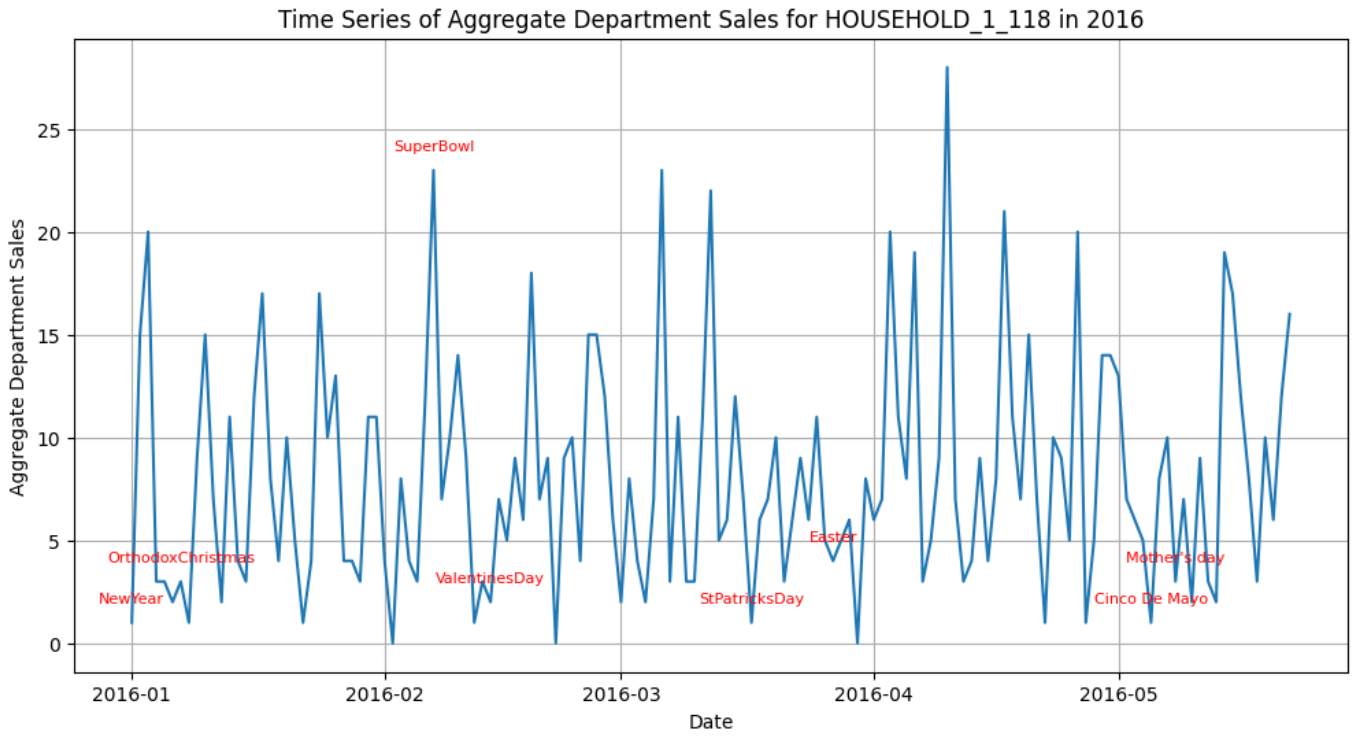


Figure 4.10: Time Series for all unit sales of HOUSEHOLD_1_118 on recorded events in 2016

Objective: Analyse the time series data and predict the projected sales of **HOUSEHOLD_1_118** at **NewYear** in 2016 based on the historical data available from the years 2011 to 2015.

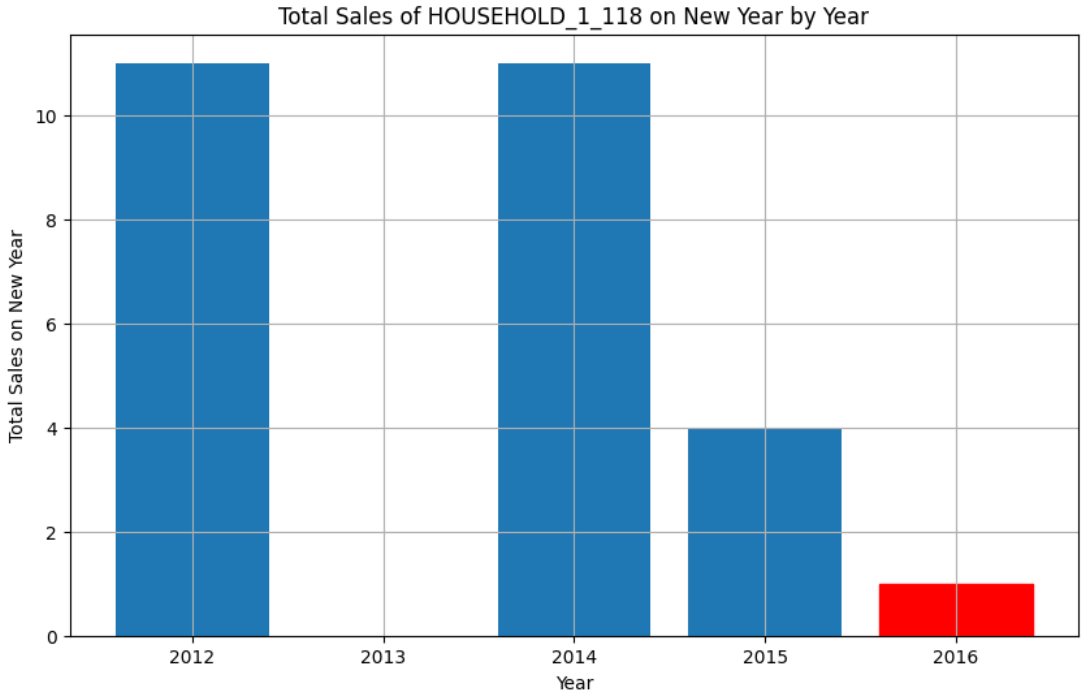


Figure 4.11: Total Sales of HOUSEHOLD_1_118 across all New Year dates, 2016 is highlighted in red and shall be considered as a test variable to calculate the accuracy of the machine learning algorithms.

To compare the accuracy of machine learning models and suggest suitable forecasting model, three algorithms are selected to analyse the time series plot. These are linear regression, ARIMA and random forest. The algorithms were selected based on their ability to process time series data effectively and provide accurate results (Kim & Shin, 2016; Sanders, 2017).

4.2.3.1 Linear Regression

To validate the models and assess their accuracy, actual data need to be tested against predicted values. For this reason, observations ranging from 2011 to 2015, (highlighted in blue in Figure 4.11) are used as training data denoted by '*X_train'* and '*y_train'* and sales in 2016 (in red) are used as the testing variable denoted by '*X_test'* and '*y_test'*. The model does not predict any values beyond 2016 because there is no way to assess the accuracy of those forecasts.

Linear regression examines the relationship between a dependent variable and an independent variable to predict outcomes. For this study, the dependent variable denoted by Y **is aggregate department sales** that depend on the independent variable **date** represented by X.

```
# Split data
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)
```

The variable '*test_size'* is 0.2 as linear regression assigns 20% of the data for testing and 80% for training by default. The random state was set to reduce the bias in the training data.

```
X = result_df[['date_ordinal']]
y = result_df['aggregate department sales countrywide for HOUSEHOLD_1_118']
```

(See Figure A.4 in Appendix A for the full code of linear regression)

4.2.3.2 ARIMA

The ARIMA model predicts future demand based on past values. It does so by averaging all the observations from the previous periods in a time series and using a weighted sum to predict future demand (G. Wang et al., 2016).

(See Figure A.5 in Appendix A for the full code of ARIMA)

"*UserWarning: Too few observations to estimate starting parameters for ARIMA and trend. All parameters except for variances will be set to zeros."*

The following error was generated by the ARIMA model due to a lack of sufficient observations. The minimum number of observations required by the basic ARIMA model is 50 (Box et al., 2008). This could be the reason why ARIMA is the least-performing model in this iteration having the maximum error percentage out of all algorithms (for full error see Figure A.6 in Appendix A).

4.2.3.3 Random Forest

Random Forest is an algorithm that is widely used in classification and regression tasks (Y. Wang et al., 2018a). It constructs multiple decision trees to reach a singular, accurate prediction. Although the algorithm predicted value that was relatively closer to the actual sales, there was a high percentage of error in results.

(See Figure A.7 in Appendix A for the full code of random forest)

Sales of HOUSEHOLD ITEM **1_118** on New Year of 2016

The Accuracy and Root Mean Squared Error of the results are as follows:

Model Performance Metrics

Algorithm	Actual Value	Predicted Value	Accuracy %	Mean Squared Error %
Linear Regression	1	2.2	-20	120
Random Forest	1	2.43	-43	143
ARIMA	1	10.221	-822.067	922.067

Figure 4.12: Python Screenshot of model performance metrics for all machine learning algorithms including the Accuracy per cent and Mean Squared Error per cent.

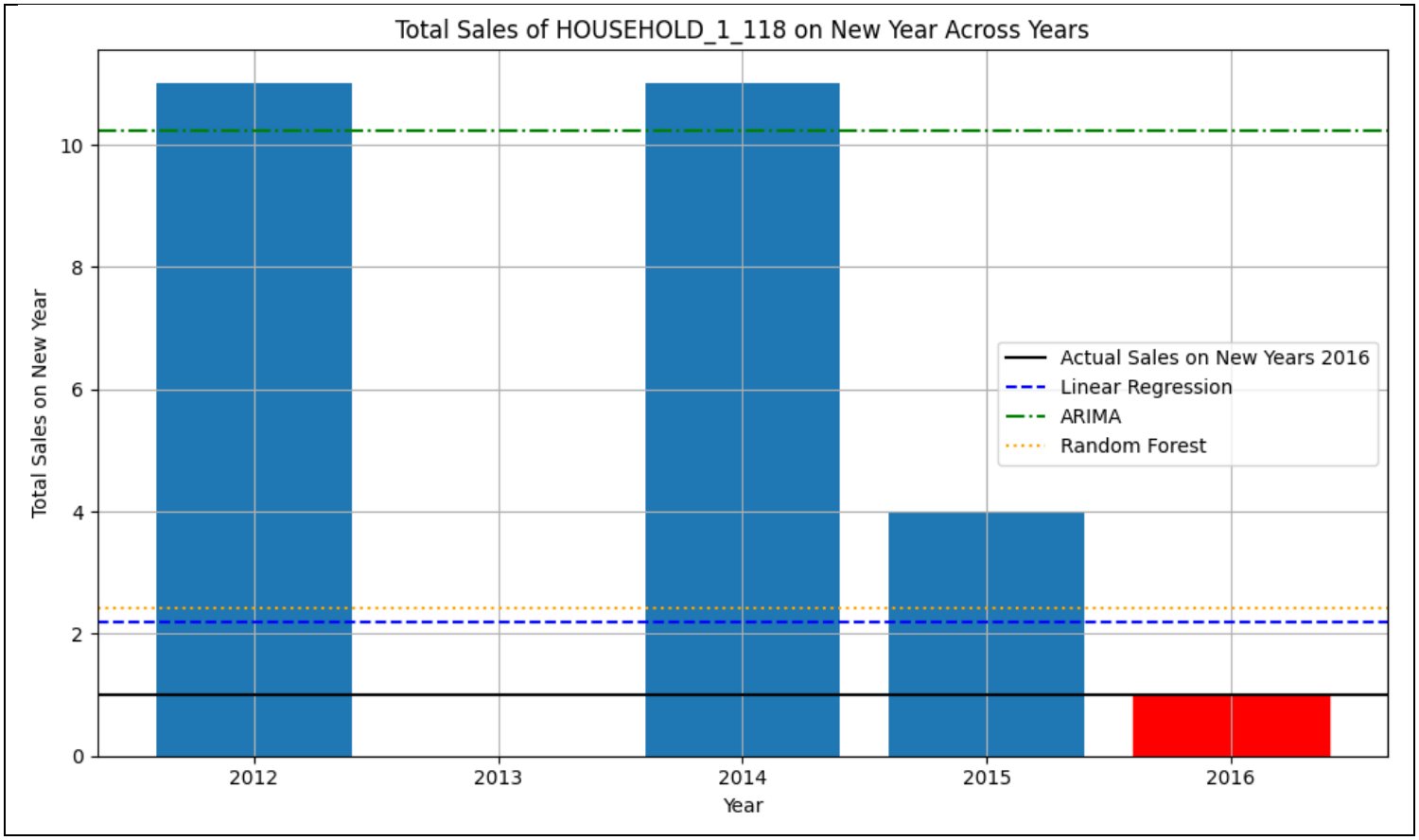


Figure 4.13: The difference between the actual value (black) and the predicted values of all algorithms using a Bar Graph

The sales on NewYear of 2016 was 1 unit. Linear regression and Random Forest predicted 3 units while ARIMA predicted 11 units. From the results in Figure 4.12, Linear Regression is found to be the least inaccurate machine learning algorithm with an accuracy of -20% on demand forecast for events that cause anomalies in the dataset.

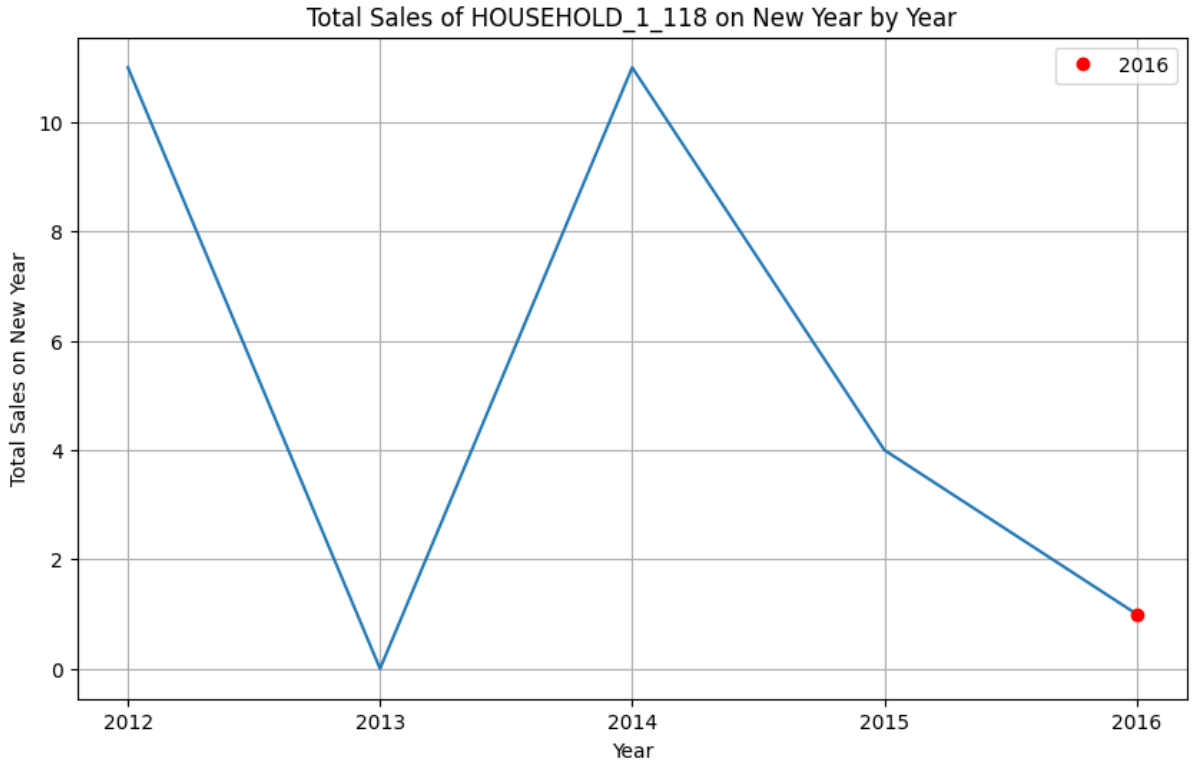


Figure 4.14, Trend of HOUSEHOLD_1_118's aggregated unit sales on new year event across the dataset. The red dot represents sales on New Year 2016 which was used as the test value to assess an algorithm's performance.

