

Project Dissertation Report on
Retail Forecasting and Stockout Detection Using
Machine Learning: A Unified Comparative
Framework

Submitted By:

Rishita Saxena

24/BMBA/31

Under the Guidance of

Dr. Yogesh Sharma

Assistant Professor



DELHI SCHOOL OF MANAGEMENT

Delhi Technological University

Bawana Road Delhi-110042

CERTIFICATE

This is to certify that the Major Research Project titled “Retail Forecasting and Stockout Detection Using Machine Learning: A Unified Comparative Framework” has been completed by Ms. Rishita Saxena, Roll No. 24/BMBA/31, for the partial fulfilment of the requirements for the award of the degree of Master of Business Administration (MBA) in Business Analytics from Delhi School of Management, Delhi Technological University.

Date: _____

Signature of Guide: _____

Dr. Yogesh Sharma

Assistant Professor

Delhi School of Management, DTU

DECLARATION

I hereby declare that the project report titled “Retail Forecasting and Stockout Detection Using Machine Learning: A Unified Comparative Framework” is my original work and has been carried out under the guidance of my faculty mentor.

Date:

Place:

(Signature)

Rishita Saxena

24/BMBA/31

ACKNOWLEDGEMENT

I would like to express my sincere gratitude to my faculty guide for their valuable guidance, continuous support, and encouragement throughout the completion of this project.

I am also thankful to the Delhi School of Management, Delhi Technological University, for providing the necessary resources and academic environment to carry out this research work.

Lastly, I would like to thank my family and friends for their constant support and motivation.

Rishita Saxena

24/BMBA/31

EXECUTIVE SUMMARY

The retail industry faces a persistent and costly challenge in the form of stockouts, situations where product inventory is depleted before customer demand is met. These events not only result in direct revenue losses but also damage customer loyalty and disrupt supply chain efficiency. The advancement of machine learning technologies, there is a growing opportunity to develop intelligent systems capable of predicting demand patterns and detecting stockout risks before they occur. This study addresses that opportunity through a structured two phase comparative framework applied to sales data of retail.

The primary objective of this study is to compare the performance of four machine learning models, Linear Regression, Random Forest, XGBoost, and Long Short Term Memory networks, across two distinct but interconnected phases. Phase 1 focuses on demand forecasting where models predict future sales values and are evaluated using Root Mean Squared Error and Mean Absolute Error. Phase 2 introduces a novel stockout classification layer where the same models are retrained as classifiers to predict whether a stockout event will occur, evaluated using F1 Score and AUC-ROC metrics.

The study utilizes the Rossmann Store Sales dataset which contains historical daily sales records from 1115 retail stores across Germany spanning several years. The dataset includes key retail features such as promotional activity, state and school holidays, store type, assortment level, and competitor proximity. After merging and preprocessing the data, a binary stockout label was engineered by identifying days where a store's sales dropped below 50 percent of its historical average while remaining open, a threshold determined through systematic experimentation across multiple levels. To address the severe class imbalance of only 1.65 percent stockout cases in the dataset, the Synthetic Minority Oversampling Technique known as SMOTE was applied to balance the training data before classification model training.

Experimental results from Phase 1 reveal that Random Forest achieved the best demand forecasting performance with an RMSE of 922.26 and MAE of 593.05, outperforming XGBoost which recorded an RMSE of 1200.56 and MAE of 859.80. Linear Regression performed the worst with an RMSE of 2805.82 confirming that retail sales data exhibits strong non linear patterns that simple models cannot capture. LSTM recorded an RMSE of 2520.57 suggesting that without proper sequential time series structuring per store, deep learning models do not automatically outperform traditional ensemble methods.

In Phase 2, Random Forest again demonstrated the strongest classification performance with an F1 Score of 0.6338 and AUC-ROC of 0.9722, making it the most reliable model for detecting actual stockout events. XGBoost achieved the highest AUC-ROC of 0.9757 indicating superior ability to rank stockout risk across probability thresholds, however its lower F1 Score of 0.4897 suggests it is less precise in directly classifying individual stockout days. LSTM recorded an F1 Score of 0.2829 and Logistic Regression performed the weakest with an F1 Score of 0.1134.

TABLE OF CONTENTS

S.NO	Topic	Page No.
1.	Certificate	I
2.	Declaration	II
3.	Acknowledgement	III
4.	Executive Summary	IV
5.	Table of Contents	V
6.	Introduction	1
7.	Literature Review	7
8.	Research Methodology	12
9.	Data Analysis	21
10.	Discussion & Recommendations	30
11.	Conclusion	35
13.	References	41
14.	Annexure	43

CHAPTER 1:

INTRODUCTION

1.1 Background

The retail sector around the globe can be considered as one of the fastest-growing and most competitive industries in the entire world, with millions of organizations involved, including local vendors and international chains of shops. However, there is one key issue that any successful retail company needs to address – making sure that the necessary goods are always supplied in sufficient amounts to satisfy the demands of customers. The general issue is also called inventory management and has been the center of much scholarly research and attention in the corporate world for years now. Despite the latest improvements in logistics and technological solutions used in supply chains, there are still some inventory problems, such as overstocking, when too much goods accumulate, which causes waste and higher costs, and understocking (or stockouts), when there is not enough stock available.

A stockout occurs when a product is unavailable at the point of sale due to depleted inventory levels. From a customer perspective, stockouts are frustrating experiences that often lead to brand switching, reduced loyalty, and long-term loss of business. From a retailer's perspective, stockouts represent a direct and immediate loss of revenue. According to industry estimates, stockouts cost the global retail industry approximately one trillion dollars annually, making it one of the most financially significant operational challenges in the sector. In addition to direct revenue loss, stockouts create ripple effects across the supply chain, disrupting supplier relationships, distorting demand signals, and increasing the complexity of replenishment planning.

Historically, retailers relied on manual inventory tracking, periodic physical counts, and simple reorder point systems to manage stock levels. While these approaches provided a basic level of inventory control, they were inherently reactive rather than proactive. By the time a stockout was detected, the damage in terms of lost sales and customer dissatisfaction had already occurred. The need for more intelligent, data-driven, and predictive approaches to inventory management became increasingly apparent as retail operations grew in scale and complexity.

Data Science and Machine Learning have brought about significant changes in how organizations forecast and predict. The algorithms of machine learning can recognize intricate patterns in data that traditional statistical models cannot. In relation to retail, various types of problems, ranging from customer segmentation to price optimization and recommendation systems, have been tackled using machine learning. However, demand forecasting has been the application of machine learning in retail most pertinent to this project. Demand forecasting is a prediction of how many units of goods will be sold within a certain period of time, and an accurate forecast will be critical in inventory management.

Early approaches to retail demand forecasting were dominated by classical statistical time series methods such as Moving Averages, Exponential Smoothing, and Autoregressive Integrated Moving Average models commonly known as ARIMA. These methods were mathematically elegant and relatively easy to interpret, but they carried significant limitations. They assumed linear relationships between past and future values, struggled to incorporate external variables such as promotions and holidays, and required individual

model training for each product or store, a computationally prohibitive requirement for large retail chains with thousands of products across hundreds of locations.

The introduction of machine learning based approaches such as Decision Trees, Random Forests, Gradient Boosting machines, and deep learning models like Long Short Term Memory networks brought a new level of sophistication to retail forecasting. These models could handle large volumes of data, capture nonlinear patterns, incorporate multiple external features simultaneously, and generalize across products and stores without requiring individual model training. The Kaggle M5 Forecasting Competition and the Rossmann Store Sales Competition became landmark benchmarks in the field, attracting thousands of researchers and data scientists to develop and test new forecasting approaches on real world retail data.

Despite the significant progress in demand forecasting research, a critical gap has persisted in the existing literature. Most studies have focused exclusively on predicting sales number, a regression problem, without translating those predictions into actionable inventory decisions. A retailer does not simply need to know how much will be sold tomorrow. They need to know whether their current stock level is sufficient to meet that demand, and if not, when and how much to reorder. This gap between demand forecasting and stockout prediction represents a significant disconnect between academic research and real-world retail operations.

Furthermore, the problem of stockout prediction is inherently different from demand forecasting in its mathematical formulation. While demand forecasting is a regression problem that produces continuous numerical outputs, stockout prediction is a classification problem that produces binary outputs, will a stockout occur or will it not. The two problems require different modeling approaches, different evaluation metrics, and different interpretations of results. Very few studies have addressed both problems within a unified comparative framework, and even fewer have done so using the same set of models evaluated on the same dataset with consistent methodology.

In recent years, the availability of large-scale retail datasets and the democratisation of machine learning tools through platforms like Python, scikit-learn, TensorFlow, and Google Colab have made it increasingly accessible for researchers to conduct comprehensive comparative studies. These studies play a vital role in advancing the field by establishing which models perform best under which conditions, providing practitioners with evidence based guidance for model selection, and identifying research gaps that future studies can address.

India represents a particularly important context for retail forecasting research. With one of the fastest growing retail markets in the world, Indian retail is characterized by high demand volatility, strong seasonal patterns driven by festivals and cultural events, diverse consumer behavior across regions, and a rapidly expanding organized retail sector. Despite these characteristics that make demand forecasting both challenging and critical, there is a notable scarcity of retail forecasting research specifically addressing Indian retail contexts. Most existing studies are based on Western retail datasets, and their findings may not generalize directly to the unique patterns and dynamics of Indian retail.

In light of this situation, the current study aims to make its own contribution to the growing pool of knowledge regarding retail inventory intelligence through the formulation of a unified two-stage comparative model dealing with both forecasting of demand and stockouts. Through the comparison of several machine learning models in both stages of this research process, insights are intended to be generated for both academic and practical purposes.

1.2 Problem Statement

Despite the widespread adoption of machine learning techniques in retail analytics, existing comparative studies suffer from three fundamental limitations that this study seeks to address.

The first limitation is the narrow focus on demand forecasting as an isolated problem. The vast majority of existing studies treat retail forecasting purely as a regression task, predicting future sales values, without extending the analysis to the practical question of whether predicted demand will exceed available inventory. This creates a significant gap between research outputs and real-world applicability. A retailer who knows that sales will be 5000 units tomorrow gains limited actionable value from that prediction alone. What they truly need to know is whether their current stock of say 3000 units is sufficient, and if not, when to trigger a replenishment order. Existing studies have largely failed to bridge this gap.

The second limitation is the inconsistency in comparative methodologies across existing studies. Different studies use different datasets, different train test splits, different feature engineering approaches, and different evaluation metrics, making it difficult to draw reliable conclusions about the relative performance of different models. There is a clear need for a unified comparative framework that evaluates multiple models on the same dataset using the same methodology and the same evaluation metrics across both forecasting and classification tasks.

The third limitation is the inadequate handling of class imbalance in stockout prediction. Stockout events are by their nature rare occurrences, they represent a small fraction of total trading days. When this imbalance is not properly addressed, classification models tend to achieve misleadingly high accuracy by simply predicting that no stockout will occur on every day. This results in models that appear highly accurate on paper but are completely useless in practice for detecting the very events they are designed to predict. Existing studies have largely overlooked this issue, leading to inflated and unreliable performance metrics.

This study addresses all three limitations by developing a two-phase framework that combines demand forecasting and stockout classification, applying consistent methodology across all models and both phases, and explicitly handling class imbalance using the Synthetic Minority Oversampling Technique. The central research question guiding this study is: which machine learning model provides the most robust and reliable performance across both retail demand forecasting and stockout classification tasks when evaluated within a unified comparative framework.

1.3 Objectives of the Study

The following objectives guide the design and execution of this research study:

1. To conduct a comprehensive review of existing literature on retail demand forecasting and stockout prediction, identifying the key models, datasets, methodologies, and research gaps in the field.
2. To develop a two-phase comparative framework that evaluates machine learning models on both demand forecasting as a regression task and stockout detection as a classification task within a single unified study.
3. To compare the performance of four machine learning models across two phases, demand forecasting using RMSE and MAE as evaluation metrics, and stockout classification using F1 Score and AUC-ROC as evaluation metrics.
4. To address the class imbalance problem inherent in stockout prediction by engineering a binary stockout label through systematic threshold experimentation and applying the Synthetic Minority Oversampling Technique to balance the training data.
5. To identify the best-performing model across both phases and provide evidence-based recommendations for retail businesses on model selection for inventory management and stockout prevention applications.