The high percentage of inaccuracy and error in performance metrics in Figure 4.12 can be explained by the inconsistent trend observed on New Year in Figure 4.14 throughout the time period. Due to randomness in data and lack of a logical pattern, the machine learning algorithms fail to detect a pattern thus providing suboptimal and inaccurate results.

4.2.4 Second Iteration & Results

Hypothesis 4: Artificial Intelligence can uncover undiscovered relationships and patterns in the dataset and improve forecast accuracy.

These results can be improved by training the algorithms with an event that has observed a stable and consistent pattern of demand throughout the years. Superbowl, one of the most watched sporting events in the U.S. is found to have a profound effect on countrywide sales of product HOUSEHOLD_1_118. As the graph in Figure 4.15 suggests, except for 2012, there is generally an increasing demand for HOUSEHOLD_1_118 on Superbowl each year. That is unlike New Year's demand where the data points deviate from the overall trend causing anomalies and reducing the accuracy of model predictions.

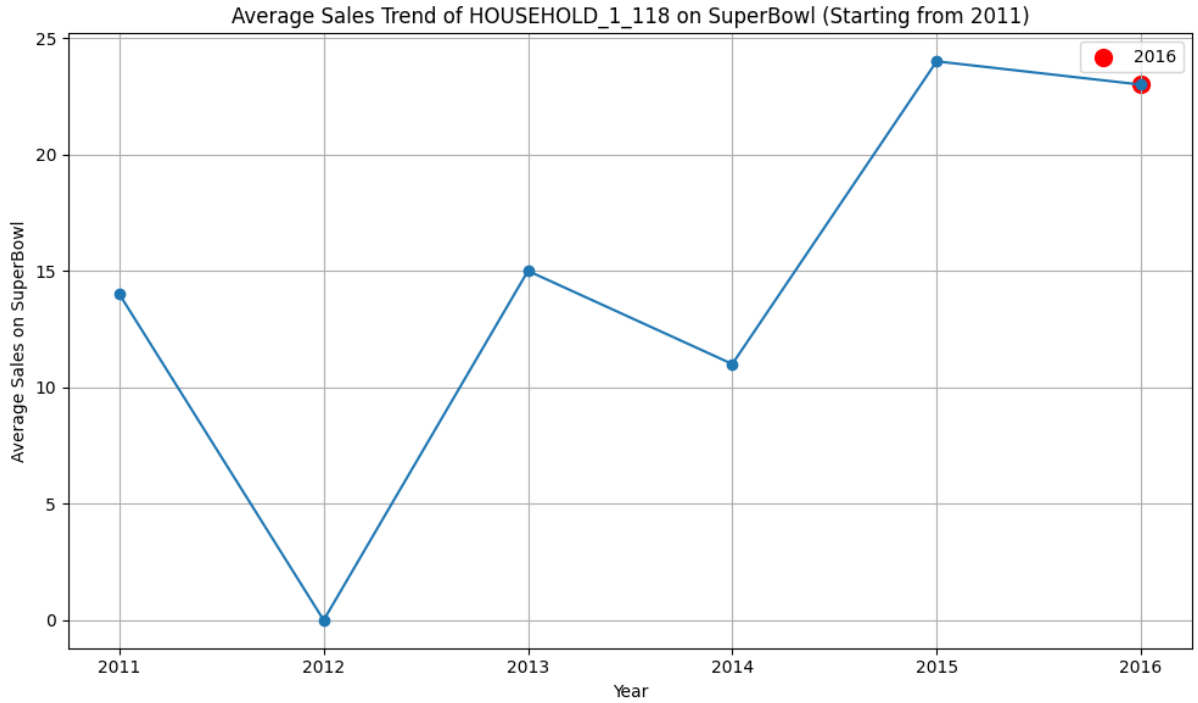


Figure 4.15: Trend of HOUSEHOLD_1_118's aggregated unit sales on SuperBowl event. The red dot represents sales on Super Bowl in 2016 which will be used as the test value to assess an algorithm's performance.

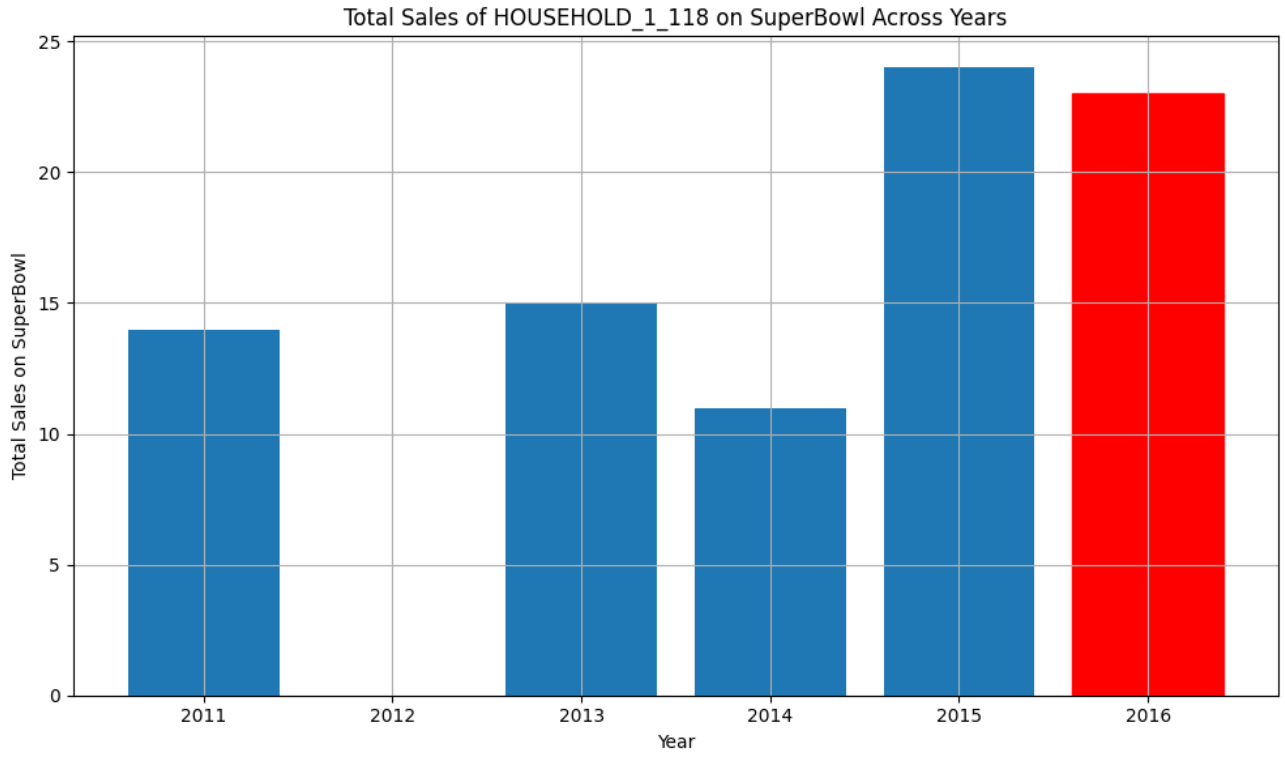


Figure 4.16: Bar Chart showing the sales of HOUSEHOLD_1_118 product on Superbowl test value is represented by red and training values are represented by blue.

Data Modelling

When Linear regression, ARIMA and Random Forests were applied to the subset of Super Bowl events all models except ARIMA yielded satisfactory results. Performance metrics in Figure 4.17 represent the predicted sales and

accuracy of all models compared to actual sales on the Super Bowl in 2016. The bar graph in Figure 4.18 provides a visual representation of the difference between predicted demand and actual demand in 2016. There is a dramatic improvement in the accuracy of all algorithms compared to New Year's sales predictions computed previously. Similar to the first iteration, linear regression is the most accurate algorithm with 96.08% accuracy and a 3.91% prediction error, followed by Random Forest and ARIMA respectively. All algorithms indicate improved results and closeness to the actual value (23 units) due to the presence of a logical trend in the training dataset.

Model Performance Metrics

Algorithm	Actual Value	Predicted Value	Accuracy %	Mean Squared Error %
Linear Regression	23.000	22.100	96.087	3.913
Random Forest	23.000	19.640	85.391	14.609
ARIMA	23.000	12.094	52.583	47.417

Figure 4.17: Python Screenshot of model performance metrics for all machine learning algorithms including the Accuracy percent and Mean Squared Error percent

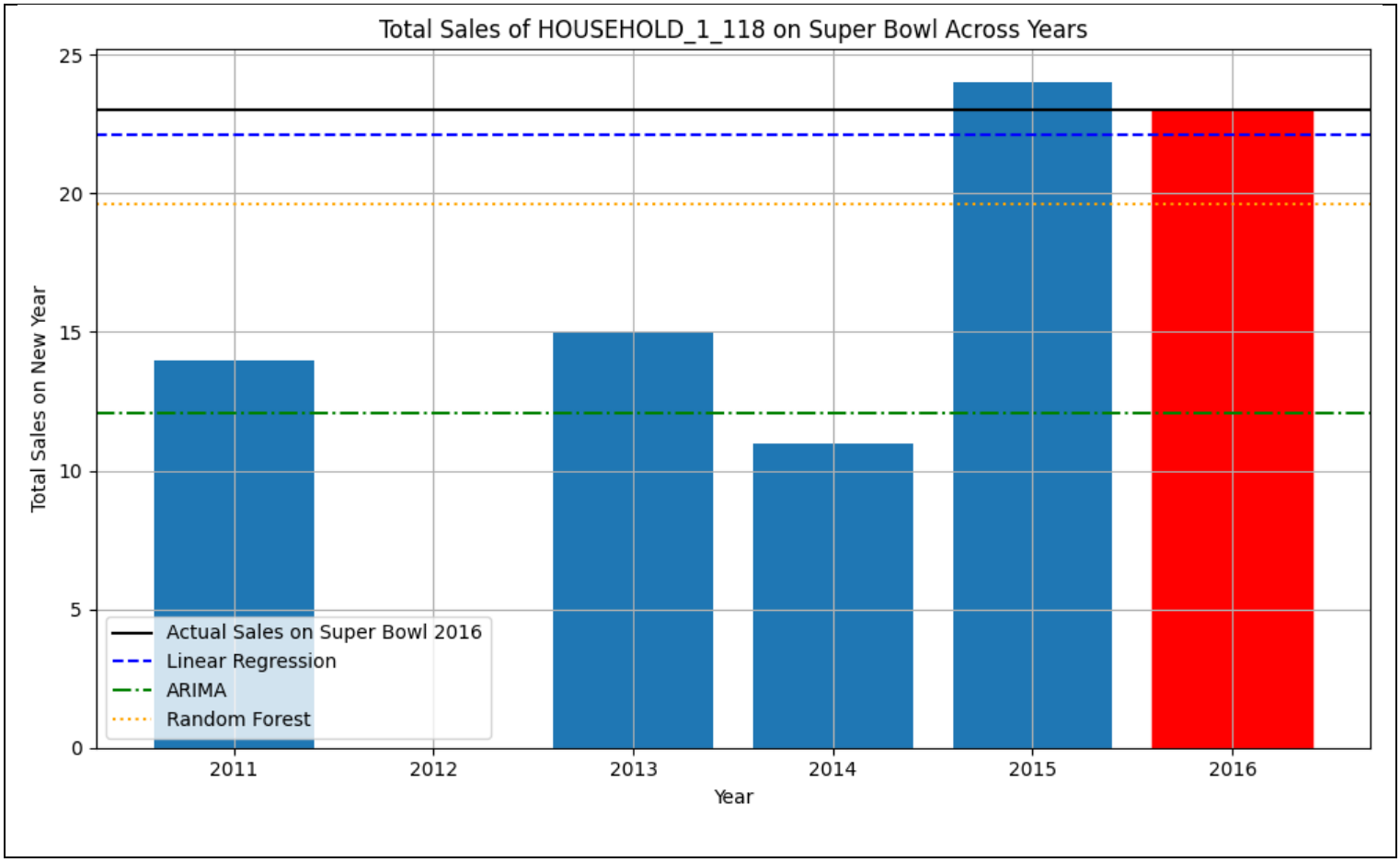


Figure 4.18: The difference between the actual value (black) and the predicted values of all algorithms using a Bar Graph.

4.3 Amazon Review Dataset

Text reviews from crowdsourced websites or online portals have become the biggest source of user sentiment for companies. Unlike customer surveys which are expensive, prone to response bias and become outdated quickly (Chakraborty et al., 2022), online reviews present a cheaper, honest and more flexible option to gather user sentiment. Using this motivation, the Amazon review dataset was selected to predict demand based on text reviews posted by Amazon customers on its online marketplace.

Limitation and Motivation

Hypothesis 2: Integrating customer reviews improves the prediction accuracy of forecasting models.

M5 Forecasting dataset although an excellent time-series data, does not embed any qualitative features or product-specific attributes. The product-specific attributes are necessary to assess the potential of qualitative data in demand forecasting. The analysis of user reviews provides a qualitative perspective to a primary quantitative analysis, offering invaluable insights into customer sentiment that numbers may fail to capture, and therefore achieving a comprehensive mixed-method approach to new product forecasting.

4.3.1 Data Preprocessing

The review dataset incorporates approximately 34 million user reviews covering over 100 gigabytes of memory. This volume of data is not suitable for an average consumer-grade computer to process. Thus, a manageable subset of the product category **Electronics** has been selected for the analysis due to its similarity with the household product in the M5 analysis. There are 7.8 million reviews related to electronics products in the dataset (refer to Appendix B Table B.1).

Feature Selection

The final subset represents 7,824,482 user reviews in total for all electronic items having the following attributes:

Attribute	**reviewerID**	**asin**	**reviewerName**	**helpful**	**reviewText**	**overall**	**reviewTime**
Description	Unique ID of user	Product ID	Name of user	Likes/dislikes on review	Product review	Rating out 5	Time the review was posted

Table 4.2: Horizontal list of all columns in Electronics, their names and correspondent descriptions.

'*Attribute'*, '*reviewerID'*, '*asin'*, '*reiewText'* and '*overall* (rating)' are relevant to the research objective and therefore selected for the data analysis. The review's name would be used if the model had to make clusters and product recommendations based on collaborative filtering. The '*reviewTime*' is not necessary because all historical reviews posted about the selected product are considered to predict demand for a line extension.

The study began with the premise of forecasting demand for tech gadgets, more specifically VR headsets. Since the review dataset belongs to the period when the VR industry was still in its early infancy (Alsop, 2024b) this dataset does not include VR headsets in the electronics category. The closest match that could be found is headset products. One drawback of this dataset is the lack of a category column that restricts the research from directly separating all the reviews based on the product category. To mitigate this problem, text mining and sentiment analysis are used to filter and classify reviews for one specific product.

Filtering

To begin the analysis, all reviews containing the word "*headset*" are selected through a filter command. This term was chosen due to its uniqueness and the limited possibility of it being used in reviews of other products.

```
most_popular_headset = headset_reviews['asin'].value_counts().idxmax()

# Create a separate DataFrame for the rows with the most repeated 'asin' value
most_popular_headset_df = headset_reviews[headset_reviews['asin'] == most_popular_headset].copy()
```

Figure 4.19: Python command creating another subset called '*most_popular_headset_df*' representing the headset product with the highest number of reviews.

79583 reviews contain the word "*headset*" constituting 1% of the whole dataset ".

```
]: headset_reviews.info()

<class 'pandas.core.frame.DataFrame'>
Index: 79583 entries, 2082 to 7824444
Data columns (total 10 columns):
 #   Column          Non-Null Count  Dtype
---  ------          --------------  -----
 0   reviewerID      79583 non-null  object
 1   asin            79583 non-null  object
 2   reviewerName    78941 non-null  object
 3   helpful         79583 non-null  object
 4   reviewText      79583 non-null  object
 5   overall         79583 non-null  float64
 6   summary         79583 non-null  object
 7   unixReviewTime  79583 non-null  int64
 8   reviewTime      79583 non-null  object
 9   year            79583 non-null  int32
dtypes: float64(1), int32(1), int64(1), object(7)
memory usage: 6.4+ MB
```

Figure 4.20: Summary of all features in the *headaset_reviews* table, including column name, data types and total count of non-null entries for each attribute.

Assumptions: It is assumed that the item with the most reviews would automatically be considered the most sold product and the highest in popularity. Moving forward with this assumption, it is sensible for a company to introduce a line extension, a new product that extends from the company's most sold item, as a better and more improved version of the original item.

	reviewerID	asin	reviewerName	helpful	reviewText
6554427	A2041QGJBXDA3N	B009A5204K	NaN	[0, 0]	I see why its best to get this one instead of ...
6554431	AK31M9XC8JRKV	B009A5204K	NaN	[0, 0]	Love the battery life on this thing. I play a...
6554433	ATAAVJ1DEACGP	B009A5204K	NaN	[0, 0]	Overall a good product. Great value for the mo...
6554443	ATCHO19Z36I1K	B009A5204K	02sassy4u	[0, 0]	The sound is clear, great battery life, and t...
6554446	A1FB12P4KSQWDV	B009A5204K	2761	[0, 0]	I really really enjoy the freedom to move abou...
...	...	...	...	...	...
6561454	A2E9PPCEWISTZX	B009A5204K	Zachary R. Sutherland	[0, 0]	*I'm using a rooted Galaxy S4 i337 with this h...
6561455	A14GIJJSMV5LO8	B009A5204K	Zach Countryman	[0, 0]	I got these for use with a galaxy s3, mostly a...
6561459	A2RFHBEQL8WPV6	B009A5204K	Zarf "-- Zarf"	[3, 4]	Pros:The microphone quality on the Tone+ HBS-7...
6561463	AG7FEPY6QFOJB	B009A5204K	ZionsvilleFuji	[3, 4]	I really wish I could have written a good revi...
6561464	A2SSH7DXZ9F29A	B009A5204K	zjl23	[0, 0]	First of all, I wanted a BT headset for workin...

2037 rows × 10 columns

Figure 4.21: Python DataFrame displaying columns and sample rows from the most sold headset product B009A5204K.

The item with the most reviews in the dataset was filtered as **B009A5204K** with a total of 2037 rows making up 2.56% of all headset reviews.

4.3.2 Data Modelling: Sentiment Analysis

A simpler sentiment analysis was performed on the headset reviews, by dividing them into the following categories based on their overall ratings.

Category	**Overall Ratings**	**Total**	**Percentage in overall forecast**
Confirmed buyers	4.0-5.0	1501	50%
Mixed reviews	3.0-3.9	181	30%
Improvements	1-2.9	68 (orig. 355)	20%

Table 4.3: Computed categories from the 'headset_reviews' with each cell representing a derived value such as overall ratings, total reviews and contribution in the overall forecast.

Confirmed Buyers: They are the most valuable customers who have rated the product as equal to or above 4 stars, representing satisfaction and a positive sentiment. From a company's perspective, these users have the potential to become loyal customers to the brand. Hence, 50% of the initial demand for the next-in-line VR product will comprise confirmed buyers.

Mixed reviews: Users who were somewhat satisfied with the product, rating it from 3 to 3.9 are put into the mixed reviews category. Assessing from the business perspective, these customers could be passive buyers or first-time users who are not loyal to the brand but can potentially become repeat consumers of the new product. They will represent 30% of the demand forecast.

Improvements: For users who have rated the product as anywhere ranging from 1 to 2.5, may have some concerns with the quality, aesthetics and ergonomics of the product. The feedback of such users is critical when promising a line extension because it provides important insights into the voice of customers. According to Packowski (2014), the voice of the customer provides an accurate production level in today's world of high uncertainty and increasing variability. Customer feedback facilitates strategic decision-making and competitive business initiatives. Considering their importance, users categorised as "improvements" will constitute about 20% of the demand forecast for the VR headset.

To remove spam reviews, a phrase dictionary has been set up with basic terminology that is commonly found in unsatisfied reviews. Some of the reviews were found to be abstract with explicit mention of the attributes that the users thought could be improved in the future.

dictionary = ["more colours", "different sizes", "new models", "additional", "upgraded", "needs", "need", "would have liked", "poor quality", "complicated", "user-friendly", "buggy", "disappointed"]

(refer to Appendix B, Figure B.2 for complete list)

After carefully selecting the reviews from the above word bank, only 68 out of 355 improvement reviews were taken into consideration for the forecast (See Figure B.3 in Appendix B for Python output).

4.3.3 Results

Confirmed buyers (1501), mixed reviews (181) and improvements (68) are aggregated with their designated percentages shown in Table 4.3, resulting in a total demand forecast of 819 units for the VR headset line extension.

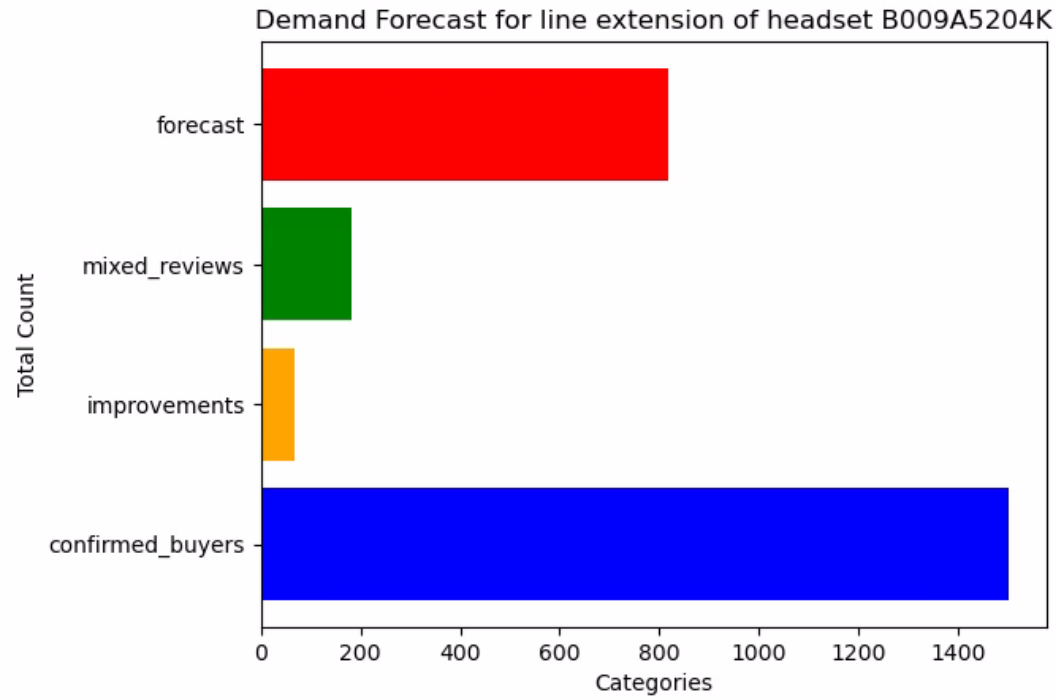


Figure 4.22: Horizontal bar graph illustrating the calculated values of confirmed_buyers, improvements, mixed_reviews and the projected value (forecast) for headset B009A5204K.

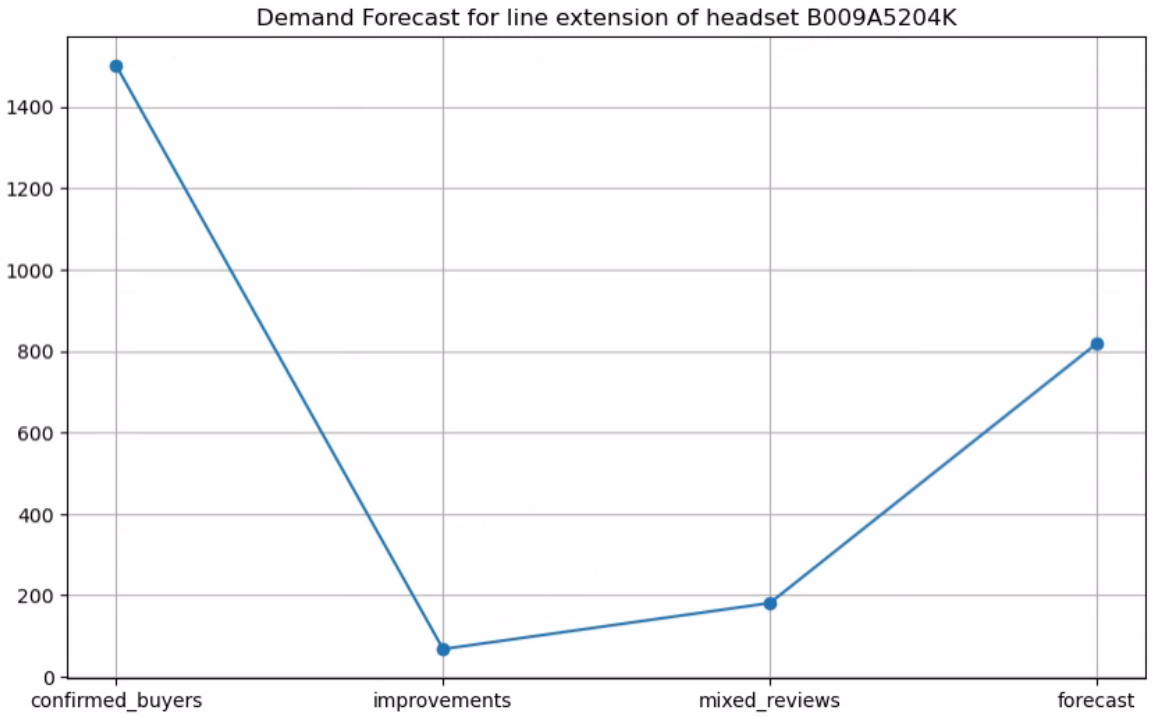


Figure 4.23: Line graph illustrating the calculated values of confirmed_buyers, improvements, mixed_reviews and the projected value (forecast) for headset B009A5204K.

4.4 Market Analysis

Macro environmental changes have a significant impact on demand. Failing to respond quickly may result in missed opportunities and increased threats. Whittington et al. (2020) suggest external factors impact an organisation in two ways:

1. Shaping supply and demand based on
    - Demographics
    - Distribution
    - Geography
    - Culture
2. Innovation
    - Organisational fields like sociograms where businesses interact with competitors to form partnerships to derive innovation.

This research study expands upon critical factors such as market growth, demographics, competitors and target audience deriving accurate demand forecasts. Many of these factors consolidate the macro environment an organisation operates within and can highlight the outliers in an otherwise consistent demand. The external analysis of the VR headset industry was conducted by taking factors such as industry growth projection, demographics information, target audience, competitor analysis and market share into consideration.

Initial Motivation & Assumptions

Hypothesis 3: Insights from market and competitor analysis increase the accuracy, reliability and relevance of demand prediction in new product forecasting.

The study initially aimed to improve new product forecasting by machine learning algorithms focusing on Virtual Reality headsets. Household items in the M5 dataset and electronics in the Amazon review dataset were selected due to their proximity to tech gadgets. Similarly, for external analysis VR headset market is explored to provide a

background to pre-launch strategies of diversification and line extension products. Improved computer-generated images and motion sensors gave rise to the Virtual Reality industry which has begun to gain significant momentum recently.

This analysis is based on the assumption that product data has been acquired by a medium-sized high-tech company called **EnvisionVR** with products ranging from laptops, mobile phones, PCs and headphones. The company is aiming to introduce a new product, a VR headset, a new-to-the-company diversification product to penetrate an unfamiliar market. To target its existing customer pool, the company must work with limited sales data and external market data to derive forecasts.

4.4.1 Industry Growth Projection

The annual sales of VR headsets worldwide reported by Statista are as follows:

Figure 4.24: A Bar Graph depicting VR headset unit sales across the globe from 2019 to 2024, (Alsop, 2024b).

Assumption: Since there is an increasing demand trend observed in Figure 4.25 it can be assumed that the market will observe a similar trend in the future.

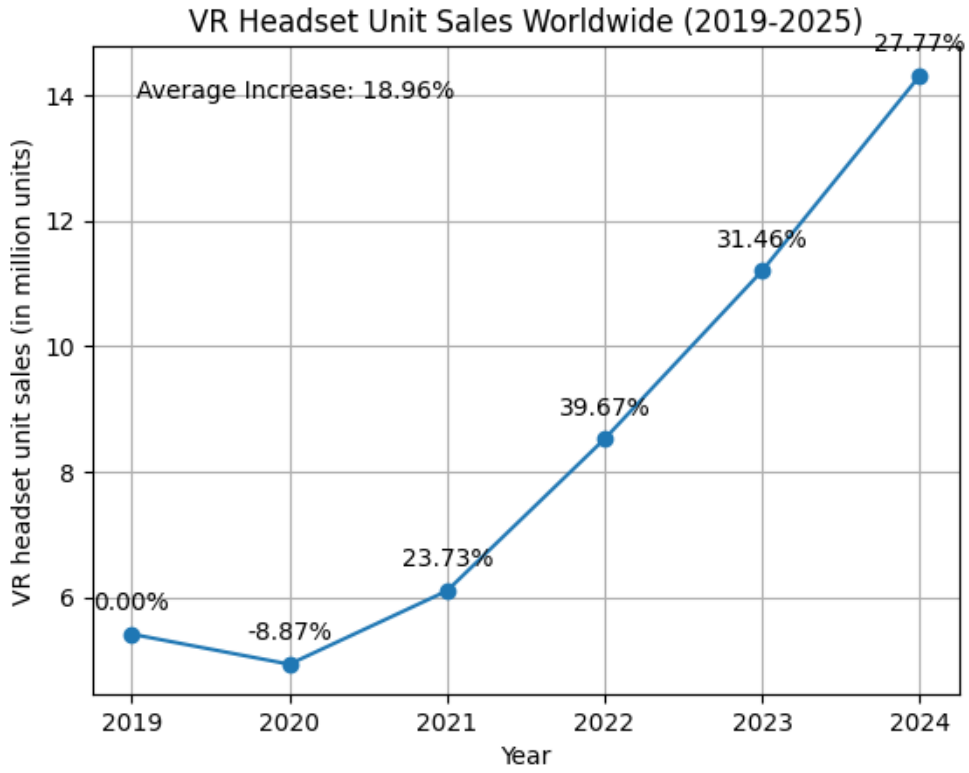


Figure 4.25: Percentage change in global VR headset unit sales from 2019 through 2024

The average percentage increase from 2019 to 2024 is calculated as 18.96%. Assuming the market continues to grow in a similar pattern, the projected worldwide sales for 2025 can be calculated as:

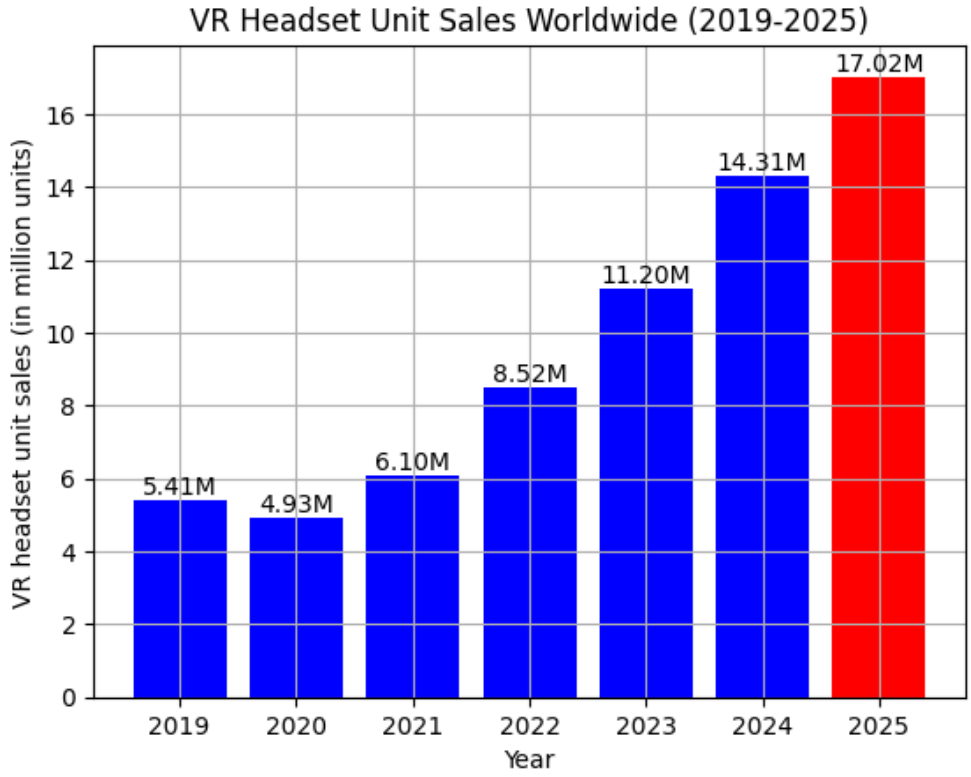


Figure 4.26: A Bar Graph depicting VR headset unit sales across the globe from 2019 to 2024 with forecast in 2025.