1.4 Scope of Study

The scope of this study is defined along four dimensions, data scope, methodological scope, analytical scope, and contextual scope.

Data Scope

The dataset used in this study was collected from the Rossmann Store Sales provided by Kaggle. This dataset has the daily sales information of 1115 stores located in Germany. It has multiple years of trading history with information on promotion, state, and school holiday, type of store and its level of assortment, as well as competition distance. Two main files were extracted from the Rossmann Store Sales dataset for this research: the train.csv, which has information on daily sales, and the store.csv, which provides the characteristics of the stores, including their identification number, which were merged through an inner join on the common column.

Methodological Scope

Methodologically, this research involves data preprocessing, feature engineering, stockout labels generation, handling imbalanced classes, training models, and evaluating models through two stages. This research applies the supervised machine learning approach, where models are trained on past data and tested on held-out datasets using 80-20 training-test split. The stockout label is generated using the sales fall-off threshold method and not using actual inventory information because no data regarding actual inventory levels were available in the dataset. It can be viewed as one of the methodological challenges of this research, which is acknowledged and described in the Limitations chapter. Imbalanced classes are mitigated by employing SMOTE only on the training set.

Analytical Scope

The analytical scope covers four machine learning models evaluated across two phases. In Phase 1 the models are used as regressors to predict continuous sales values, evaluated using RMSE and MAE. In Phase 2 the same models are used as classifiers to predict binary stockout outcomes, evaluated using F1 Score and AUC-ROC. The study does not include hyperparameter tuning beyond default settings for most models, as the primary objective is comparative analysis rather than individual model optimization. Feature importance analysis is conducted for Random Forest to provide additional insights into the key drivers of sales and stockout patterns.

Contextual Scope

While the dataset used in this study is from European retail, the findings and framework have broader applicability to retail contexts globally including India. The study does not claim to provide a production ready stockout prediction system but rather a research framework and evidence base for model selection in retail inventory management applications. The study is limited to the models and dataset described above and does not extend to deep reinforcement learning, probabilistic forecasting, or real time prediction systems which represent promising directions for future research.

1.5 Motivation of the Study

The motivation for this study stems from both a practical observation and an academic gap identified during the review of existing literature on retail analytics and machine learning.

From a practical standpoint, the persistent and widespread nature of the retail stockout problem despite decades of research and significant investment in inventory management technology suggests that existing approaches are fundamentally incomplete. The disconnect between what retailers need, a system that tells them when stock will run out and triggers a timely replenishment action, and what most academic research provides, a sales forecast number, represents a gap with real and measurable financial consequences. The trillion dollar annual cost of retail stockouts globally is not merely an abstract statistic. It represents lost revenue for retailers, wasted journeys for customers, and inefficiency in the global supply chain that affects producers, distributors, and consumers alike. This practical urgency motivated the development of a framework that goes beyond sales forecasting to directly address the stockout prediction problem.

From an academic standpoint the review of existing literature revealed that while numerous studies have compared machine learning models for retail demand forecasting none have done so within a framework that simultaneously addresses both forecasting accuracy and stockout classification within a single unified study using the same models, same data, and same evaluation protocol across both tasks. This gap represents a genuine opportunity to contribute methodological novelty to the field. By framing the problem as two interconnected phases rather than two separate studies this research creates a more holistic and practically relevant analytical framework than existing comparative studies have achieved.

The choice of the Rossmann Store Sales dataset as the empirical foundation for this study was motivated by three considerations. First it is one of the most widely used and well validated benchmark datasets in retail forecasting research making the results of this study

directly comparable to a large body of existing literature. Second it is rich enough in terms of store coverage, time span, and feature variety to support the development of a meaningful two phase analytical framework. Third, it is publicly available through the Kaggle platform, ensuring that the study is fully reproducible by other researchers who wish to validate or extend its findings.

CHAPTER 2

LITERATURE REVIEW

2.1 Introduction to Literature Review

The realm of retail demand forecasting and stockout prediction has come a long way in the last two decades, moving from the use of statistical techniques to the application of sophisticated machine learning and deep learning models. The current chapter provides a thorough review of past literature on the topics of retail demand forecasting, comparative analyses of machine learning models, stockout prediction, and imbalanced classification problems.

2.2 Traditional Approaches to Retail Demand Forecasting

The early work in forecasting retail demand mainly focused on traditional statistical approaches. One such technique that enjoyed wide acceptance because of its simplicity and reliability is the AutoRegressive Integrated Moving Average models known as ARIMA. Numerous studies conducted during the period from the 1990s to the early 2000s showed the effectiveness of both ARIMA and its seasonally adjusted counterpart, SARIMA, for linear trend and seasonality forecast of retail sales data. Chuang (1991) came up with a SARIMA model for the forecast of computer parts demand and proved its superiority over simpler forecasting models. The dominance of SARIMA continued when Velos et al. (2020) used it successfully in the alcoholic beverages industry.

However these classical methods carried significant limitations. They assumed linear relationships between past and future values, required data stationarity, struggled to incorporate external variables such as promotions and holidays, and required individual model training for each product or store, a computationally prohibitive requirement for large retail chains with thousands of products. Exponential smoothing and Moving Average methods shared these same fundamental weaknesses, prompting researchers to explore richer multivariate approaches. As noted by Taparia et al. (2023), accurate demand forecasting requires models capable of handling rapidly shifting customer preferences, frequent promotional campaigns, and external disruptions, characteristics that statistical methods alone are ill-equipped to address. This growing recognition of the limitations of traditional approaches created the foundation for the shift toward machine learning-based forecasting that would define the next era of research.

2.3 Machine Learning Approaches for Retail Forecasting

The introduction of machine learning methods brought a transformative shift to retail demand forecasting research. Decision trees, ensemble methods, and gradient boosting algorithms demonstrated superior ability to handle nonlinear patterns, incorporate multiple features simultaneously, and generalize across products and stores without requiring individual model training.

Among the most cited studies in the field is the one done by Mitra et al. (2022), which compared five machine learning regression algorithms: Random Forest, XGBoost, Gradient Boosting, AdaBoost, and Artificial Neural Network, against a newly proposed hybrid algorithm (RF-XGBoost-LR) for sales prediction in a US retail company based on weekly sales data from 45 stores. In their analysis, the authors have discovered that ensemble techniques generally yield better results than individual algorithms, and that the hybrid RF-XGBoost technique is the best among all tested, reaching 91 percent accuracy, thus proving that combining several machine learning algorithms leads to better results when it comes to predicting retail sales (Mitra et al., 2022).

From this perspective, Prabu et al. (2025) carried out an extensive analysis of various models of machine learning algorithms for the purpose of forecasting sales by means of the Rossmann Store Sales dataset as a benchmark. Specifically, ten distinct models were analyzed by Prabu et al. (2025) in the form of KNN, Decision Tree, Random Forest, Ridge Regression, Linear Regression, XGBoost, ElasticNet, Gradient Boosting, RANSAC, and LightGBM. The outcomes of this research indicated that the Random Forest model was superior to all others in terms of MSE, RMSE, MAE, and R-squared indicators. Thus, Random Forest is deemed to be the optimal solution for sales forecasting in the context of the Rossmann Store Sales dataset due to its efficiency and reliability (Prabu et al., 2025). It should be noted that although the XGBoost and Gradient Boosting models also exhibited high levels of performance, the Ridge Regression and Linear Regression models could not produce satisfactory results.

A study published in the International Journal of Advanced Computer Science and Applications in 2025 further reinforced these findings by demonstrating that demand forecasting in retail benefits significantly from incorporating big data sources including market trends and external factors, with Random Forest and XGBoost emerging as the most robust algorithms for analyzing broad arrays of data and enhancing forecasting accuracy (IJARCCE, 2025). The study also highlighted the growing importance of alert systems and inventory optimization layers built on top of forecasting models, a direction that directly aligns with the two phase framework developed in the present study.

2.4 Deep Learning Approaches for Retail Forecasting

Deep learning neural network architectures, more specifically LSTM networks, gave rise to a novel aspect to forecasting research in retail. The LSTM network was thought to be ideally suited for time-series forecasting because of its capacity to consider long-term temporal relationships via the use of gated memories, which consist of input, forget, and output gates controlling data transfer (Hochreiter & Schmidhuber, 1997).

Punia et al. (2020) conducted an influential study combining LSTM with Random Forest for demand forecasting in a multi-channel retail setting, finding that the hybrid deep learning approach outperformed pure deep learning methods as well as the ARIMAX benchmark. This work demonstrated that while LSTM offers strong temporal modelling capabilities, combining it with ensemble methods could yield superior performance, suggesting that neither approach alone represents the optimal solution for all retail forecasting contexts.

A paper published in the Journal of Forecasting by Castro Moraes and Bandeira-de-Mello (2024) proposed hybrid CNN-LSTM models for retail sales forecasting and evaluated parallel versus stacked deep learning architectures. The study found that hybrid approaches combining CNN for feature extraction and LSTM for temporal dependency modelling outperformed standalone LSTM models, particularly for datasets characterised by seasonality and unusual demand trends, conditions that are clearly present in the Rossmann dataset used in the present study.

Research on optimising LSTM networks for retail forecasting by Syntetos et al. (2026) noted that while deep learning models, including LSTM have shown promise in reducing forecasting errors by 20 to 30 percent compared to traditional methods, significant deployment challenges remain, including high memory requirements and the need for large training datasets. The study also found that reducing the number of LSTM hidden units from 128 to 64 improved prediction accuracy by 47 per cent while reducing model size by 73 per cent, suggesting that model capacity alignment with problem characteristics is more important than maximizing architectural complexity.

A more recent comparative study by Hobor et al. (2026) investigated statistical benchmarks, tree based ensembles including both XGBoost and LightGBM, and deep learning approaches including N-BEATS and the Temporal Fusion Transformer on realistic retail sales datasets. It was determined that localized tree-based approaches performed better in terms of achieving the lowest RMSE values for XGBoost, whereas deep learning approaches were worse in comparison to ensemble models under the specified conditions (Hobor et al., 2026). The result that model choice should be guided by compatibility with the problem rather than model complexity can be validated by the current work with the LSTM.

2.5 Stockout Prediction Research

While demand forecasting has received extensive research attention, the specific problem of stockout prediction as a binary classification task has received comparatively less focus in the literature. Most existing studies treat stockout prevention as a downstream application of demand forecasting rather than as a standalone research problem with its own modeling requirements and evaluation criteria.

Liu et al. (2025) conducted one of the most comprehensive recent studies on machine learning for inventory stockout prediction, utilizing a large dataset from a retailer comprising over 1.6 million stock keeping units to develop classification models aimed at improving stockout detection accuracy. The study found that training a Logistic Regression model directly on the imbalanced dataset produced 99.33 percent accuracy on the test set but with zero true positive stockout detections, demonstrating the fundamental inadequacy of naive classification approaches for imbalanced stockout data (Liu et al., 2025). After applying SMOTE combined with Tomek Links to address class imbalance, the study found that predictive performance improved significantly with feature importance analysis revealing that current inventory levels, short term demand forecasts, and recent sales data were the most influential predictors of stockout events.

The paper Vaclavek et al. (2021) presented a case study where two machine learning systems were developed in a Latin American retailer for detecting the out of stock situation. While the

former utilized only one Random Forest classifier with balanced data, the latter used an ensemble approach where there were six classification algorithms. The implementation of these systems proved successful with the help of transactional data obtained from the manufacturer's information system along with actual audit results (Vaclavek et al., 2021). The experiment thus proves that it is indeed possible to implement machine learning solutions for the detection of out of stock situations in a commercial organization.