Results

If the market continues to grow at a similar rate, the industry projection of VR headsets in the next five years can be estimated as:

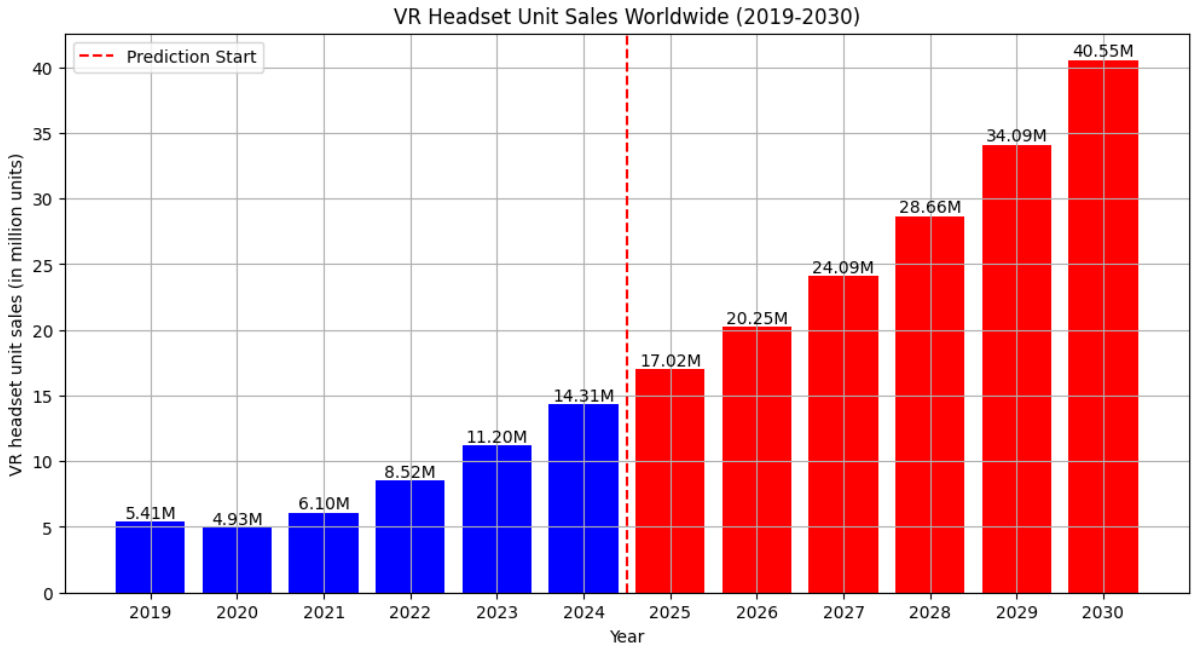


Figure 4.27: Projected Market Growth of VR Industry until 2030

Actual global headset units from 2019 to 2024 are in blue and predicted unit sales from 2015-2030 are in red. A dashed vertical red line indicates the transition from actual data to projections.

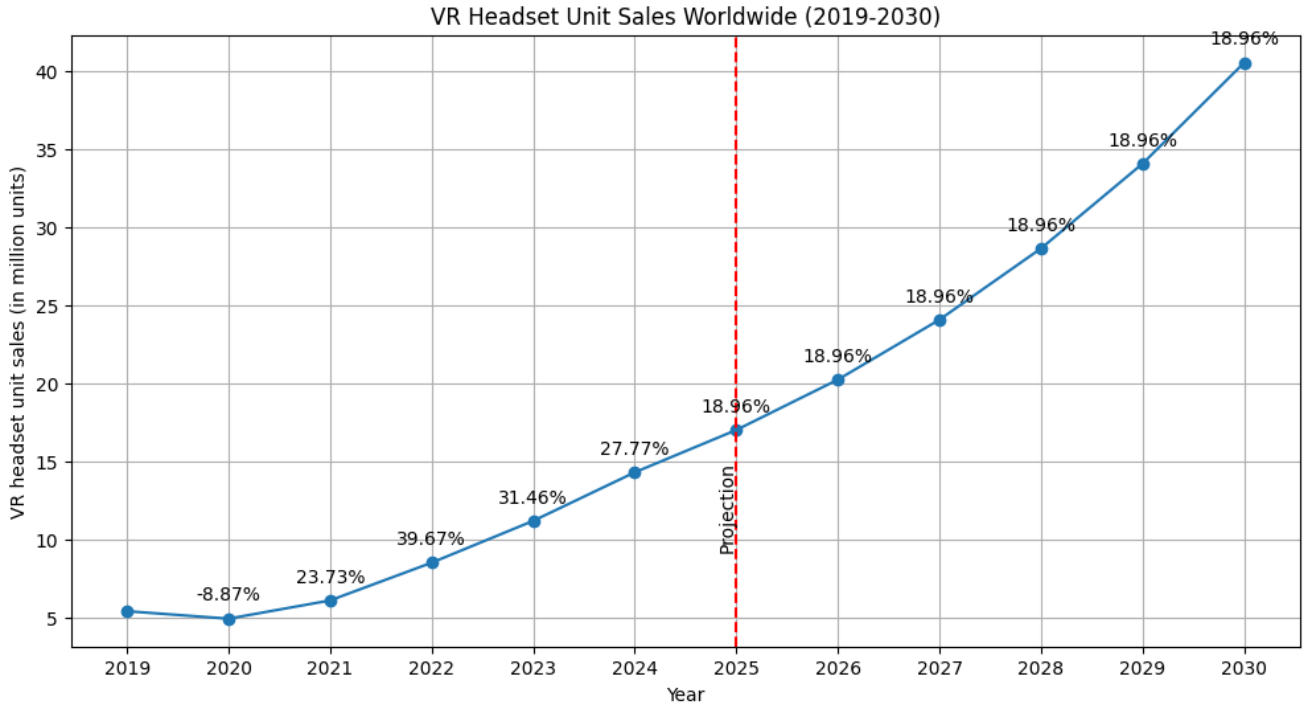


Figure 4.28: Percentage change in global VR headset unit sales from 2019 through 2030

The percentage increase from 2019 to 2024 averaged 18.96%. All years after the projection indicated by the red dashed line are increasing by the same growth rate. This is a ballpark estimation based on the industry growth in previous years. Several unseen events could significantly impact the results in future. For example, the VR headset

market fell by 8.9% in 2020 due to COVID-19. These events are natural events that cannot be predicted in advance but contingency plans can be devised during the pre-crisis stage before they lead to a major crisis (Piotrowski & Guyette, 2010). Although organisations are better equipped today to handle similar supply chain disruptions, the timing of such catastrophic events cannot be predicted. This aspect of the supply chain is beyond the scope of this study. Therefore, the results proceed with the assumption that the VR industry will continue to rise in the next five years with a calculated average of 18.96%.

4.4.2 Target Audience

The demographic information of the users of VR headsets has not yet been publicly available in industry journals. Meta, one of the biggest VR headset market shareholders, interviewed its consumer base about the use of Metaverse, its online platform of virtual and augmented reality (Petrosyan, 2024). The results were as follows:

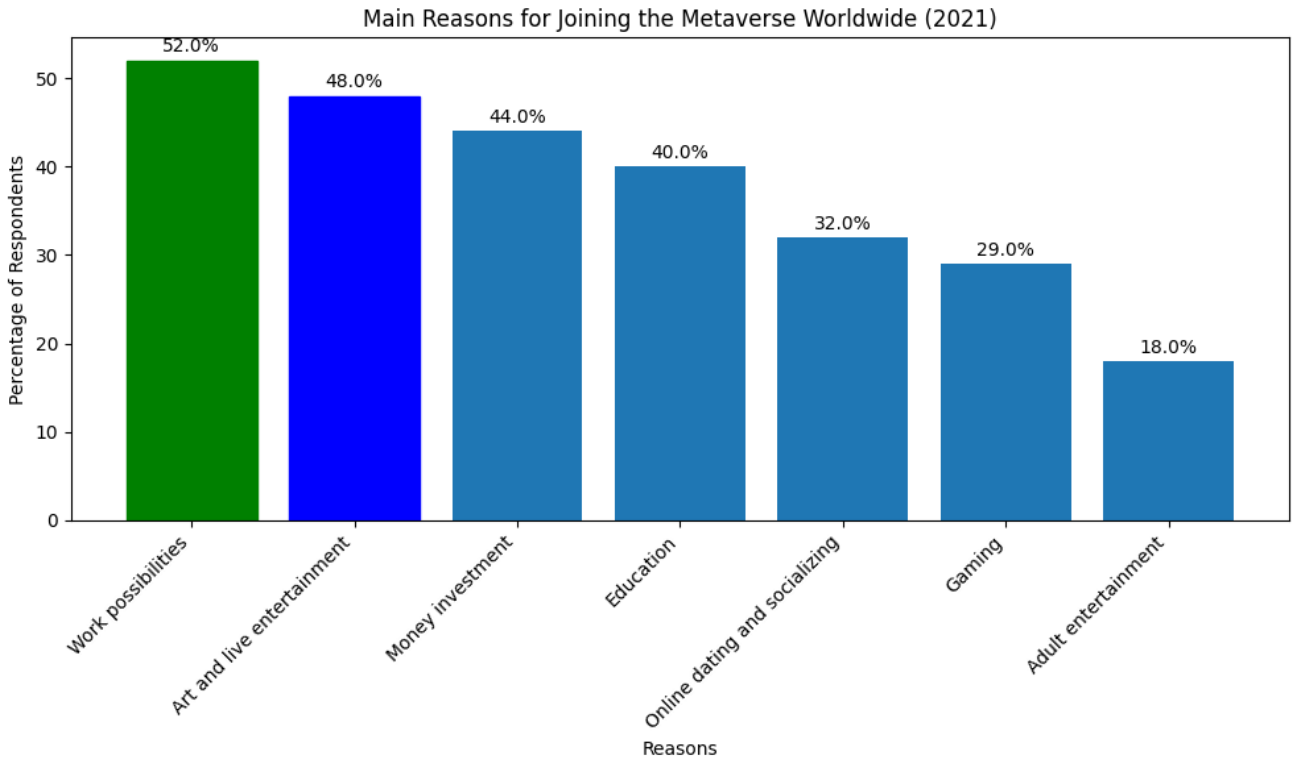


Figure 4.29: Primary reasons for joining Metaverse, a survey conducted by Facebook in 2021 (Petrosyan, 2024).

For a new company that is launching a product in an industry it is unfamiliar with, it is important to know the main customer base and target audience to meet customer demand. Since Meta has been consistently the largest shareholder in the VR industry (Figure 4.31), its consumer statistics can be generalised for the entire consumer base. As highlighted in Figure 4.29, corporate employees, leisure industry, students and investors are the largest consumers of VR headsets. These results are crucial for understanding market dynamics leading to strategic marketing campaigns and accurate demand forecasts.

4.4.3 Competitor Analysis

Market competition has a significant impact on product demand especially for a company launching a new-to-the-company diversification product. Changes in market competition and business environment create the need for more strategic, value-creating and complex supply chain networks (Slack & Lewis, 2020). Although there are pioneering advantages that directly result in “first mover” benefits, giving a competitive advantage to a company, a firm can make its position unique by satisfying the needs of its buyers better than its competitors (Porter, 1985). So, a company called **EnvisionVR** has been established that aims to penetrate the VR headset market with improved products.

The current market share distribution reported by (Alsop, 2024a) is highlighted in Figure 4.30. Meta is the biggest shareholder with about 50% of the market share followed by Apple and ByteDance having almost one-fifth of the total market. In such a concentrated market, the strategic plan for a new company would be to prevent increased inventory and storage costs.

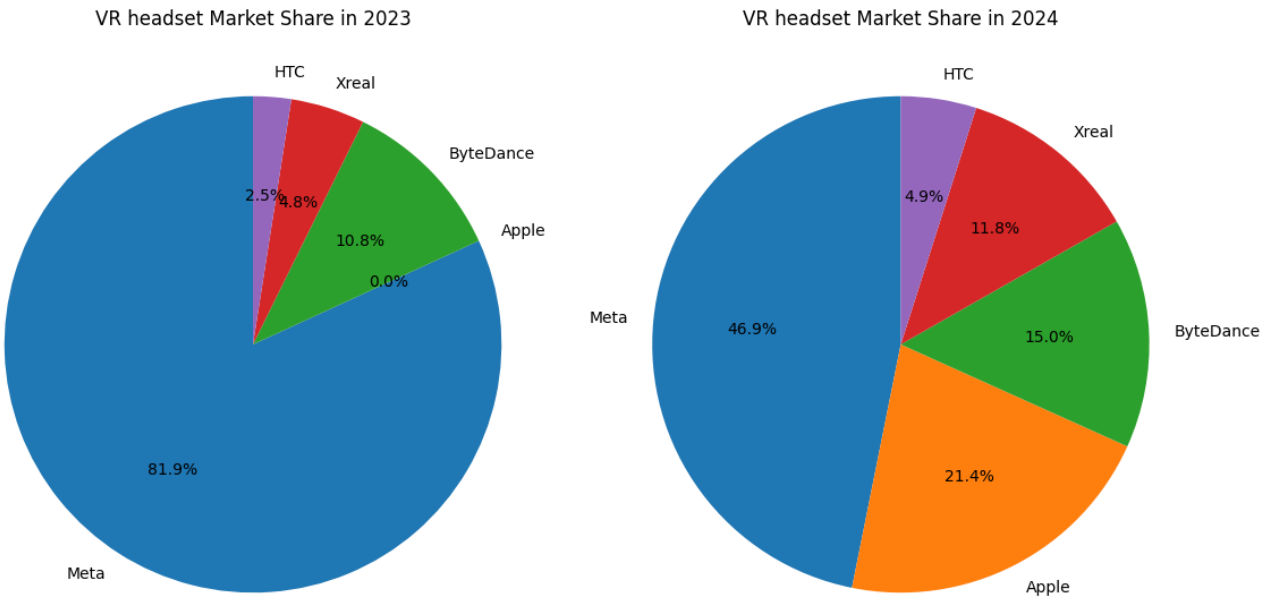


Figure 4.30: Pie Chart illustrating market share percentages of leading VR headset manufacturers in 2023, joined by Apple in 2024, (Alsop, 2024a).