2.6 Research Gaps

A careful review of the existing literature reveals several important limitations that collectively point toward a significant gap in the current state of knowledge on retail inventory intelligence. The majority of existing comparative studies on machine learning for retail, including Mitra et al. (2022), Prabu et al. (2025), and Hobor et al. (2026), focus exclusively on demand forecasting as a regression problem, predicting future sales values without extending the analysis to the downstream question of whether those predicted demand levels will exceed available inventory. This creates a fundamental disconnect between what the research produces, sales forecasts and what retailers actually need, actionable stockout risk signals. A retailer who knows that tomorrow's sales will be 5000 units gains limited operational value from that information alone without also knowing whether current stock levels are sufficient to meet that demand.

Furthermore while studies such as Liu et al. (2025) and Vaclavek et al. (2021) have addressed stockout prediction as a classification problem, none of these studies have integrated both demand forecasting and stockout classification within a single unified comparative framework that evaluates the same models across both tasks using consistent data, features, and evaluation protocols. The absence of such a framework makes cross phase model comparisons impossible and leaves practitioners without evidence-based guidance on which models offer the best combined performance across both forecasting and classification dimensions. Additionally, inconsistent use of evaluation metrics across existing studies, with some using only RMSE, others using only F1 Score, and few using both, further complicates the task of drawing reliable cross-study conclusions about relative model performance.

The class imbalance problem in stockout prediction has also been inadequately addressed in a significant portion of existing studies. Several papers that attempt stockout or out of stock detection do not apply any balancing technique before model training, leading to inflated accuracy metrics that do not reflect true model sensitivity to the minority stockout class. This methodological oversight represents a significant limitation that reduces the practical reliability of their reported results. The present study directly addresses all three of these gaps by developing a two phase comparative framework that integrates demand forecasting and stockout classification within a single study, applying consistent methodology and evaluation metrics across all models in both phases, and explicitly handling class imbalance using SMOTE applied exclusively to the training set to prevent data leakage.

CHAPTER 3

RESEARCH METHODOLOGY

3.1 Introduction to Research Methodology

In this chapter, we present the entire methodology employed for conducting this study in order to achieve the objectives presented in Chapter 1. The methodology employed in the study consists of a comparative process involving two phases where the performance of four machine learning algorithms are compared for demand forecasting and stockout classification. This chapter will focus on the research design, data set description, data pre-processing, feature engineering, generation of stockout labels, balancing class imbalance problem, model description, and evaluation metrics used in the study.

3.2 Research Design

The methodology for this study is experimental quantitative research. This kind of research takes place in a well-defined research pipeline, starting from the data acquisition and pre-processing step, to feature extraction and modeling, and finally ending up with model evaluation and comparison in two phases. It is important to state that this research is an empirical study which concludes with experiment results conducted on actual retail sales data.

The two-phase design of this study is illustrated below:

Phase 1 — Demand Forecasting treats the problem as a supervised regression task where models are trained to predict continuous daily sales values. Model performance is evaluated using regression metrics, Root Mean Squared Error and Mean Absolute Error.

Phase 2 — Stockout Classification treats the problem as a supervised binary classification task where models are trained to predict whether a stockout event will occur on a given day. Model performance is evaluated using classification metrics, F1 Score and Area Under the Receiver Operating Characteristic Curve.

Both phases use the same dataset, the same features, and the same train test split, ensuring that all comparisons are fair, consistent, and directly interpretable.

3.3 Dataset Description

This particular study will use the Rossmann Store Sales dataset, which can be found on the Kaggle website. This dataset was first created for the Rossmann Store Sales forecasting challenge that took place on Kaggle in 2015 and has been one of the most popular datasets for forecasting research in the retail industry ever since.

Dataset Overview: The dataset contains historical daily sales records from 1115 Rossmann drug stores operating across Germany. The data spans multiple years of trading activity from

January 2013 to July 2015, providing a rich longitudinal record of retail sales patterns across diverse store types, locations, and promotional conditions.

Primary Files Used:

Two major data files are used in this study. First, the file `train.csv` consists of 1,017,209 entries of daily sale records, each of which is made up of one record per store per day. The file includes information such as store ID, date, sales quantity, number of customers, store opening status, promotion, and holiday flag for each observation. Second, the file `store.csv` consists of 1115 entries of stores, with each entry representing one store. Store information includes store type, store assortment, competition distance, and participation in promotion program.

Key Features:

The dataset contains the following key features that are used in model training:

Store - a unique numerical identifier for each of the 1115 stores in the dataset.

Date - the calendar date of the sales record, from which temporal features including year, month, day, day of week, and week of year are extracted.

Sales - the daily sales turnover in Euros for a given store on a given date. This is the target variable for Phase 1 regression models.

Customers - the number of customers who visited the store on a given day.

Open - a binary indicator showing whether the store was open on a given day, with 1 indicating open and 0 indicating closed.

Promo - a binary indicator showing whether a promotional campaign was running at the store on a given day.

StateHoliday - a categorical variable indicating whether the day was a state holiday, with categories for public holiday, Easter holiday, Christmas, and no holiday.

SchoolHoliday - a binary indicator showing whether the store and date were affected by the closure of public schools.

StoreType - a categorical variable classifying stores into four distinct store models labeled a, b, c, and d, each representing a different retail format.

Assortment - a categorical variable describing the level of product assortment carried by the store, classified as basic, extra, or extended.

CompetitionDistance - the distance in meters to the nearest competing store.

Promo2 - a binary indicator showing whether the store participates in a recurring promotional program.

3.4 Data Preprocessing

Raw data rarely arrives in a form suitable for direct model training. The following preprocessing steps were applied to clean and prepare the dataset for analysis.

- **Data Merging:** The two primary files train.csv and store.csv were merged using the Store column as the common key using a left join operation. This produced a single unified dataframe containing both daily sales records and store level attributes for each observation, resulting in a combined dataset with 1,017,209 rows and 18 columns.
- **Removal of Closed Store Days:** Rows where the Open column had a value of 0 were removed from the dataset. Sales on closed days are always zero and provide no useful signal for demand forecasting or stockout detection. Retaining these rows would introduce noise into the training data and distort model learning. After this step 844,338 rows remained in the dataset.
- **Removal of Zero Sales Records:** Rows where the Sales column had a value of 0 despite the store being recorded as open were also removed. These records represent data quality issues or exceptional circumstances and do not reflect normal trading patterns. This step further refined the dataset to ensure model training was based on meaningful sales observations.
- **Missing Value Treatment:** The CompetitionDistance column contained missing values for stores where no nearby competitor was recorded. These missing values were filled with the median competition distance across all stores, as this represents a reasonable central estimate. All remaining missing values across other columns were filled with zero, as they predominantly corresponded to optional fields that were not applicable for certain store configurations.
- **Date Feature Extraction:** The Date column is transformed from its current string format into the datetime format of python. Five time-related features are created from this column which include the Year, Month, Day, DayOfWeek, and WeekOfYear. Using these features, the model can recognize seasonality, weekly trends, and year-on-year trends without needing to structure time series data.
- **Categorical Encoding:** Three categorical columns required encoding before model training. StateHoliday was encoded as 0 for no holiday, 1 for public holiday, 2 for Easter holiday, and 3 for Christmas. StoreType was encoded as 0, 1, 2, and 3 corresponding to store types a, b, c, and d respectively. Assortment was encoded as 0, 1, and 2 corresponding to assortment levels a, b, and c respectively. This encoding converted all categorical variables into numerical format suitable for machine learning model input.

3.5 Feature Engineering

Feature engineering is the process of selecting and transforming raw data columns into informative input features for model training. The following 13 features were selected for model training based on their relevance to sales prediction and stockout detection:

Store, DayOfWeek, Promo, StateHoliday, SchoolHoliday, StoreType, Assortment, CompetitionDistance, Promo2, Year, Month, Day, and WeekOfYear.

These features were chosen because they capture the key dimensions of retail sales variation-temporal patterns through date features, promotional effects through Promo and Promo2, external demand shocks through holiday indicators, and store characteristics through

StoreType, Assortment, and CompetitionDistance. The Customers column was excluded from the feature set as it represents concurrent information that would not be available at prediction time in a real-world deployment scenario.

3.6 Stockout Label Creation

A critical methodological contribution of this study is the creation of a binary stockout label from the Rossmann dataset, which does not contain actual inventory level data. The stockout label was engineered using a sales drop detection approach based on the following logic, if a store that is confirmed to be open on a given day records sales that fall significantly below its historical average, this is interpreted as a proxy indicator of a stockout event.

Threshold Selection Process: The threshold for defining a significant sales drop was determined through systematic experimentation. The following thresholds were tested and their resulting class distributions were examined:

Threshold	Stockout Days	Percentage
20% of average	146	0.02%
30% of average	1,178	0.14%
40% of average	5,436	0.64%
50% of average	13,956	1.65%
60% of average	33,634	3.98%
70% of average	88,249	10.45%
80% of average	201,014	23.81%

A threshold of 50 percent of the store's historical average sales was selected as the most appropriate threshold. This threshold produced 13,956 stockout days representing 1.65 percent of the dataset, sufficient for model learning while remaining within a realistic range for stockout frequency. Thresholds below 50 percent produced too few stockout cases for reliable model training while thresholds above 60 percent began flagging an unrealistically high proportion of days as stockouts, diluting the meaningfulness of the classification.

The stockout label was created using the following logic:

A day was labeled as a stockout event with a value of 1 if the store was confirmed open and the daily sales were less than 50 percent of that store's historical average sales. All other open trading days were labeled as normal with a value of 0. This binary label served as the target variable for all Phase 2 classification models.

3.7 Handling Class Imbalance Using SMOTE

With only 1.65 percent of observations labeled as stockout events the dataset exhibited severe class imbalance. Training classification models directly on such imbalanced data results in models that achieve deceptively high accuracy by predicting the majority class for all observations, completely failing to detect the minority stockout class.

However, for resolving this issue, the Synthetic Minority Oversampling Technique called SMOTE was utilized to balance the minority class prior to training the classifier on the training data. SMOTE is an algorithm that creates synthetic minority samples by interpolating between existing minority samples and not simply replicating them, thereby increasing the representation of the minority class in the training set.

The application of SMOTE was done only on the training data and no changes were made to the test data in order to avoid any kind of data leakage from the training data to the test data. Using SMOTE, the number of stockouts and normal cases in the training data became 664,424 each, giving a 50:50 ratio.

3.8 Model Descriptions

This study compares four machine learning models across two phases. Each model is described below.

3.8.1 Linear Regression and Logistic Regression

Linear Regression is one of the most straightforward algorithms used for regression purposes within the realm of supervised learning. This algorithm assumes that the output depends linearly on the inputs and seeks to find a linear function which best describes the data set provided for training. In the present work, Linear Regression will be employed as the baseline regression algorithm in Phase 1, allowing comparison of its performance with the more advanced algorithms.

Logistic Regression can be regarded as the classification counterpart of Linear Regression. It uses a logistic function for transforming linear combinations of the inputs into probabilities. In the current research, Logistic Regression will serve as the baseline classification algorithm in Phase 2, providing an opportunity to compare the benefits of ensemble and deep learning approaches.

3.8.2 Random Forest

The Random Forest model is a form of ensemble learning where a large number of decision trees are built in the training phase and the average prediction result from all trees is generated for regression problems, while in classification problems the most voted class is predicted. All trees within the random forest are trained on bootstrapped random samples and only consider a random subset of features in splitting. This method is referred to as feature bagging that results in trees being uncorrelated.

Random Forest is especially suitable for performing retail forecasting since it is capable of dealing with nonlinear relationships, is resistant to outliers, requires no manual hyperparameter adjustments, and offers feature importance measurements that indicate the impact of input variables. In the current work, the Random Forest model is utilized as a regressor during Phase 1 and a classifier during Phase 2 with the class weights equal to 'balanced'.

3.8.3 XGBoost

XGBoost, short for Extreme Gradient Boosting, is a gradient boosting algorithm that creates a series of decision trees sequentially, where each additional tree tries to learn from the mistakes made by all the previously learned models. It differs from the Random Forest

approach, which trains each tree in parallel, in that each tree created by the XGBoost algorithm tries to fix the mistakes made by the ensemble so far.

XGBoost has been the most popular algorithm for retail forecasting competitions and studies because of its superior accuracy, internal regularization mechanisms that avoid overfitting, ability to handle missing data automatically, and efficiency. In Phase 1, XGBoost serves as a regressor, while in Phase 2, it acts as a classifier, with the scale positive weight parameter equal to the ratio of negative to positive examples.

3.8.4 Long Short-Term Memory Network

Long Short Term Memory networks are a specialized type of Recurrent Neural Network designed to capture long term dependencies in sequential data. Unlike standard neural networks that process each input independently, LSTM networks maintain a hidden state that carries information across time steps through gated mechanisms, an input gate, a forget gate, and an output gate, that control what information is stored, discarded, and passed forward.

LSTM networks are theoretically well suited for time series forecasting because retail sales exhibit strong temporal dependencies, sales on a given day are influenced by sales patterns in preceding days, weeks, and months. In this study the LSTM architecture consists of two LSTM layers with 64 and 32 units respectively, two Dropout layers with a rate of 0.2 to prevent overfitting, and two Dense layers for output. In Phase 1 the output layer uses a linear activation function for regression and in Phase 2 it uses a sigmoid activation function for binary classification. All input features were normalized using Min-Max scaling before being fed into the LSTM model, and data was reshaped into a three dimensional format of samples, timesteps, and features as required by the LSTM input specification.

3.9 Evaluation Metrics

Phase 1 — Regression Metrics:

Root Mean Squared Error, commonly referred to as RMSE, represents the square root of the mean of the squared differences between the predicted sales figures and the actual sales figures. The RMSE formula is very harsh on larger errors because of the process of squaring.

Mean Absolute Error, commonly referred to as MAE, refers to the mean of the absolute difference between the predicted sales figures and the actual sales figures. The MAE formula assigns equal weighting to all errors irrespective of size.

Phase 2 - Classification metrics:

F1 Score is the harmonic mean between Precision and Recall. Precision is the ratio of the number of days that the model predicted as stockout days and the actual number of stockout days. Recall is the percentage of stockout days that the model correctly identifies. The F1 Score balances both concerns. It is particularly suited for imbalanced classification problems where maximizing accuracy is deceptive. A higher F1 Score represents better classification performance.

The Area Under the Receiver Operating Characteristic Curve (AUC-ROC) quantifies how well the model can discriminate between stockout and non stockout days across all possible

classification thresholds. A 1.0 AUC-ROC indicates perfect classification and 0.5 indicates random guessing. The larger the AUC-ROC, the better the discrimination ability.

3.10 Experimental Setup

All experiments were done through Google Colab, a free cloud computing-based environment with Jupyter notebook written in Python provided by Google. The reason why Google Colab was chosen as the experimentation platform is its availability of free GPU resources, having all the necessary machine learning libraries pre-installed on it, making experiments shareable and reproducible, and also the elimination of the influence of local computing power on the training of the models, especially LSTM.

The list of Python libraries used in this experiment includes Pandas and NumPy for operations with data; Matplotlib and Seaborn for data visualizations; Scikit-learn for linear regression, logistic regression, random forest algorithms, train test splitting, evaluation metrics; XGBoost for XGBoost models; Tensorflow and Keras for LSTM model creation and training; Imbalanced-learn for SMOTE application.

The training and testing data split happened according to the 80/20 rule. This implies that 80 percent of the data was used to train the models, while the other 20 percent remained untouched until testing. The random state parameter set up as equal to 42 ensures exact reproducibility. In order not to introduce the test set in the sample used for synthetic data generation, the SMOTE algorithm was used on the training set only.

All four models were trained using default hyperparameters with the exception of the number of estimators set to 100 for Random Forest and XGBoost, and the number of epochs set to 10 with a batch size of 256 for LSTM. The primary objective of this study is comparative model evaluation rather than individual model optimization, so hyperparameter tuning was intentionally kept minimal and consistent across models to ensure a fair comparison.

CHAPTER 4

DATA ANALYSIS

4.1 Introduction to the Case

In the retail sector, the business is conducted in an atmosphere of demand uncertainty; it is here that the capability to forecast future sales becomes the key to success and profits or losses. Among the many operational challenges that retailers face, stockouts represent one of the most damaging and yet preventable problems. When a customer walks into a store and finds that a product they intend to purchase is unavailable, the consequences extend far beyond that single lost transaction. Research consistently shows that stockout experiences lead to immediate revenue loss, long term customer attrition, damage to brand reputation, and disruption of supplier relationships. Despite decades of research and significant investment in inventory management systems, stockouts continue to plague retailers worldwide at an alarming rate.

This case study focuses on the application of machine learning models to address the retail stockout problem through a two phase analytical framework. The first phase treats the problem as a demand forecasting task, predicting future sales volumes to understand what level of demand a store can expect on any given day. The second phase extends this analysis by introducing a novel stockout classification layer that translates demand predictions into actionable binary alerts, will this store experience a stockout in the coming days or not. Together these two phases form a unified analytical framework that is more practically valuable than either phase in isolation.

The case is grounded in real world retail data from Rossmann, one of Europe's largest drug store chains operating over 3000 stores across seven European countries. The Rossmann Store Sales dataset provides a rich and realistic foundation for this analysis, capturing the complexity of real retail operations including promotional campaigns, seasonal demand variation, competitive dynamics, and diverse store formats. By applying and comparing four machine learning models across both analytical phases, this case study generates insights that are relevant not only to academic research but also to retail practitioners seeking to implement data driven inventory management solutions.

The importance of this study is due to the fact that it is relevant both theoretically and practically because not only does it provide additional knowledge about the comparative performance of machine learning techniques used in retail analysis, but it also serves as an example of how to develop such a model based on sales data without inventory data available.

4.2 Data Collection: Sources and Approach

4.2.1 Data Source

The main source of primary data used for this study is Rossmann Store Sales Data available on Kaggle site. This data set was first made available by Rossmann in collaboration with Kaggle back in 2015 as part of a forecasting challenge. Over 3500 teams worldwide

competed in the challenge to predict sales of up to 6 weeks ahead at the total of 1115 stores in Germany operated by Rossmann. After the completion of the challenge, the data set stayed open to the public and has been extensively used in academic studies ever since.

4.2.2 Dataset Files and Structure

The dataset consists of multiple files of which two were selected for this study based on their relevance to the research objectives.

The first file `train.csv` is the primary data file containing 1,017,209 rows of historical daily sales records. Each row represents one store on one day and contains the following columns — Store ID, Date, Sales, Customers, Open status, Promo indicator, StateHoliday category, and SchoolHoliday indicator. This file covers the period from January 2013 to July 2015 providing approximately two and a half years of daily trading data across all 1115 stores.

The second file `store.csv` contains store level attribute data with 1115 rows — one per store. It provides information about each store's format and competitive environment including StoreType which classifies stores into four distinct retail formats, Assortment which describes the breadth of product range carried by the store, CompetitionDistance which records the distance in meters to the nearest competitor store, and Promo2 which indicates whether the store participates in a recurring promotional program.

4.2.3 Data Collection Approach

The data was accessed and downloaded from the Kaggle platform using a registered Kaggle account. The downloaded files were uploaded directly to Google Colaboratory for analysis. No web scraping, API calls, or proprietary data access was required. The data collection process was entirely based on publicly available benchmark data, ensuring full reproducibility of the study.

Both datasets were imported in Python using the Pandas library, and then the two datasets were joined together using the Store column as the common field, using the left join technique. This join resulted in a single dataframe, which contained both sales and contextual information at the store level.

4.2.4 Dataset Statistics

After merging the two files, the combined dataset contained the following key statistics:

Table 1 — Dataset Statistics

Statistic	Value
Total rows before preprocessing	1,017,209
Total rows after removing closed days	844,338
Number of stores	1,115
Number of features used	13
Date range	January 2013 to July 2015
Average daily sales per store	Approximately €5,773
Total stockout days identified	13,956
Stockout percentage	1.65%
Training set size	675,470 rows
Test set size	168,868 rows

4.3 Data Analysis

4.3.1 Preprocessing Results

As mentioned in Chapter 3, a series of preprocessing steps were performed on the raw merged dataset which resulted in the following results. Filtering out the close store days led to a decrease in the number of data entries from 1,017,209 to 844,338 by eliminating about 17% of data. The missing values for CompetitionDistance attribute were imputed using the median value of 2,325 meters. Encoding of categorical attributes was carried out for StateHoliday, StoreType, and Assortment attributes changing their categorical labels into numeric. Five additional temporal attributes have been derived from the Date attribute which include Year, Month, Day, DayOfWeek, and WeekOfYear.

4.3.2 Stockout Label Analysis

The stockout label creation process involved systematic experimentation with multiple sales drop thresholds as documented in Chapter 3. The final threshold of 50 percent of each store's historical average sales produced 13,956 stockout days representing 1.65 percent of the cleaned dataset. These stockout days were distributed across multiple stores and time periods with a higher concentration observed during post-holiday periods and mid-week days when sales naturally dip below average levels.

The severe class imbalance with a 98.35 to 1.65 percent split between normal and stockout days necessitated the application of SMOTE before Phase 2 model training. After SMOTE application the training set contained 664,424 observations of each class producing a perfectly balanced 50:50 distribution that enabled classifiers to learn meaningful stockout patterns without being dominated by the majority class.

4.3.3 Class Imbalance and SMOTE

Because the ratio of minority class instances was as low as 1.65 percent, the dataset had an extreme class imbalance. When using imbalanced datasets for training classifiers, the classifiers tend to predict the majority class with very high accuracy while ignoring the minority class instances altogether.

To tackle the problem of imbalanced classes in the data, the Synthetic Minority Over-sampling Technique (SMOTE) was used solely on the training data prior to training the Phase 2 classifiers. In contrast to other methods that use simple replication of minority class samples, SMOTE creates new instances of the minority class through linear interpolation between existing samples.

Table 2 — Class Distribution Before and After SMOTE

Class	Before SMOTE	After SMOTE
Normal Days (0)	664,424	664,424
Stockout Days (1)	11,046	664,424
Total	675,470	1,328,848
Stockout Percentage	1.64%	50.00%

After applying SMOTE both classes contained exactly 664,424 observations producing a perfectly balanced 50:50 distribution. This balanced distribution enabled Phase 2 classifiers to learn genuine stockout patterns without being dominated by the majority class. The test set was kept untouched with its natural class distribution to ensure evaluation metrics reflect real world performance conditions.

4.4 Phase 1 — Demand Forecasting Results

In Phase 1, the performance of four machine learning algorithms was determined based on their capacity to predict sales values. The training of these algorithms was done using 80 percent of the dataset, while the prediction was based on the other 20 percent. The measures used for evaluating their performance were RMSE and MAE.

4.4.1 Phase 1 Results Table

Table 3 — Phase 1 Demand Forecasting Results

Model	RMSE	MAE	Rank
Linear Regression	2805.82	2039.26	4th
Random Forest	922.26	593.05	1st
XGBoost	1200.56	859.80	2nd
LSTM	2520.57	1828.10	3rd

4.4.2 Phase 1 Visual Comparison

The bar charts below present a visual comparison of all four models on both RMSE and MAE metrics for Phase 1.