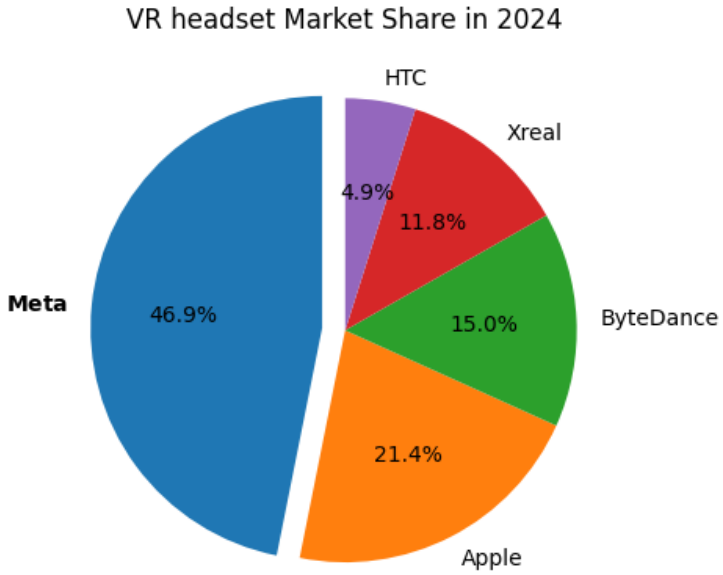


Figure 4.31: Leading VR headset manufacturers in 2024. The industry has been dominated by meta with about 50% of the total market share.

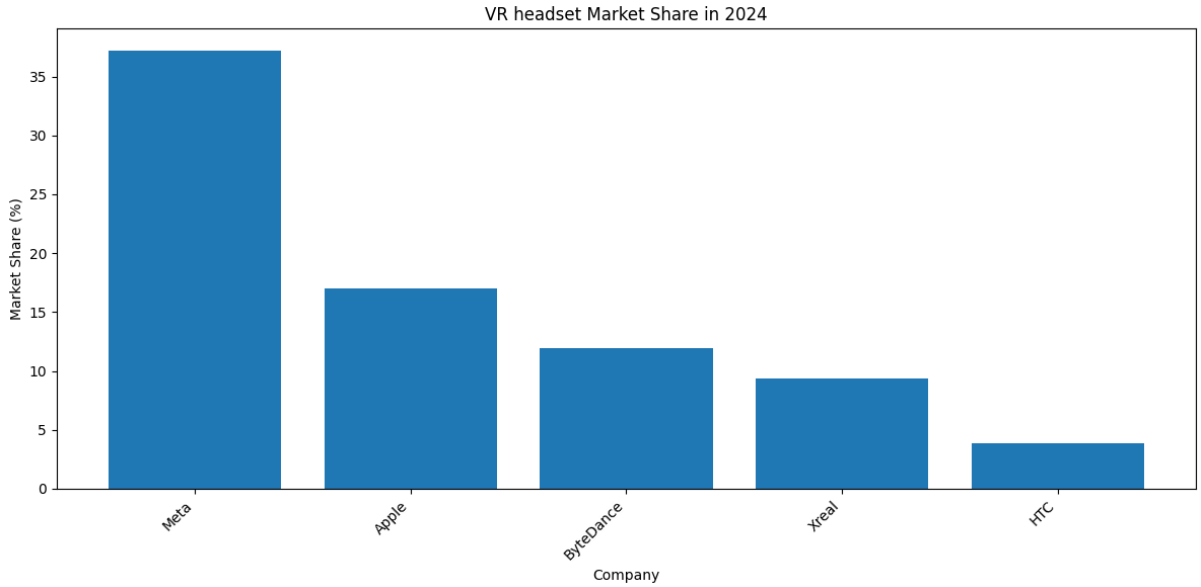


Figure 4.32: A bar graph depicting market share percentages of leading VR manufacturers in 2024.

Assumption: Considering all factors, the total market share for Envision VR is estimated to be around 1% at the launch of its new product. Previously, predicted annual sales of VR headsets in 2025 were estimated to be around 17.02 million (page 46, Figure 4.26). When this value is mapped onto the market share the percentage distribution of all companies is as follows:

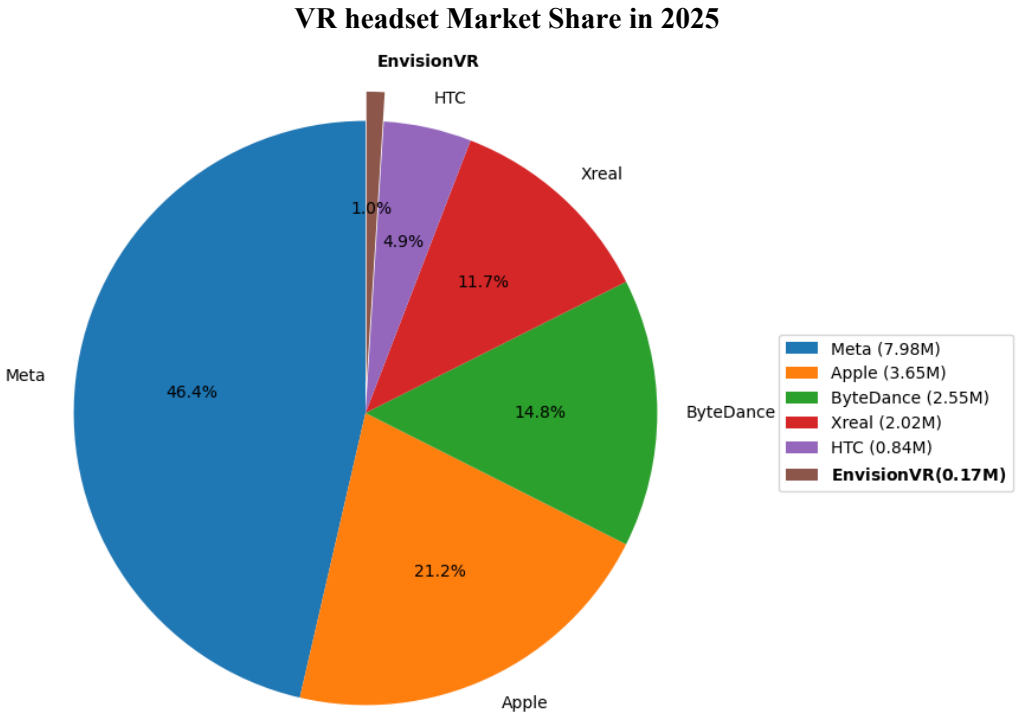


Figure 4.33: Projected VR Headset Market Share in 2025 in percentage with the addition of new company EnvisionVR. The ledger indicates project unit sales in millions.

The total market share of EnvisionVR in the VR headset industry in 2025 is expected to be 1% with an estimated demand of 170,000 units for a new-to-the-market diversification product.

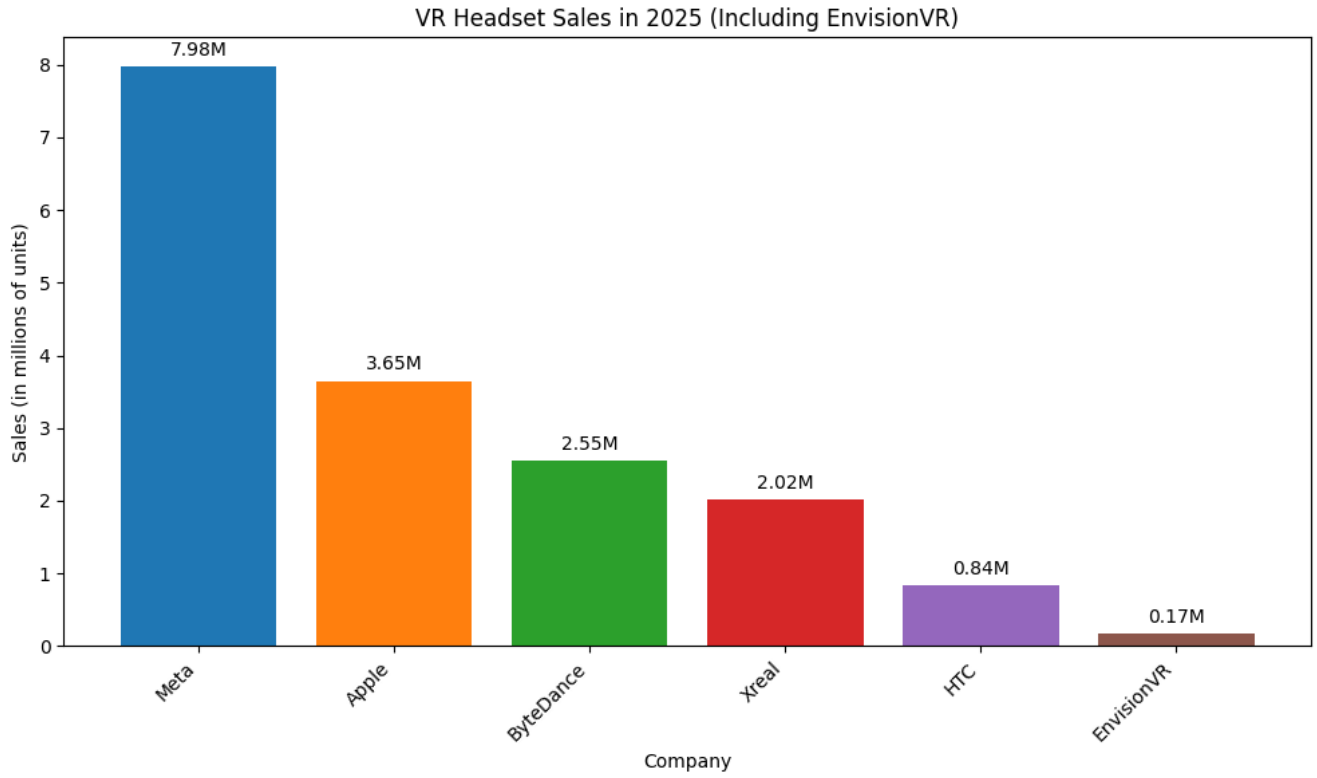


Figure 4.34: Projected VR Headset Unit Sales in 2025 including EnvisionVR

Chapter 5: Discussion & Recommendations

This research demonstrates the impact of machine learning in improving the accuracy of traditional forecasting algorithms and sheds light on the challenges in new product forecasting. A holistic forecasting framework was developed and implemented using secondary data sources. The use of secondary data sources was guided by the time, budget and proprietary disclosure constraints. M5 Forecasting dataset from Walmart, Review dataset from Amazon and Market Analysis from industry journals combined constitute the data collection approach. Besides the challenge of acquiring primary data, secondary data has its own set of limitations. The M5 dataset was divided into multiple hierarchical levels, with states, departments, stores and product categories. The drawback of this dataset was the missing granularity in product features. The segmentation stopped at the product category level, but further analysis was limited beyond that point. Product differentiation, although a suitable method for these types of datasets, could not be implemented due to limited product attributes. Therefore, the analysis of the M5 dataset was restricted to trend analysis.

Product Differentiation

Product Differentiation is a widely used pre-launch strategy used by companies as a determinant of success and failure (Davcik & Sharma, 2015). Originating from marketing, this strategy assesses the confidence factor of a product based on measures like market share, brand loyalty and pricing. In other approaches, it uses differentiation to determine the commonality between products. Afrin et al. (2018) used design specifications to predict demand for automobiles by calculating the product differentiation index (PDI) between old and new vehicles. Kanehara & Kamei (2022) analysed differentiation levels in oligopolies and concluded that decreased levels of product differentiation lead to decreased marginal revenue, profit and transportation prices, ultimately affecting trade costs. This solidifies the importance of product differentiation in new product forecasting.

The Amazon review dataset on the other hand was rich with attribute data and was eligible for product differentiation. The differentiation was performed on an existing product to predict the demand for a line extension. Using sentiment analysis on specific product reviews, unique features, feedback, success and consumer preferences of existing products resulted in context-aware forecasts of an extended product. Data size and complexity were major issues in the analysis of the Amazon review dataset. The total size of the dataset in terms of computer storage is over 100 gigabytes which was computationally intensive to process for the machine. Even when the dataset was reduced to one specific category (*Electronics*) the Python kernel crashed several times due to data-intensive processing tasks of Natural Language Processing and sentiment analysis. For this reason, the computing power was upgraded from 32 to 64 Gigabytes of RAM, but at times even that was not enough for the system to perform complex calculations on millions of records. This emphasises a requirement for the Big Data Analytics model to have appropriate systems and resources in place as it is unfeasible to perform on a typical consumer-grade machine.

Scalability is another factor that must be considered when developing machine learning forecasting models (Hodge et al., 2015). Big data infrastructure is costly and requires components working in parallel at all stages of the data analysis pipeline (Figure 4.1). The storage and processing resources must be established before the loading stage of ETL to ensure the data does not get lost.

Data Quality

Real-world data contains noise and discrepancies and measures should be taken to improve its quality (Kaleem et al., 2023). Data Quality was a significant hurdle in M5 dataset analysis. Although observations were recorded throughout the year, many events with extremely high-low demand levels were left **unnamed**. Data points with missing labels were not included in the analysis and ultimately decreased the overall forecast accuracy that could have been achieved with complete data. To achieve the highest confidence levels the data collection should be consistent, reliable and complete.

Data Aggregation

According to (Hodge et al., 2015; Kaleem et al., 2023; Valenzuela et al., 2023), aggregation is one of the biggest challenges in time series forecasting. Either the disaggregated data is not available for aggregation or the results are only preliminary estimates, producing low-frequency results. As time series is constantly evolving these forecasts

become quickly outdated due to revisions. This issue arose when attempting to integrate the M5 dataset with the Amazon review dataset for improved results, but it was not feasible due to primary differences between the datasets, the attributes and their purpose. Data collection begins with a definite objective and this objective derives the attributes, volume, speed and quality of data (Clark et al., 2019). Since the two datasets were collected for two different organisations with two different purposes; Amazon review for sentiment analysis and M5 forecasting to predict Walmart's sales data, the amalgamation of both datasets to achieve better accuracy was simply not feasible. For future studies that focus on measuring the impact of different types of data, an empirical approach as opposed to secondary data would be more complete, feasible and less time-consuming.

Twitter Data

The New Product Forecasting framework is a complete, robust framework that was developed to derive context-aware, sentiment-focused, pre-launch forecasts using various social media prompts. There are data sources in the framework such as Twitter that could not be incorporated into the solution due to timing and resource constraints. In the empirical research of big data systems, there is growing evidence of social media data improving the precision, accuracy and performance of forecasting models (Srinivas et al., 2021; Welch & Widita, 2019). Zheng et al. (2016) developed an Intelligent Transportation System (ITS) by aggregating social media data to predict road traffic, waiting times and scheduling routes with real-time computing. According to Welch & Widita (2019), despite its predictive benefits, the use of social media in decision-support systems is not common. Provided the advantages of social media prompts, Twitter reviews can be used to determine the sentiment of users on existing products to be aggregated with current insights about a new extension. The tweets can be segmented based on *keywords*, *hashtags* and *geography*. This would provide specific demographic information about product consumption in specific consumer groups. This detailed consumer differentiation was missing in the market analysis of the Virtual Reality headset product. Apart from improving the forecasting accuracy, it would also aid in creating targeted marketing strategies and campaigns for the product launch.