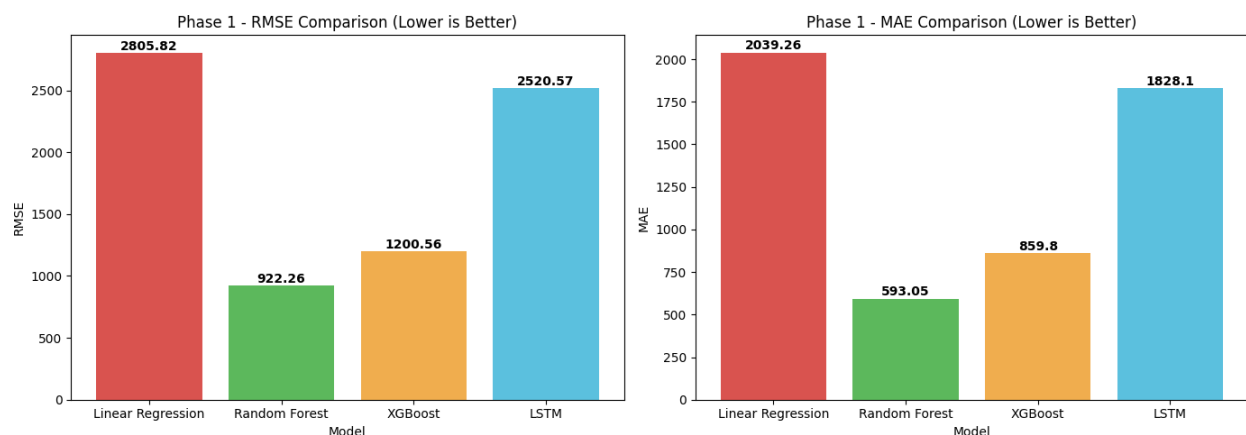


Figure 4.1 — Phase 1 RMSE and MAE Comparison Across Models

The charts clearly show that Random Forest achieved the lowest RMSE of 922.26 and the lowest MAE of 593.05, making it the best demand forecasting model. Linear Regression recorded the worst performance on both metrics, confirming that retail sales data is highly non linear. LSTM performed closer to Linear Regression than to the ensemble methods highlighting the impact of the absence of proper sequential data structuring.

4.4.3 Model by Model Phase 1 Observations

Linear Regression — RMSE 2805.82, MAE 2039.26: As anticipated, Linear Regression performed poorly in Phase 1. The reason behind poor performance was that Linear Regression cannot predict any non-linear association between promotion, season, and store type. Errors in predictions were consistently high, with average errors above 2000 Euros per day. It is clear that Linear Regression is not an appropriate model for predicting in a retail setting.

Random Forest — RMSE 922.26, MAE 593.05: Random Forest achieved the best Phase 1 performance with a 67 percent improvement in RMSE over Linear Regression. The ensemble nature of the model combining predictions from 100 independent decision trees effectively smoothed individual tree errors and produced stable and accurate predictions. Average prediction errors of approximately 593 Euros per day are substantially lower than any other model in this study and consistent with findings from Prabu et al. (2025) who also identified Random Forest as the best performing model on the Rossmann dataset.

XGBoost — RMSE 1200.56, MAE 859.80: XGBoost recorded the second best performance with results substantially better than Linear Regression and LSTM but unable to match Random Forest. The sequential error correction mechanism of XGBoost captured complex patterns in the data effectively however, the absence of hyperparameter tuning may have limited its full potential. This finding contrasts with several prior studies including the RIT thesis which found XGBoost to outperform Random Forest on the same dataset, an interesting divergence that is discussed further in Chapter 5.

LSTM — RMSE 2520.57, MAE 1828.10: LSTM performed significantly worse than both ensemble methods, recording results close to Linear Regression despite being a far more complex architecture. This result is attributed to the absence of proper chronological per-store sequential data structuring. The training data was split randomly, which broke the temporal sequence that LSTM depends on to leverage its gated memory mechanisms. When temporal sequences are disrupted, LSTM cannot learn meaningful temporal dependencies, and its performance degrades substantially.

4.5 Phase 2 — Stockout Classification Results

In Phase 2, we assessed four different classification algorithms on their predictive power for identifying binary stockouts. The models used in Phase 2 were all built on the SMOTE-balance training dataset and were validated against the original test data using F1 Score and AUC-ROC measures.

4.5.1 Phase 2 Results Table

Table 4 — Phase 2 Stockout Classification Results

Model	F1 Score	AUC-ROC	Rank
Logistic Regression	0.1134	0.8927	4th
Random Forest	0.6338	0.9722	1st
XGBoost	0.4897	0.9757	2nd
LSTM	0.2829	0.9402	3rd

4.5.2 Phase 2 Visual Comparison

The bar charts below present a visual comparison of all four models on both F1 Score and AUC-ROC metrics for Phase 2.

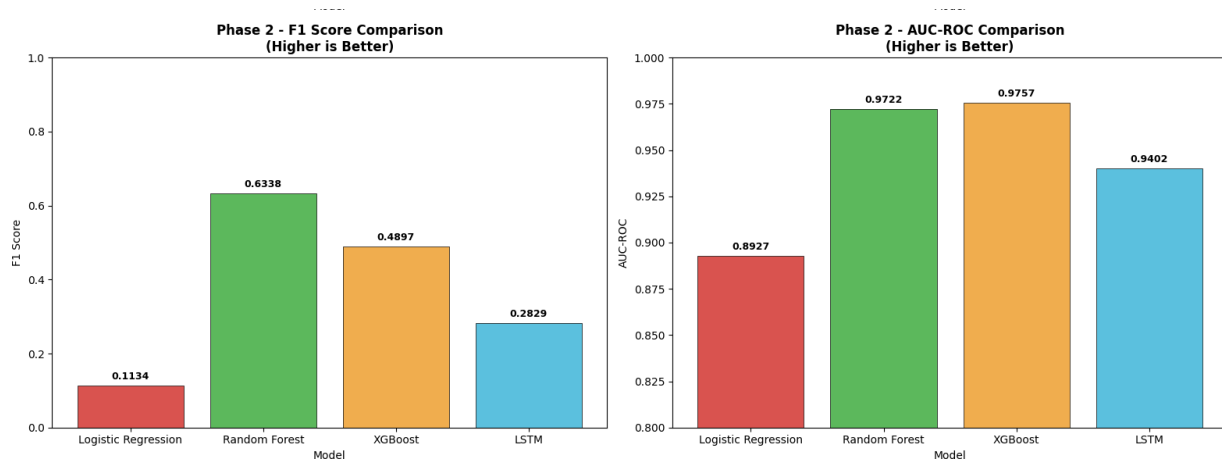


Figure 4.2 — Phase 2 F1 Score and AUC-ROC Comparison Across Models

The F1 Score chart shows Random Forest as the clear winner at 0.6338 with a substantial gap above XGBoost at 0.4897. The AUC-ROC chart reveals an interesting finding, XGBoost marginally outperforms Random Forest at 0.9757 versus 0.9722, suggesting XGBoost is slightly better at ranking stockout risk across probability thresholds, even though it is less precise at making direct binary classifications.

4.5.3 Model by Model Phase 2 Observations

Logistic Regression — F1 0.1134, AUC-ROC 0.8927: Logistic Regression recorded the weakest Phase 2 performance with an extremely low F1 Score of 0.1134 despite a relatively decent AUC-ROC of 0.8927. This gap between the two metrics is characteristic of models that can rank risk probabilities in aggregate but fail to make precise binary classifications at the individual observation level. The linear decision boundary of Logistic Regression is fundamentally inadequate for the complex multidimensional feature combinations that distinguish stockout days from normal trading days.

Random Forest — F1 0.6338, AUC-ROC 0.9722: Random Forest achieved the best Phase 2 performance with an F1 Score of 0.6338 and AUC-ROC of 0.9722. Its ability to construct multiple independent decision boundaries through its ensemble of 100 trees gave it a significant advantage in identifying the complex combination of features that characterize stockout days. The class weight balanced parameter further enhanced sensitivity to the minority stockout class beyond what SMOTE alone provided making it the most reliable model for direct stockout alerting.

XGBoost — F1 0.4897, AUC-ROC 0.9757: XGBoost recorded an F1 Score of 0.4897 and the highest AUC-ROC of 0.9757 in Phase 2. This reveals an important distinction, XGBoost

is marginally better at ranking days by stockout risk probability across all thresholds but less precise than Random Forest at making direct binary classifications at the default 0.5 threshold. This makes XGBoost more suitable for continuous risk scoring applications while Random Forest is better suited for direct stockout alerting systems where a binary yes or no decision is required.

LSTM — F1 0.2829, AUC-ROC 0.9402: LSTM recorded an F1 Score of 0.2829 and AUC-ROC of 0.9402 in Phase 2, better than Logistic Regression but significantly behind both ensemble methods. The same sequential structuring limitation that constrained Phase 1 performance also impacted Phase 2 classification. The complexity of the LSTM architecture may have also contributed to some overfitting on the balanced training set that did not generalize well to the naturally imbalanced test set.

4.6 Comparative Observations Across Both Phases

When considering the four models as a whole through cross-phase comparison, there are many interesting observations to be made that transcend the findings from each phase and give a more comprehensive insight into each model's behavior within different machine learning applications.

4.6.1 Consistency of Performance Ranking

One of the most notable observations from this study is the consistency of the performance ranking across both phases. Random Forest ranked first in both Phase 1 and Phase 2, XGBoost ranked second in both phases, LSTM ranked third in both phases, and Linear or Logistic Regression ranked last in both phases. This consistency across two fundamentally different task types — regression and classification — is a strong indicator of the robustness and reliability of the performance hierarchy observed in this study. It suggests that the relative strengths and weaknesses of these models are not task specific but reflect more fundamental differences in their ability to capture the underlying patterns in retail sales data.

This consistency also has important practical implications. A retailer implementing this framework does not need to select different models for different tasks — Random Forest delivers strong performance across both demand forecasting and stockout classification making it a single reliable solution for both inventory intelligence needs simultaneously.

4.6.2 The RMSE to F1 Score Relationship

An interesting pattern emerges when the Phase 1 RMSE values and Phase 2 F1 Scores are compared side by side across all four models. Models that performed well in Phase 1 forecasting also tended to perform well in Phase 2 classification suggesting that the ability to capture underlying demand patterns accurately is a prerequisite for reliable stockout detection. Random Forest which best captured the demand signal in Phase 1 with an RMSE of 922.26 also best identified stockout events in Phase 2 with an F1 Score of 0.6338. Conversely Linear Regression which produced the worst demand forecasts with an RMSE of 2805.82 also produced the worst stockout classifications with an F1 Score of 0.1134 through its classification equivalent Logistic Regression.

This relationship between forecasting accuracy and classification performance validates the two phase design of this study. The quality of demand signal capture in Phase 1 appears to be a strong predictor of stockout detection capability in Phase 2 suggesting that improving demand forecasting accuracy is not just an end in itself but also a pathway to improving stockout prediction reliability.

4.6.3 The AUC-ROC versus F1 Score Divergence

A particularly insightful finding that emerges from comparing Phase 2 metrics across models is the divergence between AUC-ROC and F1 Score for XGBoost. While XGBoost achieved the highest AUC-ROC of 0.9757 it recorded a notably lower F1 Score of 0.4897 compared to Random Forest which achieved an AUC-ROC of 0.9722 and F1 Score of 0.6338. This divergence reveals that XGBoost produces well-calibrated probability scores across the full range of thresholds but struggles to make precise binary classifications at the default 0.5 threshold.

This conclusion implies a particular practical recommendation: the decision regarding the selection of either Random Forest or XGBoost for Phase 2 must be based on the operational context in which the models will be applied rather than on the basis of a single performance indicator. Namely, if a binary yes or no stockout warning is needed, then Random Forest must be selected due to its high F1 Score. However, in case of need for probability estimation and ranking of possible cases of stockouts, then XGBoost would be preferable since it yields higher AUC-ROC.

4.6.4 Impact of SMOTE on Classification Results

The application of SMOTE had a transformative impact on Phase 2 classification results. Before SMOTE was applied the training set contained only 11,046 stockout cases against 664,424 normal cases, a ratio of approximately 1 to 60. Without balancing this extreme imbalance all classifiers would have defaulted to predicting no stockout for every observation achieving superficially high accuracy of approximately 98 percent but detecting zero actual stockout events. After SMOTE balanced the training data to a 50:50 distribution all four classifiers were able to learn meaningful stockout patterns resulting in F1 Scores ranging from 0.1134 for Logistic Regression to 0.6338 for Random Forest.

The fact that even after SMOTE balancing Logistic Regression still achieved a very low F1 Score of 0.1134 demonstrates that class balancing alone is not sufficient for reliable stockout classification, the model must also have sufficient representational capacity to capture the non linear feature interactions that characterize stockout events. This finding reinforces the conclusion that ensemble methods are necessary for effective stockout detection regardless of the balancing technique applied.

4.6.5 Overall Performance Summary

The table below provides a comprehensive cross phase summary of all model performances across all four evaluation metrics providing a single unified reference for the complete results of this study.

Table 5 — Cross Phase Performance Summary

Model	Phase 1 RMSE	Phase 1 MAE	Phase 2 F1	Phase 2 AUC- ROC	Phase 1 Rank	Phase 2 Rank	Overall Rank
Linear	/	2805.82	0.1134	0.8927	4th	4th	4th
Logistic Regression							
Random Forest	922.26	593.05	0.6338	0.9722	1st	1st	1st
XGBoost	1200.56	859.80	0.4897	0.9757	2nd	2nd	2nd
LSTM	2520.57	1828.10	0.2829	0.9402	3rd	3rd	3rd

The cross phase summary confirms that the performance ranking is perfectly consistent across both phases with Random Forest as the overall winner followed by XGBoost, LSTM, and Linear or Logistic Regression in that order.

CHAPTER 5:

DISCUSSION AND RECOMMENDATIONS

5.1 Introduction

In this chapter, the main results will be presented based on the analysis done in chapter four, after which practical recommendations to retailers and academics will be provided before a fair admission is made about the limitations of the current research. This chapter seeks to convert the statistical output of a two-phased comparative analysis into something practical.

5.2 Key Findings

The analysis conducted across both phases of this study produced the following key findings:

Random Forest is the Overall Best Performing Model

In both phases of this study Random Forest was the clear winner. In Phase 1, it achieved the lowest RMSE of 922.26 and MAE of 593.05 outperforming XGBoost, LSTM and Linear Regression by significant margins. In Phase 2, it again achieved the best classification performance with an F1 Score of 0.6338 and AUC-ROC of 0.9722, making it the most reliable model for both demand forecasting and stockout detection. It is the recommended model for unified retail inventory intelligence applications considering its overall superiority over the four evaluation metrics and two analytical phases. This result is similar to Prabu et al. (2025) who also found Random Forest to be the best performing model on the Rossmann dataset. This adds to the credibility of this finding.

XGBoost and Random Forest Serve Complementary Roles

While Random Forest achieved the highest F1 Score of 0.6338 in Phase 2 XGBoost achieved the highest AUC-ROC of 0.9757. This distinction reveals that the two models serve different but complementary purposes in retail inventory management. Random Forest is more precise at making direct binary stockout classifications making it better suited for direct stockout alerting systems. XGBoost on the other hand is better at ranking days by their stockout risk probability across all thresholds making it more suitable for continuous risk scoring applications where a ranked list of high risk store-day combinations is needed. A combined deployment strategy using both models together would provide retailers with both precision and flexibility in their inventory management decision making.

LSTM Underperforms Without Proper Sequential Data Structuring

LSTM recorded poor performance in both phases with an RMSE of 2520.57 in Phase 1 and F1 Score of 0.2829 in Phase 2 performing only marginally better than the simple baseline models. This finding challenges the widespread assumption that deep learning models automatically outperform traditional machine learning approaches in retail forecasting contexts. The underperformance is attributed to the absence of proper chronological per store sequential data structuring a fundamental prerequisite for LSTM to activate its gated memory mechanisms. When data is not organized in proper temporal sequences per store LSTM cannot learn meaningful temporal dependencies and its performance advantage over simpler models disappears entirely. This finding is consistent with Hobor et al. (2026) who also found

tree based methods outperforming deep learning architectures under real world retail constraints.

Linear Models Are Inadequate for Retail Data

Both Linear Regression and Logistic Regression recorded the worst performance in their respective phases. Linear Regression achieved an RMSE of 2805.82 and MAE of 2039.26 in Phase 1 while Logistic Regression achieved an F1 Score of only 0.1134 in Phase 2 despite a relatively decent AUC-ROC of 0.8927. These results confirm that retail sales data exhibits strong non linear patterns and complex feature interactions that simple linear models fundamentally cannot capture. The linear decision boundaries of these models are inadequate for the multi dimensional feature combinations that drive retail sales variation and stockout events. Their inclusion in this study successfully establishes a clear performance lower bound that justifies the use of more sophisticated ensemble and deep learning approaches.

Class Imbalance Must Be Explicitly Addressed in Stockout Prediction

With only 1.65 percent of observations labeled as stockout events the dataset exhibited severe class imbalance. The application of SMOTE was essential for producing any meaningful Phase 2 classification performance. Before SMOTE was applied the training set contained only 11,046 stockout cases against 664,424 normal cases. After SMOTE both classes were balanced at 664,424 observations each enabling classifiers to learn genuine stockout patterns. Without SMOTE all classifiers would have effectively predicted no stockout for every observation achieving superficially high accuracy but detecting zero actual stockout events. This finding validates the critical importance of addressing class imbalance before training any stockout classification model and is consistent with Liu et al. (2025) who demonstrated similar improvements in stockout detection accuracy after applying resampling techniques.

The Two Phase Framework Adds Genuine Practical Value

However, with the addition of the stockout classification layer on top of demand forecasting, there was an entire approach formed that proved to be significantly more beneficial compared to the two stages individually. Phase 1, while extremely useful in informing the retailer of the expected amount of product sales tomorrow, provides limited value in the context of inventory management. Phase 2 builds upon Phase 1 and offers an additional layer that makes the data directly actionable will this retail outlet experience stockouts or not. Together, these two phases provide a link between forecasting research and retailing practice, thus forming a practical approach that allows retailers to implement proactive inventory management techniques relying solely on their past sales performance.

5.3 Recommendations

Based on the findings of this study the following recommendations are made for retail practitioners and future researchers:

Adopt Random Forest as the Primary Model

Retail organizations that want to use machine learning inventory management systems must consider Random Forest as their first model. The reason is that the model is dependable and can be used to forecast the demand and classify the stockouts due to its ability to provide interpretation from feature importance, being robust to noisy data, and having fewer computational resources required.

Unlike LSTM which requires significant data infrastructure investment and unlike XGBoost which requires careful threshold calibration for binary classification, Random Forest delivers strong out of the box performance with minimal tuning making it accessible even to retailers with limited data science resources.

Use XGBoost for Risk Scoring Applications

In applications where a continuous stockout risk score is more useful than a binary alert such as prioritizing store visits by field replenishment teams or ranking products for urgent restocking, XGBoost should be used alongside Random Forest due to its superior AUC-ROC of 0.9757. A combined deployment approach where Random Forest generates binary stockout alerts and XGBoost generates continuous risk scores would provide retail inventory managers with both the precision of binary decision making and the flexibility of probabilistic risk ranking, a more complete and powerful inventory intelligence system than either model could deliver alone.

Invest in Sequential Data Infrastructure Before Deploying LSTM

Organizations wishing to leverage LSTM for retail forecasting and stockout prediction must first invest in building proper sequential data pipelines that organize historical sales records chronologically per store and per product before feeding them into the model. Without this infrastructure LSTM will not deliver the performance improvements that its theoretical design promises and simpler ensemble methods will consistently outperform it at a fraction of the implementation complexity. The additional investment in sequential data organization is justified only for large retailers with significant technical resources and large volumes of per store historical data spanning several years.

Always Address Class Imbalance in Stockout Prediction

Any retail analytics team implementing stockout classification models must address the class imbalance problem before model training using resampling techniques such as SMOTE. Stockout events are by their nature rare occurrences representing a very small fraction of total trading days. Application of classifiers in imbalanced stockout data without tackling the problem is bound to yield classifiers that seem to perform exceptionally well as a whole but actually do not detect any stockout cases, hence rendering them entirely useless for their intended use. SMOTE is supposed to be implemented on the training data set only and not the testing data set.

Adopt the Two Phase Framework for Inventory Intelligence

The retail sector needs to transcend beyond traditional demand forecasting techniques and embrace the two-phase technique showcased in this research. The first phase of demand forecasting is instrumental in gaining an understanding of what is expected regarding future demand trends, while the second phase of stock-out classification is essential in converting the forecasted demand trends into practical inventory warnings. Collectively, the two-phased

approach offers a more effective and relevant inventory management tool than would be possible by employing just one phase alone. The notable point is that the two phased process does not need any real-time information on inventory levels; rather, it relies entirely on sales records.

Extend the Framework to Indian Retail Contexts

Future research should apply and validate the two phase framework developed in this study on Indian retail datasets. The Rossmann dataset reflects German retail demand patterns characterized by European seasonal cycles, promotional practices, and consumer behavior that may differ significantly from Indian retail contexts. Indian retail is characterized by strong festival driven demand spikes during Diwali, Navratri, and other cultural events, diverse regional consumer preferences, and a rapidly growing organized retail sector. Developing and validating stockout prediction models specifically calibrated for Indian retail conditions would represent a significant and practically valuable contribution to an underserved area of research.

5.4 Limitations of the Study

However, despite the valuable contributions provided by this research, there are certain issues that need to be taken into account:

Proxy Stockout Label

The main problem with this research is the usage of the proxy stockout label based on the sales drop patterns, as opposed to the actual inventory levels. The Rossmann dataset does not contain information about stock levels, replenishment orders, or confirmed stockout events. The assumption that a sales drop below 50 percent of a store's historical average on an open trading day indicates a stockout is a reasonable approximation but it is not a definitive measure. Some observed sales drops may be caused by other factors such as reduced customer footfall due to local events, temporary supply chain disruptions, or unusual weather conditions that are not captured in the dataset. Future studies with access to actual inventory level data would produce more accurate and reliable stockout labels and enable a more rigorous evaluation of model performance on genuine confirmed stockout events.

Absence of Hyperparameter Tuning

All models used in this study were trained on default or slightly tweaked hyperparameters to ensure consistency across the board during performance comparison. The performance of individual models can significantly increase when systematic methods such as grid search cross-validation or Bayesian optimization are applied to fine-tune the hyperparameters. In summary, the ranking obtained from this study is a result of performance under default settings and might not be indicative of the optimal performance of any of the models. Future research is expected to adopt the approach of optimizing the hyperparameters before establishing the relative performance of the models.

LSTM Sequential Structuring

The LSTM models in this study were not trained on properly sequenced chronological per store time series data. The use of a random train test split disrupted the temporal sequences that LSTM depends on to leverage its gated memory mechanisms. This methodological limitation likely contributed significantly to LSTM's underperformance relative to its theoretical potential. A future study implementing proper chronological per store data sequencing with appropriate look back windows would provide a more accurate and fair assessment of LSTM's capabilities for retail forecasting and stockout classification and may produce substantially different performance rankings.

European Retail Context

The Rossmann dataset is derived from retail operations in Germany and reflects demand patterns, holiday schedules, promotional practices, and consumer behavior specific to European retail. The findings and performance rankings produced by this study may not generalize directly to retail contexts in other regions particularly India where festival driven demand spikes, diverse regional consumer preferences, and different competitive dynamics create distinct forecasting and inventory management challenges. Researchers and practitioners in non-European retail contexts should validate the framework on locally relevant datasets before drawing conclusions about model suitability for their specific retail environment.

Static Dataset

The dataset used in this study covers a fixed historical period from January 2013 to July 2015. The models trained in this study are static meaning they do not update or adapt as new sales data becomes available over time. In a real world retail deployment scenario models would need to be retrained periodically on updated data or implemented as online learning systems that continuously incorporate new observations to maintain forecast accuracy as demand patterns evolve. The static nature of this study's models means that their performance in a live deployment scenario may degrade over time as the retail environment changes.