YouTube/WeChat Reviews

Other data sources present in the framework that were not implemented are YouTube video reviews and WeChat, a Chinese e-commerce platform, both sources having user sentiment filtered by product types. YouTube videos can be downloaded and converted into text by using web scrapping, a software model that parses the website data from HTML format to textual extensions such as *txt* or *JSON*. However, there are legal and ethical concerns related to web scrapping, especially in Europe with strict GDPR rules (Jennings & Yates, 2009). Similarly, obtaining data from a Chinese platform does not only involve copyright laws but also requires another translation layer to convert the Mandarin language into English. Translation models add another level of complexity to sentiment analysis that is out of the scope of this study.

Data Updating & Processing

The M5 dataset is a time series data with product sales plotted throughout the calendar years. Time series forecasting is one of the most widely used methods to identify trends and predict future demand. Even so, for longest factor prediction horizons the forecasts become uncertain because the data is constantly being collected and updated (Valenzuela et al., 2023). The objective for analysing the M5 dataset was to predict demand and compare the accuracy of machine learning algorithms. In a broader context with different objectives, this dataset would be continuous data that is collected at regular intervals and updated constantly to keep the accuracy consistent. To rectify this problem, there are two possible ways to update time series data, each with its storage and processing requisites. Online dataset is processed on arrival and does not need to be stored for reprocessing (Akata et al., 2014). These datasets are better for optimisation as they take minimal computing power. Batch processing, on the other hand, unlearns and recalculates learning each time it is updated. For example, astronomical data is collected and updated in real-time and is reprocessed constantly (Valenzuela et al., 2023). The choice of updating mechanism depends on the data analysis objectives.

Random Forests

Three widely used machine learning algorithms namely Linear Regression, ARIMA and Random Forests were applied to the dataset to calculate the accuracy of demand forecasts. With inconsistent sales on New Year in Figure

4.14, Random Forest proved to be relatively accurate in predicting demand and making associations in otherwise noisy and inconsistent data. Y. Wang et al. (2018b) argued that although RF has been widely researched, its practicability is mostly theoretical and usually performs poorly in practice. The authors (ibid) suggested the Bernoulli Random Forest (BRF), an improved and simplified version of random forests, for better performance and empirical accuracy. Cheng et al. (2020) analysed the association between walking time and environment attributes (land use, distance, bikes, crowd, population density) for adults in Nanjing (China) and noted that Random Forest with transformed independent variables outperforms linear regression. Based on the analysis of various authors, the applications and effectiveness of random forests remain debatable and should be considered a topic of future research in new product forecasting.

Linear Regression

Linear Regression showed extraordinary results with **96.08%** accuracy of demand forecast on Super Bowl (Figure 4.17). It was also the model having least inaccuracy when performed on inconsistent data (Figure 4.12). Albeit the satisfactory results and wide application of linear regression, it is only applicable to data with variables having linear relationships. According to (Ari & Güvenir, 2002; Tharenou et al., 2007; Thomas, 2004), real-world datasets have multiple variable relationships and are much more complex and sophisticated. Linear regression also suffers from missing features loss of information, irrelevant features, noisy data and partitioning and normalisation requirements.

Ari & Güvenir (2002) suggest Cluster Linear Regression (CLR) as an alternative to linear regression to improve the accuracy of predictions. It is an extension of linear regression, that improves the accuracy by dividing training datasets into subspace partitions called clusters. Research suggests that CLR outperforms linear regression and improves model learning as long as there is a linear approximation for the determined function in each subspace (Ari & Güvenir, 2002). It is an iterative algorithm where partitioning continues until the approximation fits all data instances and each new subspace has a better-fitting linear approximation than the previous one, thus improving forecast accuracy. Similar to ARIMA, CLR requires a substantial number of data points to make clusters, which was not possible with only six observations for yearly recurring events in the M5 dataset.

ARIMA

ARIMA performed poorly in predicting product demand in both iterations of M5 data analysis. There could be two potential reasons for this performance. ARIMA assumes there is a stable time series with constant variance over time. This was not true for New Year events which showed an irregular trend and highly volatile demand (see Figure 4.14 on page 37). The results were slightly better in the second iteration with 52% accuracy but compared to other models these were sub-optimal figures. ARIMA model requires a reasonable amount of sales data (fifty data points) to estimate the parameters. The Superbowl historical data with only five data points was not sufficient enough for ARIMA to identify patterns and provide an accurate forecast. To get accurate results in future, the data collection strategy should be revised and data size should be devised and validated during the selection stage of data collection.

Application

This research study started with the premise of the consumer electronics industry mainly focusing on predicting demand for technology gadgets. This is a fast-moving industry with a shorter product lifecycle and steep decline compared to other product categories (Hu et al., 2019). Due to the lack of readily available primary electronic sales data, the research was guided by secondary data in the electronics category having limited granularity. If implemented on first-hand, highly detailed, fine-grained and segmented product data, more favourable results can be achieved.

The New Product Forecasting framework has the potential to be adopted by other industries due to its flexible and scalable nature. The retail industry relies heavily on fast-moving trends and user reviews. Incorporating sentiment from users, analysing the latest trends and making product differentiation based on specific attributes like colour, size and style will enhance product launch strategies and supply chain effectiveness. Similarly, e-commerce websites usually contain an abundance of user reviews. Obtaining online sentiment has become cheaper because of crowdsourcing platforms, providing e-commerce businesses with user feedback that is rich with sentiment. There is

a common theme of combining expert knowledge with historical sales data in the existing literature in new product forecasting (Fye et al., 2013; Kahn, 2006; Lee et al., 2014; Yamamura et al., 2022). The use of impersonal sources like online reviews has not been sufficiently investigated in previous studies (Welch & Widita, 2019). This research recommends adding a customer-centric layer of user reviews to increase the efficiency of demand forecasts.

The datasets when analysed individually offer limited benefit when predicting demand for novel products but when merged extract comprehensive, data-driven and sentiment-based insights. The growing use of big data and artificial intelligence has been greatly documented in recent forecast research studies (Aggarwal, 2016; Choi et al., 2022; Gao et al., 2019; Hodge et al., 2015; Kaleem et al., 2023; Kuo & Kusiak, 2019; Laranjeiro et al., 2019; Shen & Chan, 2017). This study builds upon the existing evidence and contributes to further enhancing the forecasting framework by combining personal and impersonal data sources. The premise of machine learning algorithms is validated by comparing the model accuracy and recommending the most suitable machine learning techniques based on data type and performance metrics.

Chapter 6: Conclusion

This dissertation investigated the role of machine learning algorithms in improving new product forecasting. By identifying the challenges currently faced by organisations in the forecasting realm and devising mitigation strategies, a motivation was developed to introduce artificial intelligence and machine learning in the traditional statistical forecasting frameworks. Unlike current literature in pre-launch forecasting, this study proposes using diverse data sources to be used in conjunction to improve forecasting practices. Quantitative data such as historical sales, demographics, social media clicks and keywords provide numerical metrics that are transformed into specific figures to produce a measure of interest. Qualitative data acquired from crowdsource websites, YouTube videos and Twitter presents a valuable opportunity to innumerate customer interest and embed it into projections.

The availability of personal and impersonal data sources establishes the concept of Big Data with large and complex datasets. Data is being collected at unprecedented speed and in enormous volume. Big Data Analytics provides solutions to businesses that are finding ways to gather valuable insights from this data and guide their decision-making. One of the most important decisions for any product-based business is to maintain its inventory levels according to forecasted demand. Traditional forecasting models were not designed to process today's high volume and high-speed data. Building upon this necessity, the study investigates the role of machine learning in demand forecasting and subsequently demonstrates its benefit in improving forecasting accuracy.

Algorithms such as Linear Regression, ARIMA and Random Forest were applied to Walmart's time-series data, combining traditional statistics with machine learning techniques. The results indicated that these algorithms forecasted accurate demand on unseen events and showcased their ability to handle large volumes of data. Sentiment analysis was applied to text reviews posted online by users on Amazon. Natural Language Processing combined with results of sentiment analysis provided three distinct review categories namely positive, negative and neutral reviews. The categories constitute a demand forecast and further solidified the importance of customer preferences in forecasting decisions. Market analysis was mainly focused on identifying market growth projections, target demographics and competitors to develop accurate forecasts of new products.

Future Research

The research focused on improving the demand forecast for diversification and line extensions. Originally conceived with the idea of an electronics company that is introducing a tech gadget and wants to systematically derive the demand for their new product. The datasets that are used to achieve research objectives are unique in their motivation, objectives and content. Although there was a significant amount of data disparity, the analysis was generalised to the initial idea of electronics and technology products. The benefit of this research study is its expandability to other industries. Consumer electronics, retail and e-commerce operate in a similar Business-to-Consumer (B2C) model. Whether it's a retail store or online marketplace, customers are buying products and companies have to ensure the products are available at the time of purchase. Apart from the business models, all firms generally obtain their data from similar data sources. Sales history, sensors, IoT devices, social media analytics, customer preferences and buying behaviour are common data sources among businesses. Despite variations in industry type and business models, the fundamental sources of data remain the same. This research study presents an opportunity to explore the proposition of big data in several business paradigms.

The findings underscore the importance of impersonal data sources for new product forecasting when there is limited historical data available. The study demonstrates key factors that impact demand and highlights the role of algorithms in enhancing forecast accuracy. While it provides significant insights, the findings are limited by the unavailability of primary data, complex aggregation, data quality, size and complexity of secondary data, limited computational resources, time and proprietary concerns. Despite all limitations and challenges, the generalised new product forecasting framework, collected datasets and implementation were able to synthesise information and deliver actionable results, successfully meeting all research objectives.

Known to be the most challenging and risky, diversification remains a challenge to be solved and the research is expected to expand further in the near future. The current literature on new products and machine learning proposes endless opportunities for improving forecasting models. Now that the world is leaning towards Industry 5.0, the availability of data is not a problem anymore. The focus of future research should be how this data can be

effectively used to improve business decisions. In summary, this research study has advanced an understanding of machine learning algorithms in the application of new product forecasting. Obtaining data from multiple sources and integrating traditional statistical forecasting models with machine learning techniques can significantly improve forecast accuracy. This research not only pinpoints the applications of machine learning models in supply chain forecasting but also lays the groundwork for future exploration of various data sources within the domain of Big Data.

References

Afrin, K., Nepal, B., & Monplaisir, L. (2018). A data-driven framework to new product demand prediction: Integrating product differentiation and transfer learning approach. *Expert Systems with Applications*, *108*, 246–257. https://doi.org/10.1016/j.eswa.2018.04.032

Aggarwal, C. C. (2016). *Recommender systems: the textbook*. Springer. https://doi.org/10.1007/978-3-319-29659-3

Akata, Z., Perronnin, F., Harchaoui, Z., & Schmid, C. (2014). Good Practice in Large-Scale Learning for Image Classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *36*(3), 507–520. https://doi.org/10.1109/TPAMI.2013.146

Alsop, T. (2024a). *AR/VR headset companies shipment share 2023*. https://www.statista.com/statistics/1407105/ar-vr-headset-companies-shipment-share/

Alsop, T. (2024b). *VR headset unit sales worldwide 2024*. https://www.statista.com/statistics/677096/vr-headsets-worldwide/

Amazon. (2019). *Real-time personalization and recommendation | Amazon Personalize | AWS*. aws.amazon.com. https://aws.amazon.com/personalize/

Amazon. (2023). *Introduction - Amazon Reviews'23*. https://amazon-reviews-2023.github.io/main.html

Ari, B., & Güvenir, H. A. (2002). Clustered linear regression. *Knowledge-Based Systems*, *15*(3), 169–175. https://doi.org/https://doi.org/10.1016/S0950-7051(01)00154-X

Bilger, M., & Manning, W. G. (2015). MEASURING OVERFITTING IN NONLINEAR MODELS: A NEW METHOD AND AN APPLICATION TO HEALTH EXPENDITURES. *Health Economics*, *24*(1), 75–85. https://doi.org/10.1002/hec.3003

Box, G. E. P., Jenkins, G. M., & Reinsel, G. C. (2008). *Time series analysis: forecasting and control* (Fourth). J. Wiley & Sons.

Buchanan, P. D., & Bryman, P. A. (2009). *The Sage Handbook of Organizational Research Methods*. SAGE Publications Ltd.

Ćalasan, M., Abdel Aleem, S. H. E., & Zobaa, A. F. (2020). On the root mean square error (RMSE) calculation for parameter estimation of photovoltaic models: A novel exact analytical solution based on Lambert W function. *Energy Conversion and Management*, *210*, 112716. https://doi.org/10.1016/j.enconman.2020.112716

Chakraborty, I., Kim, M., & Sudhir, K. (2022). Attribute Sentiment Scoring with Online Text Reviews: Accounting for Language Structure and Missing Attributes. *Journal of Marketing Research*, *59*(3), 600–622. https://doi.org/10.1177/00222437211052500

Chen, S. H., & Pollino, C. A. (2012). Good practice in Bayesian network modelling. *Environmental Modelling & Software : With Environment Data News*, *37*, 134–145. https://doi.org/10.1016/j.envsoft.2012.03.012

Cheng, L., De Vos, J., Zhao, P., Yang, M., & Witlox, F. (2020). Examining non-linear built environment effects on elderly's walking: A random forest approach. *Transportation Research. Part D, Transport and Environment*, *88*, 102552. https://doi.org/10.1016/j.trd.2020.102552

Ching-Chin, C., Ka Ieng, A. I., Ling-Ling, W., & Ling-Chieh, K. (2010). Designing a decision-support system for new product sales forecasting. *Expert Systems with Applications*, *37*(2), 1654–1665. https://doi.org/10.1016/j.eswa.2009.06.087

Choi, T., Kumar, S., Yue, X., & Chan, H. (2022). Disruptive Technologies and Operations Management in the Industry 4.0 Era and Beyond. *Production and Operations Management*, *31*(1), 9–31. https://doi.org/10.1111/poms.13622

Choy, M. (2021). *Council Post: What Netflix's Recommendation Systems Can Teach Us About The Computing Challenges Of The Near Future*. https://www.forbes.com/sites/forbestechcouncil/2021/02/19/what-netflixs-recommendation-systems-can-teach-us-about-the-computing-challenges-of-the-near-future/.

Clark, T., Foster, L., & Bryman, A. (2019). *How to do your social research project or dissertation* (First). Oxford University Press. https://doi.org/10.1093/hepl/9780198811060.001.0001

Dang, S., Peng, L., Zhao, J., Li, J., & Kong, Z. (2022). A Quantile Regression Random Forest-Based Short-Term Load Probabilistic Forecasting Method. *Energies (Basel)*, *15*(2), 663. https://doi.org/10.3390/en15020663

Davcik, N. S., & Sharma, P. (2015). Impact of product differentiation, marketing investments and brand equity on pricing strategies: A brand level investigation. *European Journal of Marketing*, *49*(5/6), 760–781. https://doi.org/10.1108/EJM-03-2014-0150

Eroglu, C., Hofer, C., Hofer, A. R., & Hou, Y. (2023). "Cultural inventories": How dimensions of national culture moderate the effect of demand unpredictability on firm-level inventories. *International Journal of Production Economics*, *264*, 108984. https://doi.org/10.1016/j.ijpe.2023.108984

Feiler, D., & Tong, J. (2022). From Noise to Bias: Overconfidence in New Product Forecasting. *Management Science*, *68*(6), 4685–4702. https://doi.org/10.1287/mnsc.2021.4102

Fye, S. R., Charbonneau, S. M., Hay, J. W., & Mullins, C. A. (2013). An examination of factors affecting accuracy in technology forecasts. *Technological Forecasting & Social Change*, *80*(6), 1222–1231. https://doi.org/10.1016/j.techfore.2012.10.026

Galic, K., Curic, D., & Gabric, D. (2009). Shelf Life of Packaged Bakery Goods--A Review. *Critical Reviews in Food Science and Nutrition*, *49*(5), 405–426. https://doi.org/10.1080/10408390802067878

Gandomi, A., & Haider, M. (2015). Beyond the hype: Big data concepts, methods, and analytics. *International Journal of Information Management*, *35*(2), 137–144. https://doi.org/10.1016/j.ijinfomgt.2014.10.007

Gao, G., Wang, Z., Liu, X., Li, Q., Wang, W., & Zhang, J. (2019). Travel Behavior Analysis Using 2016 Qingdao's Household Traffic Surveys and Baidu Electric Map API Data. *Journal of Advanced Transportation*, *2019*, 1–18. https://doi.org/10.1155/2019/6383097

Gartner. (2023). *Three Factors Weighing on Growth Rates in 2023*. https://www.gartner.com/en/insights

Gartner, W. B., & Thomas, R. J. (1993). Factors affecting new product forecasting accuracy in new firms. *The Journal of Product Innovation Management*, *10*(1), 35–52. https://doi.org/10.1016/0737-6782(93)90052-R

GDPR. (2013). *General Data Protection Regulation (GDPR)* . https://gdpr-info.eu

Ge, M., Bangui, H., & Buhnova, B. (2018). Big Data for Internet of Things: A Survey. *Future Generation Computer Systems*, *87*, 601–614. https://doi.org/10.1016/j.future.2018.04.053

Goetze, F. (2011). Understanding Chinese Consumer Behaviour Towards New Products. *Marketing*, *33*(2), 147–158. https://doi.org/10.15358/0344-1369-2011-2-147

harvard business review. (2016). *About Us*. https://hbr.org/corporate/about

Hassani, H., Beneki, C., Unger, S., Mazinani, M. T., & Yeganegi, M. R. (2020). Text Mining in Big Data Analytics. *Big Data and Cognitive Computing*, *4*(1), 1. https://doi.org/10.3390/bdcc4010001

Hodge, V. J., O'Keefe, S., Weeks, M., & Moulds, A. (2015). Wireless Sensor Networks for Condition Monitoring in the Railway Industry: A Survey. *IEEE Transactions on Intelligent Transportation Systems*, *16*(3), 1088–1106. https://doi.org/10.1109/TITS.2014.2366512

Hu, K., Acimovic, J., Erize, F., Thomas, D. J., & Van Mieghem, J. A. (2019). Forecasting New Product Life Cycle Curves: Practical Approach and Empirical Analysis. *Manufacturing & Service Operations Management*, *21*(1), 66. https://doi.org/10.1287/msom.2017.0691

Huang, Y., Wang, R., Huang, B., Wei, B., Zheng, S. L., & Chen, M. (2021). Sentiment Classification of Crowdsourcing Participants' Reviews Text Based on LDA Topic Model. *IEEE Access*, 1. https://doi.org/10.1109/ACCESS.2021.3101565

IBM. (2023). *What is Data Mining? | IBM*. https://www.ibm.com/topics/data-mining

Jennings, F., & Yates, J. (2009). Scrapping over data: are the data scrapers' days numbered? *Journal of Intellectual Property Law & Practice*, *4*(2), 120–129. https://doi.org/10.1093/jiplp/jpn232

Jo, T. (2021). *Machine learning foundations: supervised, unsupervised, and advanced learning*. Springer. https://doi.org/10.1007/978-3-030-65900-4

Jung, A. (2022). *Machine learning: the basics*. Springer. https://doi.org/10.1007/978-981-16-8193-6

Kahn, K. B. (2002). An exploratory Investigation of new product forecasting practices. *The Journal of Product Innovation Management*, *19*(2), 133–143. https://doi.org/10.1111/1540-5885.1920133

Kahn, K. B. (2006). *New product forecasting: an applied approach*. M.E. Sharpe, Inc. https://doi.org/10.4324/9781315702117

Kaleem, S., Sohail, A., Tariq, M. U., & Asim, M. (2023). An Improved Big Data Analytics Architecture Using Federated Learning for IoT-Enabled Urban Intelligent Transportation Systems. *Sustainability (Basel, Switzerland)*, *15*(21), 15333. https://doi.org/10.3390/su152115333

Kanehara, D., & Kamei, K. (2022). Transportation price, product differentiation, and R&D in an oligopoly. *PloS One*, *17*(9), e0273904–e0273904. https://doi.org/10.1371/journal.pone.0273904

Kato, A. (2012). Productivity, returns to scale and product differentiation in the retail trade industry: an empirical analysis using Japanese firm-level data. *Journal of Productivity Analysis*, *38*(3), 345–353. https://doi.org/10.1007/s11123-011-0251-1

Kim, S., & Shin, D. H. (2016). Forecasting short-term air passenger demand using big data from search engine queries. *Automation in Construction*, *70*, 98–108. https://doi.org/10.1016/j.autcon.2016.06.009

Kolluri, J., Kotte, V. K., Phridviraj, M. S. B., & Razia, S. (2020). Reducing Overfitting Problem in Machine Learning Using Novel L1/4 Regularization Method. *2020 4th International Conference on Trends in Electronics and Informatics (ICOEI)(48184)*, 934–938. https://doi.org/10.1109/ICOEI48184.2020.9142992

Kuo, Y.-H., & Kusiak, A. (2019). From data to big data in production research: the past and future trends. *International Journal of Production Research*, *57*(15–16), 4828–4853. https://doi.org/10.1080/00207543.2018.1443230

Laranjeiro, P. F., Merchán, D., Godoy, L. A., Giannotti, M., Yoshizaki, H. T. Y., Winkenbach, M., & Cunha, C. B. (2019). Using GPS data to explore speed patterns and temporal fluctuations in urban logistics: The case of São Paulo, Brazil. *Journal of Transport Geography*, *76*, 114–129. https://doi.org/10.1016/j.jtrangeo.2019.03.003

Lee, H., Kim, S. G., Park, H., & Kang, P. (2014). Pre-launch new product demand forecasting using the Bass model: A statistical and machine learning-based approach. *Technological Forecasting & Social Change*, *86*, 49–64. https://doi.org/10.1016/j.techfore.2013.08.020

Levitt, T. (1965). *Exploit the product life cycle*. https://hbr.org/1965/11/exploit-the-product-life-cycle

LU, J., CHEN, W., MA, Y., KE, J., LI, Z., ZHANG, F., & MACIEJEWSKI, R. (2017). Recent progress and trends in predictive visual analytics. *Frontiers of Computer Science*, *11*(2), 192–207. https://doi.org/10.1007/s11704-016-6028-y

Ma, C., Zhao, M., & Zhao, Y. (n.d.). An overview of Hadoop applications in transportation big data. *Journal of Traffic and Transportation Engineering (English Edition)*. https://doi.org/10.1016/j.jtte.2023.05.003

Makridakis, S., Spiliotis, E., & Assimakopoulos, V. (2022). M5 accuracy competition: Results, findings, and conclusions. *International Journal of Forecasting*, *38*(4), 1346–1364. https://doi.org/https://doi.org/10.1016/j.ijforecast.2021.11.013

Matthias, O., Fouweather, I., Gregory, I., & Vernon, A. (2017). Making sense of Big Data – can it transform operations management? *International Journal of Operations & Production Management*, *37*(1), 37–55. https://doi.org/10.1108/IJOPM-02-2015-0084

McCarthy, R. V, McCarthy, M. M., Ceccucci, W., & Halawi, L. (2019). *Applying predictive analytics: finding value in data*. Springer. https://doi.org/10.1007/978-3-030-14038-0

Meeran, S., Jahanbin, S., Goodwin, P., & Quariguasi Frota Neto, J. (2017). When do changes in consumer preferences make forecasts from choice-based conjoint models unreliable? *European Journal of Operational Research*, *258*(2), 512–524. https://doi.org/10.1016/j.ejor.2016.08.047

Mills, T. C. (2019). *Applied time series analysis: a practical guide to modeling and forecasting*. Academic Press. https://doi.org/10.1016/C2016-0-03956-6

Nielson. (2024). *About*. https://www.nielsen.com/about-us/about/

Nixtla. (2023). *Nixtla*. https://github.com/Nixtla/nixtla

Oracle. (2022). *What is data mining?* https://www.oracle.com/uk/big-data/what-is-data-mining/

Packowski, J. (2014). *LEAN supply chain planning : the new supply chain management paradigm for process industries to master today's VUCA world*. Crc Press.

Pazzani, M. J. (2000). Knowledge discovery from data? *IEEE Intelligent Systems & Their Applications*, *15*(2), 10–12. https://doi.org/10.1109/5254.850821

Petrosyan, A. (2024). *Main reasons for joining the metaverse 2021*. https://www.statista.com/statistics/1288870/reasons-joining-metaverse/

Piotrowski, C., & Guyette, R. W. (2010). Toyota Recall Crisis: Public Attitudes on Leadership and Ethics. *Organization Development Journal*, *28*(2), 89.

Porter, M. E. (1985). *Competitive advantage: creating and sustaining superior performance*. Free Press. https://go.exlibris.link/dT76p9gM

Pullman, M. E., Moore, W. L., & Wardell, D. G. (2002). A comparison of quality function deployment and conjoint analysis in new product design. *The Journal of Product Innovation Management*, *19*(5), 354–364. https://doi.org/10.1016/S0737-6782(02)00152-2

Rawlinson, J. G. (2017). *Creative Thinking and Brainstorming* (First). Taylor and Francis. https://doi.org/10.4324/9781315259000

Robert Jacobs, F., & 'Ted' Weston, F. C. (2007). Enterprise resource planning (ERP)—A brief history. *Journal of Operations Management*, *25*(2), 357–363. https://doi.org/10.1016/j.jom.2006.11.005

Roettig, E. (2016). *Inventory management and optimization in SAP ERP* (1st ed.). Rheinwerk. https://go.exlibris.link/zGvLJTQt

Russell, S. J., & Norvig, P. (2021). *Artificial intelligence: a modern approach* (Global;Fourth;). Pearson. https://go.exlibris.link/yH5fZfNq

Sakr, S. (2016). *Big data 2.0 processing systems: a survey*. Springer. https://doi.org/10.1007/978-3-319-38776-5

Sanders, N. R. (2017). *Forecasting fundamentals* (First). Business Expert Press. https://go.exlibris.link/Jpt0g6Lf

Saunders, M. N. K., Lewis, P., & Thornhill, A. (2019). *Research methods for business students* (Eighth). Pearson. https://go.exlibris.link/WRMtZFSH

Sawhney, M., Damkroger, L., McGuirk, G., Milbratz, J., & Rountree, J. (2017). *Illinois Superconductor Corp: forecasting demand for superconducting filters*. SAGE Publications Ltd.

Saxena, D., & Cao, J. (2022). Generative Adversarial Networks (GANs): Challenges, Solutions, and Future Directions. *ACM Computing Surveys*, *54*(3), 1–42. https://doi.org/10.1145/3446374

Scher, S., & Messori, G. (2019). Weather and climate forecasting with neural networks: using general circulation models (GCMs) with different complexity as a study ground. *Geoscientific Model Development*, *12*(7), 2797–2809. https://doi.org/10.5194/gmd-12-2797-2019

Shah, S., & Theodosoulaki, A. (2018). An Exploratory Study to Examine Big Data Application on Services and SCM. *2018 IEEE International Conference on Technology Management, Operations and Decisions (ICTMOD)*, 155–159. https://doi.org/10.1109/ITMC.2018.8691279

Shen, B., & Chan, H.-L. (2017). Forecast Information Sharing for Managing Supply Chains in the Big Data Era: Recent Development and Future Research. *Asia-Pacific Journal of Operational Research*, *34*(1), 1740001. https://doi.org/10.1142/S0217595917400012

Skenderi, G., Joppi, C., Denitto, M., & Cristani, M. (2024). Well googled is half done: Multimodal forecasting of new fashion product sales with image-based google trends. *Journal of Forecasting*, *43*(6), 1982–1997. https://doi.org/10.1002/for.3104

Slack, N., & Lewis, M. (2020). *Operations strategy* (Sixth). Pearson Education Limited. https://go.exlibris.link/H1vQxR4q

Smirnov, P. S., & Sudakov, V. A. (2021). Forecasting new product demand using machine learning. *Journal of Physics. Conference Series*, *1925*(1). https://doi.org/10.1088/1742-6596/1925/1/012033

Son, M., & Han, K. (2011). Beyond the technology adoption: Technology readiness effects on post-adoption behavior. *Journal of Business Research*, *64*(11), 1178–1182. https://doi.org/10.1016/j.jbusres.2011.06.019

Sondhi, R. (2008). *Total strategy* (3rd ed.). BMC Global Services Publications.

Srinivas, J., Das, A. K., Wazid, M., & Vasilakos, A. V. (2021). Designing Secure User Authentication Protocol for Big Data Collection in IoT-Based Intelligent Transportation System. *IEEE Internet of Things Journal*, *8*(9), 7727–7744. https://doi.org/10.1109/JIOT.2020.3040938

Tashakkori, A., & Teddlie, C. (Eds.). (2016). *SAGE handbook of mixed methods in social & behavioral research* (Second). SAGE. https://doi.org/10.4135/9781506335193

Tharenou, P., Donohue, R., & Cooper, B. (2007). *Management research methods*. Cambridge University Press. https://doi.org/10.1017/CBO9780511810527

Thomas, A. (2004). *Research skills for management studies*. Routledge. https://doi.org/10.4324/9780203006146

Tsafarakis, S., Grigoroudis, E., & Matsatsinis, N. (2011). Consumer choice behaviour and new product development: an integrated market simulation approach. *The Journal of the Operational Research Society*, *62*(7), 1253–1267. https://doi.org/10.1057/jors.2010.70

Valenzuela, O., Rojas Ruiz, F., Herrera, L. J., Pomares, H., & Rojas, I. (Eds.). (2023). *Theory and applications of time series analysis: selected contributions from ITISE 2022*. Springer. https://go.exlibris.link/QN0hZ8SM

van Steenbergen, R. M., & Mes, M. R. K. (2020). Forecasting demand profiles of new products. *Decision Support Systems*, *139*, 113401. https://doi.org/10.1016/j.dss.2020.113401

Vollmann, T. E., Berry, W. L., & Whybark, D. C. (2005). *Manufacturing planning and control systems for supply chain management: the definitive guide for professional* (Fifth). McGraw-Hill Education. https://go.exlibris.link/f6dLZMXq

Vrbka, J. (2021). *Using artificial neural networks for timeseries smoothing and forecasting: case studies in economics: Vol. 979.* Springer. https://doi.org/10.1007/978-3-030-75649-9

Wang, G., Gunasekaran, A., Ngai, E. W. T., & Papadopoulos, T. (2016a). Big data analytics in logistics and supply chain management: Certain investigations for research and applications. *International Journal of Production Economics*, *176*, 98–110. https://doi.org/https://doi.org/10.1016/j.ijpe.2016.03.014

Wang, G., Gunasekaran, A., Ngai, E. W. T., & Papadopoulos, T. (2016b). Big data analytics in logistics and supply chain management: Certain investigations for research and applications. *International Journal of Production Economics*, *176*, 98–110. https://doi.org/10.1016/j.ijpe.2016.03.014

Wang, Y., Xia, S.-T., Tang, Q., Wu, J., & Zhu, X. (2018a). A Novel Consistent Random Forest Framework: Bernoulli Random Forests. *IEEE Transaction on Neural Networks and Learning Systems*, *29*(8), 3510–3523. https://doi.org/10.1109/TNNLS.2017.2729778