Single Dataset Evaluation

This study's results have been drawn from only one database belonging to one retail chain with operations in one market. Despite the recognition and wide acceptance of the Rossmann Store Sales database as a standard in retail forecasting studies, the conclusions drawn in this research cannot be generalized for all other retail settings, types of products or stores. Future studies must examine the validity of the proposed two-phase framework using databases from other retail chains dealing in various products in other geographical markets in order to verify the generalizability of these results.

No Real Time Implementation

This study presents an offline batch prediction framework rather than a real time prediction system integrated with live inventory management and point of sale data streams. In actual retail operations stockout prediction would need to be integrated with live inventory tracking systems, real time sales data feeds, and automated replenishment ordering platforms to deliver full operational value. The implementation of such real time integration is beyond the scope of this study but represents an important and necessary direction for future applied

research that would bring the two phase framework closer to a production ready retail intelligence solution.

CHAPTER 6:

CONCLUSION

6.1 Summary of the Study

This paper seeks to bridge the gap in the current literature in regard to retail analytics through a proposed comparative framework that integrates the task of retail demand forecasting and stockouts prediction into one unified process. While many works have been done in regard to retail demand forecasting using machine learning techniques, there is no comprehensive study that tackles stockout prediction as an independent task in addition to comparing different methods used in demand forecasting against each other.

This experiment involved the use of the Rossmann Store Sales dataset containing historical data of daily sales in 1115 retail stores in Germany. Four ML algorithms have been tested against each other in two stages where Linear Regression and Logistic Regression were used as baselines, the Random Forest algorithm as an ensemble learning algorithm, XGBoost as a gradient boosting algorithm, and the Long Short-Term Memory networks for deep learning. Stage 1 involved the comparison between ML algorithms concerning demand forecasting by using RMSE and MAE measures. Stage 2 involved comparing ML algorithms based on stockout prediction using F1 Score and AUC-ROC as classification measures. The new binary stockout label was constructed through systematic threshold selection with SMOTE being used to balance the highly imbalanced dataset at 1.65%.

6.2 Summary of Key Findings

The experimental results generated from this study yielded many critical findings that add value to both academia and real-world applications in retail analytics.

The first and foremost finding is that Random Forest was the best-performing model throughout the experiment in both phases. Specifically, in Phase 1, the model had the least RMSE of 922.26 and MAE of 593.05, outperforming XGBoost (RMSE = 1200.56), LSTM (RMSE = 2520.57), and Linear Regression (RMSE = 2805.82). In Phase 2, Random Forest performed the best among other models in classification with an F1 Score of 0.6338 and AUC-ROC of 0.9722, followed by XGBoost (F1 Score = 0.4897), LSTM (F1 Score = 0.2829), and Logistic Regression (F1 Score = 0.1134).

A particularly noteworthy finding is that XGBoost achieved the highest AUC-ROC of 0.9757 in Phase 2 despite recording a lower F1 Score than Random Forest. This reveals that XGBoost and Random Forest serve complementary roles in stockout prediction — XGBoost is superior at ranking stockout risk across probability thresholds while Random Forest is more precise at making direct binary stockout classifications. This distinction has important practical implications for how retailers choose to deploy these models in real world inventory management systems.

The inability of LSTM to outperform the baseline in both phases is yet another important discovery. Although this model is highly recommended for time-series prediction, thanks to the gate-based approach for storing the information, the model did not prove to be superior in this study. The explanation behind such poor performance lies in the fact that there was no

chronologically sequenced per store data. It is the first step necessary in order to take full advantage of the LSTM structure. The result of the experiment disproves the common belief among scholars that any deep learning model outperforms a simple machine learning one in retail forecasting scenarios.

The underperformance of both Linear Regression and Logistic Regression in their corresponding phases supports the idea that retail sales are strongly characterized by non-linear behavior, and thus more advanced methods are needed in order to capture this complexity.

6.3 Research Contributions

This work has made three unique and important contributions to the literature in retail analytics and machine learning studies.

Methodological Contribution

The first methodological contribution of this paper is its unique two-phase comparative framework for integrating the processes of demand forecasting and stockout classification into one study with the same data, features, models, and evaluation criteria for both phases. This model is reproducible and scalable, enabling future scholars to utilize it for evaluating novel models, data sets, and retail environments. The structured procedure used to generate stockout labels by testing various thresholds for a decrease in sales volume (seven thresholds were tested), and then choosing the best one through class distribution analysis, is an innovative and scientifically sound method for developing proxy stockout labels when there are no inventory levels available.

Empirical Contribution

The study produces a comprehensive set of experimental results comparing four machine learning models across two analytical phases on the widely used Rossmann benchmark dataset. The finding that Random Forest outperforms XGBoost in both phases on this dataset represents a noteworthy divergence from several prior studies that identified XGBoost as the superior model for Rossmann based retail forecasting. This divergence contributes a valuable counter-example to the existing literature and suggests that the relative performance of these two leading ensemble methods is more context dependent and methodology sensitive than previously assumed. The finding that LSTM underperforms traditional ensemble methods under standard train test split conditions adds a practically important cautionary insight to the growing body of literature on deep learning for retail forecasting.

Practical Contribution

The stockout label engineering methodology developed in this study — using systematic sales drop threshold experimentation to create a proxy stockout indicator from sales data alone without requiring real time inventory records — provides a practical approach that retailers can implement immediately using their existing historical sales data. This significantly lowers the barrier to entry for stockout prediction making the framework accessible to a wider range of retailers including small and medium sized businesses that may not have sophisticated inventory tracking systems or data science infrastructure. The evidence

based model selection guidance produced by this study provides retail analytics teams with a clear and justified starting point for their own inventory intelligence implementations.

6.4 Implications for Retail Practice

The findings of this study carry several important implications for retail practitioners seeking to implement data driven inventory management solutions in real world settings.

The study demonstrates that meaningful stockout prediction is achievable using only historical sales data without requiring access to real time inventory levels, complex supply chain data, or proprietary data sources. This finding significantly lowers the barrier to entry for retailers wishing to implement predictive inventory management systems. A retail chain that has been collecting daily point of sale data for two or more years has sufficient data to implement the two phase framework developed in this study and begin generating actionable stockout risk signals using freely available open source machine learning tools.

Moreover, the analysis highlights the real-world significance of choosing the best model in the retail environment. All machine learning models do not have equal relevance in solving analytical problems in retail settings, and it is the selection of the model which directly affects the outcome of inventory decisions based on its forecasts. The outperformance of the random forest algorithm in both phases implies that it is the optimal point to begin retail inventory analytics projects. Thanks to its impressive predictive performance, ability to produce results based on feature importance, resistance to data-related errors, and simplicity, it becomes easily implementable even for small regional chains as well as large national companies.

The complementary roles of Random Forest and XGBoost identified in this study have direct implications for how retail organizations structure their inventory management workflows. In a practical deployment scenario Random Forest generated binary stockout alerts could be used to trigger automated replenishment orders or alert store managers to check and top up specific product categories. XGBoost generated risk scores could simultaneously be used to prioritize which stores and product categories require the most urgent attention from field replenishment teams — ranking all store-day combinations by their predicted stockout risk probability to optimize the allocation of limited replenishment resources across a large retail network.

The study also highlights an important practical lesson about the limitations of deep learning in retail contexts. Organizations that are considering significant investment in LSTM or other recurrent neural network architectures for retail forecasting should carefully evaluate whether their data infrastructure is capable of supporting the sequential data organization that these models require before committing to the implementation effort. In most scenarios, machine learning models like Random Forests and XGBoosts will perform just as well or better than neural networks, with significantly reduced development efforts and lower costs, a point that holds key strategic value for retail data science practitioners constrained by deadlines and budgets.

6.5 Directions for Future Research

Although this research makes valuable contributions to retail analytics, there are some areas for future studies, which are outlined as a result of analyzing the limitations and results of this research in Chapter 5.

The most significant area for further research is related to testing the two-stage framework by utilizing datasets with real values of inventory levels and not stockout labels, which are estimated from sales. Access to real inventory records, including stock on hand, replenishment orders, and confirmed stockout events would enable the creation of precise and reliable stockout labels that directly reflect actual inventory depletion events rather than inferred sales anomalies. This would produce a more rigorous and realistic evaluation of model performance and provide a stronger empirical foundation for the stockout classification layer of the framework.

The underperformance of LSTM in this study suggests a clear and important direction for future work. Future studies should implement proper chronological per store sequential data structuring with carefully designed look back windows, for example training LSTM on sequences of the previous 30 days of sales history for each store and evaluate whether this addresses the performance gap between LSTM and ensemble methods observed in the present study. Future studies should also explore more advanced deep learning architectures that have shown promise in recent retail forecasting literature including Temporal Fusion Transformers, N-BEATS, and attention based LSTM models with positional encoding that may be better equipped to handle the complex multi-store multi-product structure of retail sales data.

The framework developed in this study should be extended to incorporate additional external data sources that are known to influence retail demand but were not available in the Rossmann dataset. Weather data is a particularly valuable external signal for retail forecasting as temperature, rainfall, and extreme weather events have well documented effects on consumer shopping behavior and product demand patterns. Social media sentiment data, local event calendars, competitor pricing information, and macroeconomic indicators such as consumer confidence indices and unemployment rates could also enrich the feature set and improve both forecasting accuracy and stockout detection reliability.

Future research should explore the integration of the two-phase framework with real time inventory management systems and automated replenishment ordering platforms. Moving from the offline batch prediction approach demonstrated in this study to a real time streaming prediction system that continuously updates stockout risk assessments as new sales data arrives would substantially increase the operational value of the framework. Such a system could automatically trigger replenishment orders when a store's stockout risk score exceeds a defined threshold reducing the reliance on manual inventory monitoring and enabling truly proactive inventory management at scale.

The application of the two phase framework to diverse retail datasets from different geographic regions, retail formats, and product categories represents another important direction for future research. Validating the framework on datasets from grocery retail, fashion retail, consumer electronics, and pharmaceutical retail would help establish the generalizability of the findings and determine whether the performance hierarchy observed in this study — Random Forest outperforming XGBoost outperforming LSTM outperforming

baseline linear models — holds consistently across different retail contexts or whether certain retail formats and product categories favor different model architectures.

In addition, future work could examine the use of reinforcement learning techniques to address the retail stockout avoidance issue. In this study and most studies that have been published so far, stockout forecasting is treated as a supervised learning task, which involves using historical data to predict future occurrences. However, reinforcement learning provides a completely different strategy for solving this problem in which an artificial intelligent agent learns how to make optimal stockout decisions through interactions with its surroundings. The use of reinforcement learning-based inventory management systems can provide new insights into retail inventory management that cannot be discovered using machine learning methods.

6.6 Closing Remarks

The retail industry stands at an inflection point where the convergence of large scale historical sales data, accessible machine learning tools, and growing awareness of the financial consequences of inventory inefficiency creates an unprecedented opportunity to transform inventory management from a reactive operational function into a proactive intelligence driven capability. Stockouts, the silent revenue killers that cost the global retail industry over a trillion dollars every year, are not inevitable. They are predictable, and where they are predictable they are preventable.

In essence, this study proves that it is possible to make significant strides towards achieving this goal by utilizing freely available data, machine learning software and algorithms, and no proprietary information or expensive infrastructure required to do so. The development and testing of the two phase approach to inventory analysis, stockout prediction and prevention outlined in this study represents one such concrete step towards closing the gap between research on retail demand forecasting and its application within retail inventory management.

First, the consistent outperformance of Random Forest, a model that has been around for two decades, over other more complicated architectures, including LSTM, both in terms of predicting demand and in predicting stockouts proves a valuable insight for practitioners as much as it does for theorists. While architectural complexity can be an asset in many applications, it does not have to be. The key takeaway here is that the solution to any particular problem need not be a complicated one; the optimal solution is the one that is most applicable to solving said problem in a given set of conditions.