Wang, Y., Xia, S.-T., Tang, Q., Wu, J., & Zhu, X. (2018b). A Novel Consistent Random Forest Framework: Bernoulli Random Forests. *IEEE Transaction on Neural Networks and Learning Systems*, *29*(8), 3510–3523. https://doi.org/10.1109/TNNLS.2017.2729778

Welch, T. F., & Widita, A. (2019). Big data in public transportation: a review of sources and methods. *Transport Reviews*, *39*(6), 795–818. https://doi.org/10.1080/01441647.2019.1616849

Whittington, R., Regnér, P., Angwin, D., Johnson, G., & Scholes, K. (2020). *Exploring strategy: text and cases* (Twelfth). Pearson Education Limited. https://go.exlibris.link/rRHkZp6P

Wu, X., Zhu, X., Wu, G.-Q., & Ding, W. (2014). Data mining with big data. *IEEE Transactions on Knowledge and Data Engineering*, *26*(1), 97–107. https://doi.org/10.1109/TKDE.2013.109

Yamamura, C. L. K., Santana, J. C. C., Masiero, B. S., Quintanilha, J. A., & Berssaneti, F. T. (2022). Forecasting New Product Demand Using Domain Knowledge and Machine Learning. *Research Technology Management*, *65*(4), 27–36. https://doi.org/10.1080/08956308.2022.2062553

Zhang, S., Zhou, L., Chen, X. (Michael), Zhang, L., Li, L., & Li, M. (2020). Network-wide traffic speed forecasting: 3D convolutional neural network with ensemble empirical mode decomposition. *Computer-Aided Civil and Infrastructure Engineering*, *35*(10), 1132–1147. https://doi.org/10.1111/mice.12575

Zhao, M., Shen, X., Liao, H., & Cai, M. (2022). Selecting products through text reviews: An MCDM method incorporating personalized heuristic judgments in the prospect theory. *Fuzzy Optimization and Decision Making*, *21*(1), 21–44. https://doi.org/10.1007/s10700-021-09359-8

Zhao, P., & Cao, Y. (2020). Commuting inequity and its determinants in Shanghai: New findings from big-data analytics. *Transport Policy*, *92*, 20–37. https://doi.org/10.1016/j.tranpol.2020.03.006

Zheng, X., Chen, W., Wang, P., Shen, D., Chen, S., Wang, X., Zhang, Q., & Yang, L. (2016). Big Data for Social Transportation. *IEEE Transactions on Intelligent Transportation Systems*, *17*(3), 620–630. https://doi.org/10.1109/TITS.2015.2480157

Appendices

Appendix A: M5 Forecasting Accuracy Analysis

Table A.1: Sample Data Rows from Calendar.csv

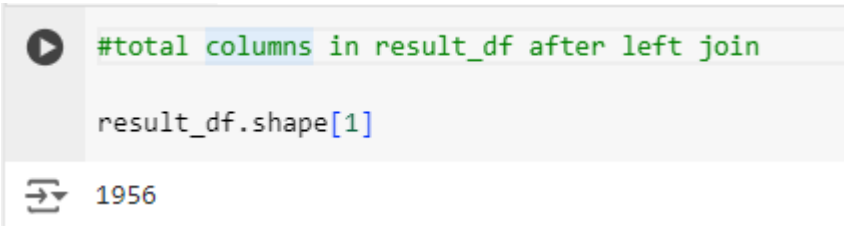
date	wm_yr_wk	weekday	wday	month	year	d	event_nan	event_type	event_nan	event_type	snap_CA	snap_TX	snap_WI
06/02/2011	11102	Sunday	2	2	2011	d_9	SuperBow	Sporting			1	1	1
07/02/2011	11102	Monday	3	2	2011	d_10					1	1	0
08/02/2011	11102	Tuesday	4	2	2011	d_11					1	0	1
09/02/2011	11102	Wednesda	5	2	2011	d_12					1	1	1
10/02/2011	11102	Thursday	6	2	2011	d_13					1	0	0
11/02/2011	11102	Friday	7	2	2011	d_14					0	1	1
12/02/2011	11103	Saturday	1	2	2011	d_15					0	1	1
13/02/2011	11103	Sunday	2	2	2011	d_16					0	1	0
14/02/2011	11103	Monday	3	2	2011	d_17	Valentines	Cultural			0	0	1

Table A.2: Sample Data Rows from Sales_train_evaluation.csv

id	item_id	dept_id	cat_id	store_id	state_id	d_1	d_2 …
HOBBIES_1_001_CA_1_evaluation	HOBBIES_1_001	HOBBIES_1	HOBBIES	CA_1	CA	0	0
HOBBIES_1_002_CA_1_evaluation	HOBBIES_1_002	HOBBIES_1	HOBBIES	CA_1	CA	0	0
HOBBIES_1_003_CA_1_evaluation	HOBBIES_1_003	HOBBIES_1	HOBBIES	CA_1	CA	0	0
HOBBIES_1_004_CA_1_evaluation	HOBBIES_1_004	HOBBIES_1	HOBBIES	CA_1	CA	0	0
HOBBIES_1_005_CA_1_evaluation	HOBBIES_1_005	HOBBIES_1	HOBBIES	CA_1	CA	0	0
HOBBIES_1_006_CA_1_evaluation	HOBBIES_1_006	HOBBIES_1	HOBBIES	CA_1	CA	0	0
HOBBIES_1_007_CA_1_evaluation	HOBBIES_1_007	HOBBIES_1	HOBBIES	CA_1	CA	0	0
HOBBIES_1_008_CA_1_evaluation	HOBBIES_1_008	HOBBIES_1	HOBBIES	CA_1	CA	12	15
HOBBIES_1_009_CA_1_evaluation	HOBBIES_1_009	HOBBIES_1	HOBBIES	CA_1	CA	2	0

An excerpt of rows from the sales_train_evaluation dataset, showcasing the features used in the analysis. It has 1947 columns and 30490 rows in total before data aggregation.

Figure A.1: Columns after left join on two datasets



This Figure illustrates the total number of columns of the resulting dataset after performing a SEQUEL Left Join between the calendar data frame and sales_train_evaluaton.csv. The left join operation retains all columns from calendar.csv and appends columns from s sales_train_evaluaton.csv matching value "**HOUSEHOLD_1_118**".

Figure A.2: Aggregate sales of HOUSEHOLD_1_118 on Christmas and New Year

date	event_name_1	aggregate department sales countrywide for HOUSEHOLD_1_118
2011-12-25	Christmas	0
2012-01-01	NewYear	11
2012-12-25	Christmas	0
2013-01-01	NewYear	0
2013-12-25	Christmas	0
2014-01-01	NewYear	11
2014-12-25	Christmas	0
2015-01-01	NewYear	4
2015-12-25	Christmas	0
2016-01-01	NewYear	1

Figure A.3: Effect of Key Events on Demand

Event Name	Aggregate Sales
SuperBowl	87
OrthodoxEaster	81
LaborDay	79
MemorialDay	61
Easter	55
Pesach End	55
Eid al-Fitr	54
Purim End	54
VeteransDay	51
PresidentsDay	49

This figure illustrates how specific events influence demand. On the left are events labelled in the calendar dataset while on the right are aggregate sales of HOUSEHOLD_1_118 associated with each event.

Figure A.4: Python Code for Linear Regression

```
[ ]  from sklearn.linear_model import LinearRegression

     # Prepare data
     new_year_data = result_df_filtered[result_df_filtered['event_name_1'] == 'NewYear']
     X = new_year_data.index.year.values.reshape(-1, 1)
     y = new_year_data["aggregate department sales countrywide for HOUSEHOLD_1_118"].values

     model = LinearRegression()
     model.fit(X, y)

     # Predict for 2016
     prediction_2016_LR_ = model.predict([[2016]])
     print("Predicted sales for New Year 2016:", prediction_2016_LR_[0])

Predicted sales for New Year 2016: 2.199999999999818
```

Figure A.5: Python Code for ARIMA

```
     from statsmodels.tsa.arima.model import ARIMA
     from sklearn.metrics import mean_squared_error

     # Prepare data
     new_year_data = result_df_filtered[result_df_filtered['event_name_1'] == 'NewYear']
     y = new_year_data["aggregate department sales countrywide for HOUSEHOLD_1_118"].values

     model = ARIMA(y, order=(5, 1, 0))
     model_fit = model.fit()

     # Forecast for 2016
     prediction_2016_AR_ = model_fit.forecast(steps=1)[0]
     print("Predicted sales for New Year 2016 (ARIMA):", prediction_2016_AR_)
```

Figure A.6: Error Code for ARIMA

```
/usr/local/lib/python3.10/dist-packages/statsmodels/tsa/statespace/sarimax.py:866: UserWarning:
Too few observations to estimate starting parameters for ARMA and trend. All parameters except for variances will be set to zeros.
Predicted sales for New Year 2016 (ARIMA): 10.220674856150586
/usr/local/lib/python3.10/dist-packages/statsmodels/base/model.py:607: ConvergenceWarning:

Maximum Likelihood optimization failed to converge. Check mle_retvals
```

This error code has been generated by the ARIMA model indicating that the algorithm could not proceed because there are too few observations. The last error line signifies its failure to optimize the likelihood.

Figure A.7: Python Code for Random Forest

```
[ ]  from sklearn.ensemble import RandomForestRegressor

     # Prepare data
     new_year_data = result_df_filtered[result_df_filtered['event_name_1'] == 'NewYear']
     X = new_year_data.index.year.values.reshape(-1, 1)
     y = new_year_data["aggregate department sales countrywide for HOUSEHOLD_1_118"].values

     model = RandomForestRegressor(n_estimators=100, random_state=42)
     model.fit(X, y)

     # Predict for 2016
     prediction_2016_RF_ = model.predict([[2016]])
     print("Predicted sales for New Year 2016 (Random Forest):", prediction_2016_RF_[0])
```

```
Predicted sales for New Year 2016 (Random Forest): 2.43
```

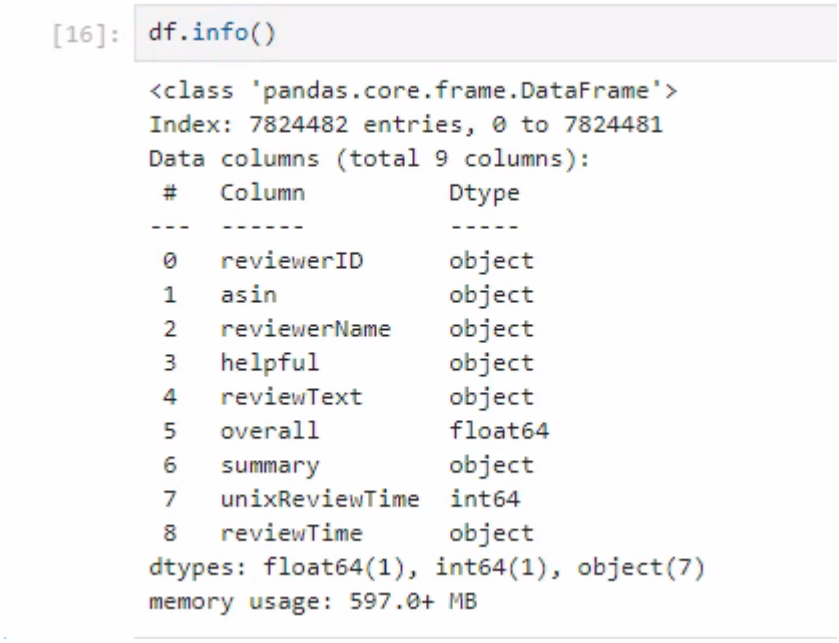
Appendix B: Amazon Review Dataset

Table B.1: Example View Amazon Review Dataset, Electronics

reviewerID	asin	reviewerName	helpful	reviewText	overall	summary	reviewTime
A2N2L3DD0XTLAH	B0039ADJHO	Brad	[0, 0]	Great headset	5	Great headset	07 2, 2014
A16YX9NHQ1EZRG	B0032HM6JG	Lee Family	[0, 0]	Great headset	5	Five Stars	07 10, 2014
ATLRO6T0UBB2Y	B003VKVXF2	Neil "Redman"	[0, 0]	Great headset.	5	Five Stars	07 1, 2014
A2U6JFE9NWNJ7Q	B000CNAEEW	Andrew G.	[0, 0]	Awesome headsets	5	Five Stars	07 5, 2014
A172VMVPOCM0NQ	B00I51BWAS	Dafull97	[0, 0]	Awesome headset!	4	Awesome Headset!	07 21, 2014
ALX4PN3JYHOGI	B004RKQM8I	tom raleigh	[0, 0]	Excellent headset	5	Great	07 12, 2014
A1U4I1IVTXDES1	B000UXZQ42	Richard C. Hallowell	[0, 0]	Excellent headset	5	Five Stars	07 4, 2014

This table displays a visual representation of the Amazon Review Dataset, for the category Electronics, showcasing columns such as ‘reviewerID’, ‘asin’, ‘reviewerName’, ‘helpful’, ‘reviewText’, ‘overall’, ‘summary’ and ‘reviewTime’.

Figure B.1: Electronics Summary Generated by Python



This screenshot displays Python output describing the electronics data frame with its statistics, features, memory, datatypes and metadata.

Figure B.2: Complete List of Phrases for Review Filtering

```
for phrase in phrases:
    print(phrase)
```

```
more colors
different sizes
new models
additional
upgraded
needs
need
would have liked
poor quality
complicated
user-friendly
buggy
disappointment
```

The figure represents a complete list of phrases used to remove spam from the improvements review subset and save relevant user reviews.

Figure B.3: Final Count of Improvements Reviews

```
[62]: len(improvements)

[62]: 68
```

This is a Python screenshot representing the total number of reviews remaining in improvements after applying the filtering criteria shown in Figure C2 in Appendix C.

Appendix C: Ethics Training

Ethical Approval Number: WMG-R_2iydcYUpUFk0cOy

Figure C1: Email Confirmation

Ethical approval is granted (WMG Taught Student Research)

Qualtrics Survey Software <noreply@qemailserver.com>
Wed 22/05/2024 13:04
To: DAL, JAVERIA (PGT) <Javeria.Dal@warwick.ac.uk>

Date: **May 22, 2024**

Student: **Javeria Dal**
Student ID number: **5508978**
Project title: **Predicting future demand using machine learning**
Your ethical approval number is **WMG-R_2iydcYUpUFk0cOy**

Your supervisor **Dave Food** has **granted ethical approval for your project.**

This means you have consent to conduct your research.

You now have the appropriate approval in place to begin your data collection. It is advisable to note the actual dates of data collection in the final project submission to evidence that your data was collected after the date of ethical approval (as stated in this email).

You are reminded that you must now adhere to the answers and detail given in the completed ethics form. If anything changes in your research such that any of your answers change to the form for which you received **ethical approval** for, then you must contact your supervisor to check if you need to reapply for or update your **ethical approval** before you proceed with data collection.

When you submit your project, please write **your ethical approval number against the ethical approval field** on the cover page of the submission and include a copy of this email in the Appendices of the submission.

Kind regards,

WMG Projects Team

Download as PDF

Figure C2: SPA Badges

Download

2024 Ethics in Research

Awarded to Javeria Dal

Issued 11 March 2024, 10:36 AM

Issued by moodle.warwick

Course: WM999: Study, Professional & Analytical Skills

2024 Ethics in Research on FT MSc SPA Moodle site

Criteria

- The following activity has to be completed:
  "Quiz - 22-23 Research Ethics - Part C"

Download

2024 Ethical Approval Process

Awarded to Javeria Dal

Issued 15 May 2024, 9:24 PM

Issued by moodle.warwick

Course: WM999: Study, Professional & Analytical Skills

2024 Ethical Approval Process on FT MSc SPA Moodle site

Figure C3: Research Integrity Training Programme: Epigeum

Figure C4: Protecting Human Participants Module: Epigeum