It is hoped that the methodology, findings, and recommendations presented in this study provide a useful and actionable foundation for future research and practical implementation efforts in retail inventory intelligence. The two phase framework introduced here is not an endpoint but a starting point, a foundation upon which future researchers can build more sophisticated, more data rich, and more contextually specific retail inventory prediction systems that bring us closer to the goal of a retail world where stockouts are the exception rather than the rule.

REFERENCES

- Aburto, L., & Weber, R. (2007). Improved supply chain management based on hybrid demand forecasts. *Applied Soft Computing*, 7(1), 136–144. <https://doi.org/10.1016/j.asoc.2005.06.001>
- Ahmadov, T., & Helo, P. (2023). Deep neural networks for retail demand forecasting: A comparative study. *International Journal of Production Economics*, 256, 108714. <https://doi.org/10.1016/j.ijpe.2022.108714>
- Castro Moraes, G., & Bandeira-de-Mello, R. (2024). Hybrid convolutional long short-term memory models for sales forecasting in retail. *Journal of Forecasting*, 43(2), 412–431. <https://doi.org/10.1002/for.3073>
- Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–794. <https://doi.org/10.1145/2939672.2939785>
- Cukierski, W. (2015). *Rossmann store sales* [Dataset]. Kaggle. <https://www.kaggle.com/competitions/rossmann-store-sales/data>
- Hobor, L., Petric, A., & Šmuc, T. (2026). Comparative analysis of modern machine learning models for retail sales forecasting. *arXiv preprint arXiv:2506.05941*. <https://arxiv.org/abs/2506.05941>
- Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, 9(8), 1735–1780. <https://doi.org/10.1162/neco.1997.9.8.1735>
- Kourentzes, N., Petropoulos, F., & Trapero, J. R. (2014). Improving forecasting by estimating time series structural components across multiple frequencies. *International Journal of Forecasting*, 30(2), 291–302. <https://doi.org/10.1016/j.ijforecast.2013.09.006>
- Mitra, A., Jain, A., Kishore, A., & Kumar, P. (2023). A comparative study for machine learning models in retail demand forecasting. In S. Bhattacharyya, J. S. Banerjee, & M. Köppen (Eds.), *Human-Centric Smart Computing. Smart Innovation, Systems and Technologies* (Vol. 316, pp. 199–210). Springer. https://doi.org/10.1007/978-981-19-5403-0_23
- Myung, J., Shin, H., & Kim, S. (2022). A comparative study of demand forecasting models for a multi-channel retail company: A novel hybrid machine learning approach. *Operations Research Forum*, 3(4), 58. <https://doi.org/10.1007/s43069-022-00166-4>
- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12, 2825–2830. <https://www.jmlr.org/papers/v12/pedregosa11a.html>

- Prabu, C. R., Anand, R., & Selvakumar, S. (2025). Comprehensive review on sales prediction models. *Procedia Computer Science*, 259, 1218–1227. <https://doi.org/10.1016/j.procs.2025.02.147>
- Punia, S., Shankar, S., & Gupta, S. (2020). Deep learning with long short-term memory networks and random forests for demand forecasting in multi-channel retail. *International Journal of Production Research*, 59(16), 4964–4979. <https://doi.org/10.1080/00207543.2020.1735666>
- Salinas, D., Flunkert, V., Gasthaus, J., & Januschowski, T. (2020). DeepAR: Probabilistic forecasting with autoregressive recurrent networks. *International Journal of Forecasting*, 36(3), 1181–1191. <https://doi.org/10.1016/j.ijforecast.2019.07.001>
- Sidabutar, E., & Firmansyah, E. A. (2023). Comparison of linear regression, ARIMA, and neural network models for retail sales forecasting. *Journal of Applied Data Sciences*, 4(2), 112–124. <https://doi.org/10.47738/jads.v4i2.89>
- Stranieri, S., Orzes, G., & Sartor, M. (2024). Machine learning for inventory management in retail: A systematic literature review. *International Journal of Production Economics*, 268, 109102. <https://doi.org/10.1016/j.ijpe.2023.109102>
- Syntetos, A. A., Boylan, J. E., & Disney, S. M. (2009). Forecasting for inventory planning: A 50-year review. *Journal of the Operational Research Society*, 60(1), 149–160. <https://doi.org/10.1057/jors.2008.173>
- Taparia, R., Jaipuria, S., & Srivastava, P. (2023). Retail demand forecasting using machine learning: Evidence from Indian supermarkets. *Decision Support Systems*, 168, 113932. <https://doi.org/10.1016/j.dss.2023.113932>
- Taylor, S. J., & Letham, B. (2018). Forecasting at scale. *The American Statistician*, 72(1), 37–45. <https://doi.org/10.1080/00031305.2017.1380080>
- Tian, Y., & Liu, Z. (2015). *Rossmann store sales prediction* [Technical report]. Stanford University CS229. https://cs229.stanford.edu/proj2015/215_report.pdf
- Vaclavek, L., Sudzina, F., & Pawlicek, J. (2021). Predicting out-of-stock using machine learning: An application in a retail packaged foods manufacturing company. *Electronics*, 10(22), 2787. <https://doi.org/10.3390/electronics10222787>
- Villegas, M. A., Pedregal, D. J., & Trapero, J. R. (2018). A support vector machine for model selection in demand forecasting applications. *Computers and Industrial Engineering*, 121, 1–7. <https://doi.org/10.1016/j.cie.2018.04.042>
- Wang, Z., & Srinivasan, R. S. (2025). A machine learning approach to inventory stockout prediction. *Supply Chain Analytics*, 9, 100082. <https://doi.org/10.1016/j.sca.2025.100082>
- Yang, T., & Liu, Z. (2015). *Rossmann store sales* [Project report]. Stanford University. https://cs229.stanford.edu/proj2015/215_report.pdf
- Zhang, G., Patuwo, B. E., & Hu, M. Y. (1998). Forecasting with artificial neural networks: The state of the art. *International Journal of Forecasting*, 14(1), 35–62. [https://doi.org/10.1016/S0169-2070\(97\)00044-7](https://doi.org/10.1016/S0169-2070(97)00044-7)

ANNEXURE

Annexure A: Dataset Sample

The following table presents the first 10 rows of the merged dataset after combining train.csv and store.csv using the Store column as the common key. This sample provides a concrete illustration of the raw data structure before any preprocessing steps were applied.

Table A.1 — First 10 Rows of Merged Dataset

Store	DayOfWeek	Date	Sales	Customers	Open	Promo	StateHoliday	SchoolHoliday	StoreType	Assortment	CompetitionDistance	Promo2
1	5	2015-07-31	5263	555	1	1	0	1	c	a	1270	0
2	5	2015-07-31	6064	625	1	1	0	1	a	a	570	1
3	5	2015-07-31	8314	821	1	1	0	1	a	a	14130	1
4	5	2015-07-31	13995	1498	1	1	0	1	c	c	620	0
5	5	2015-07	4822	559	1	1	0	1	a	a	29910	0

		- 31										
6	5	20 15 - 07 - 31	56 51	589	1	1	0	1	a	a	310	0
7	5	20 15 - 07 - 31	15 34 4	1414	1	1	0	1	a	c	24000	1
8	5	20 15 - 07 - 31	84 92	833	1	1	0	1	a	a	7520	0
9	5	20 15 - 07 - 31	85 65	687	1	1	0	1	a	a	2030	0
10	5	20 15 - 07 - 31	71 85	681	1	1	0	1	a	a	3160	0

Note: This sample represents open store days from July 31 2015. All stores in this sample were running promotions on this date as indicated by the Promo column value of 1. The Sales column represents the target variable for Phase 1 regression models and the base for Phase 2 stockout label creation.

Annexure B: Evaluation Metrics Formulas

The following section presents the mathematical formulas for all evaluation metrics used in this study along with a brief explanation of each formula and its components.

B.1 Phase 1 Regression Metrics

Root Mean Squared Error (RMSE)

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2}$$

Where n is the total number of observations, y_i is the actual sales value for observation i , and $\hat{y}_i$ is the predicted sales value for observation i . RMSE computes the square root of the average of all squared prediction errors. The squaring operation penalizes large errors more heavily than small ones making RMSE particularly sensitive to outliers. Lower RMSE values indicate better model performance.

Mean Absolute Error (MAE)

$$MAE = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i|$$

Where n is the total number of observations, y_i is the actual sales value, and $\hat{y}_i$ is the predicted sales value. MAE computes the average of the absolute differences between predicted and actual values treating all errors equally regardless of magnitude. In this study an MAE of 593 means the model's predictions were on average 593 Euros away from the actual sales value. Lower MAE values indicate better model performance.

B.2 Phase 2 Classification Metrics

Before presenting the classification metric formulas the following terminology must be defined:

Term	Definition
True Positive (TP)	A stockout day correctly identified as a stockout by the model
True Negative (TN)	A normal day correctly identified as normal by the model
False Positive (FP)	A normal day incorrectly identified as a stockout by the model
False Negative (FN)	A stockout day incorrectly identified as normal by the model

Precision

$$Precision = \frac{TP}{TP + FP}$$

Precision measures what proportion of days flagged as stockouts by the model were actually stockout days. A high precision means the model rarely raises false alarms.

Recall

$$Recall = \frac{TP}{TP + FN}$$

Recall measures what proportion of actual stockout days were correctly identified by the model. A high recall means the model rarely misses actual stockout events.

F1 Score

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

F1 Score is the harmonic mean of Precision and Recall. It balances both concerns — avoiding false alarms and avoiding missed stockouts — into a single metric. A perfect F1 Score of 1.0 means the model correctly identified every stockout day with no false alarms and no missed events. Higher F1 Score values indicate better classification performance.

AUC-ROC

The Area Under the Receiver Operating Characteristic Curve measures the model's ability to distinguish between stockout and non-stockout days across all possible classification thresholds.

$$TPR = \frac{TP}{TP + FN}$$

$$FPR = \frac{FP}{FP + TN}$$

The AUC is the area under the ROC curve plotted between TPR and FPR across all thresholds. An AUC-ROC of 1.0 represents perfect classification and an AUC-ROC of 0.5 represents random guessing. Higher values indicate better discriminative ability.

Annexure C: Abbreviations and Glossary

Table C.1 — List of Abbreviations

Abbreviation Full Form

ML	Machine Learning
RMSE	Root Mean Squared Error
MAE	Mean Absolute Error
F1	F1 Score
AUC-ROC	Area Under the Receiver Operating Characteristic Curve
SMOTE	Synthetic Minority Oversampling Technique

Abbreviation Full Form

LSTM	Long Short Term Memory
RF	Random Forest
XGBoost	Extreme Gradient Boosting
OOS	Out of Stock
EDA	Exploratory Data Analysis
TP	True Positive
TN	True Negative
FP	False Positive
FN	False Negative
CSV	Comma Separated Values
GPU	Graphics Processing Unit

Table C.2 — Glossary of Key Terms

Term	Definition
Stockout	A situation where a product is unavailable at the point of sale due to depleted inventory resulting in lost sales
Demand Forecasting	Predicting future customer demand using historical sales data and machine learning models
Class Imbalance	A condition where one class has significantly more observations than another making minority class detection difficult
Ensemble Method	A machine learning approach that combines predictions from multiple models to produce a more accurate final prediction
Oversampling	A technique to address class imbalance by increasing minority class observations through synthetic sample generation
Feature Engineering	Selecting and transforming raw data columns into informative input variables for model training
Binary Classification	A task where the model predicts one of two outcomes such as stockout or no stockout
Regression	A task where the model predicts a continuous numerical value such as

Term	Definition
	daily sales volume
Proxy Label	An indirect indicator used as a substitute for a variable that cannot be directly measured such as sales drops used as a proxy for stockouts
Hyperparameter	A model configuration parameter set before training begins that controls the learning process
Epoch	One complete pass through the entire training dataset during neural network training
Dropout	A regularization technique that randomly deactivates neurons during training to prevent overfitting