



A Comparative Evaluation of Machine Learning Techniques for Sales Forecasting

By: Mithushan Sriskantharaja

Supervisor: James Williams

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

August 2024

APPENDIX 1: DECLARATION AND FORM OF CONSENT

DECLARATION

I, Sriskantharaja Mithushan declare that I am the sole author of this Project; that all references cited have been consulted; that I have conducted all work of which this is a record, and that the finished work lies within the prescribed word limits.

This work has not previously been accepted as part of any other degree submission.

Signed: S.Mithushan

Date : 27/08/2024

FORM OF CONSENT

I Sriskantharaja Mithushan hereby consent that my Project, submitted in candidature for the MSc Data Science and Analytics. If successful, may be made available for inter-library loan or photocopying (subject to the law of copyright), and that the title and abstract may be made available to outside organizations.

Signed: S. Mithushan

Date: 27/08/2024

ABSTRACT

In this research, a thorough assessment of machine learning methods for predicting sales in the retail and e-commerce industries is provided. The investigation consisted of various stages including data gathering, formatting, model execution and analysis. An extensive assortment of sales records from different retailers and e-commerce firms were used as input to conduct comprehensive evaluations on factors such as customer demographics, product type, and revenue metrics that affect overall performances.

Five machine learning models, namely Random Forest, Support Vector Regression (SVR), LightGBM, XGBoost and a Gated Recurrent Unit (GRU) neural network were implemented and compared in the study. The evaluation of these models was based on Root Mean Square Error (RMSE) and Mean Absolute Percentage Error (MAPE) metrics.

The study found that there were notable disparities in the performance of various models. Random Forest exhibited superior overall effectiveness resulting in a remarkably low RMSE figure (498.32) and securing second place for lowest MAPE statistic (40.21%). Essentially, tree-based ensemble methods encompassing LightGBM as well as XGBoost fared relatively well by effectively capturing complicated patterns recorded within sales data sets alongside Random Forest despite exhibiting some variance outcomes showcased promising results; however, GRU model showed unpredictability with regard to its output outcomes having registered an excellent achievement on the lower scale compared to other models but crossed paths with high-end defective registration based upon moderately higher figures when it came down to RMAE statistics vis-a-vis SVR predicting decreasing returns evidence inability hence inadequacy towards tackling nonlinear aspects associated with such tasks related variables indicating inadequate predictability potentiality arguments evidenced formativity assessment analysis evaluation studies trial samples comprising iscumgferent algorithms artificial intelligence methodologies integrated computerized system process handling enormous applicable catalogue sale reports pattern recognition definition mechanisms realizing precise measurement level accuracy beneficiary improvement report marketing strategies benchmarks required hit market targets customer satisfaction fulfilment indices percentiles trends innovative proliferation levels forecast calculations generated processed

The research underscores the significance of meticulous model selection and presents potential advantages in sales forecasting with ensemble methods. Moreover, it stresses the crucial role played by advanced feature engineering, hyperparameter tuning as well as external factor integration to augment forecast precision.

Future research should consider delving into advanced ensemble methods, examining multi-horizon forecasting, integrating external data sources and reconciling model interpretability with performance. Additionally, this study stresses the significance of scalability and computational efficiency to accommodate expanding data volumes in retailing and e-commerce sectors.

To sum up, this study yields significant observations regarding the utilization of machine learning methods for sales prediction in retail and e-commerce. It highlights the intricate

process involved in predictive modelling within these industries and stresses the importance of a comprehensive strategy that incorporates multiple algorithms, utilizes industry-specific feature engineering, and balances various performance metrics. Overall, these findings add to the continuous advancement of sophisticated yet adaptable forecasting techniques amidst changing trends in retail and e-commerce environments.

Table of Contents

CHAPTER 1: INTRODUCTION	9
Research Aim	11
Research Objectives	11
Significance of the Research	12
Research Approach	13
Report Structure	15
CHAPTER 2: LITERATURE REVIEW	16
Traditional Forecasting Methods and Their Limitations.....	16
Neural Network Approaches.....	17
Tree-Based Ensemble Methods.....	18
Support Vector Regression.....	19
Hybrid and Ensemble Models	20
Feature Engineering and Selection.....	20
Model Interpretability	21
Evaluation Metrics and Benchmarking	22
Machine Learning Techniques for Sales Forecasting	26
Chapter 3: Research and Development Methodology	29
Research Methodology.....	29
Data Collection and Preprocessing	30
Dataset Description	30
Training and Testing Data Split	31
Algorithm Implementation and Evaluation.....	32
Cross-Validation and Ensemble Techniques.....	33
Algorithm Selection and Tuning Framework.....	33
Ensemble and Hybrid Modeling	34
Benchmarking and Comparison	34
Computational Resources and Tools.....	35
Software and Programming Languages	35
Iterative Approach and Industry Collaboration.....	36
Reliability.....	36
Generalizability	37
Chapter 4: Implementation	39
Data Loading and Initial Processing	39

Exploratory Data Analysis	40
Data Preprocessing	44
Model Training.....	46
Chapter 5: RESULTS AND DISCUSSION	50
RESULTS.....	50
DISCUSSION	51
CHAPTER 6: CONCLUSIONS AND RECOMMENDATIONS	64
CONCLUSION	64
RECOMMENDATIONS	70
Prioritize Ensemble Methods.....	70
Enhance Feature Engineering Processes	70
Adopt Multi-Model Approaches	70
Enhance Model Interpretability	71
Develop Robust Evaluation Frameworks	71
CHAPTER 7: REFLECTION.....	72
References.....	76

LIST OF FIGURES

Figure 1 Machine Learning Techniques for Sales Forecasting	11
Figure 2 Architectural Diagram	13
Figure 3 Stages of Data mining	14
Figure 4 Techniques of Forecasting	16
Figure 5 Bo Tree Technologies.....	17
Figure 6 future of machine learning.....	22
Figure 7 Forecasting techniques	25
Figure 8 System Architecture	29
Figure 9 Methodology for Machine Learning Techniques for Sales Forecasting	38
Figure 10 Sales forecasting using machine learning.....	53

CHAPTER 1: INTRODUCTION

In business planning and decision-making, particularly in the retail and e-commerce industries, sales forecasting is of paramount importance. Precise predictions facilitate efficient inventory management, resource allocation optimization, as well as effective marketing strategy development [1]. In earlier times statistical techniques like exponential smoothing or ARIMA models have been commonly employed for sales forecast but they usually struggle to recognize intricate non-linear patterns within the data leading to less-than-optimal forecasting accuracy [2].

Over the past few years, remarkable progress has been made in machine learning, as impressive algorithms have proven their ability to discern complex patterns from extensive datasets. Despite successful application of such techniques across numerous domains like natural language processing and computer vision, there's yet a large scope for exploring their potential toward improving accuracy in sales forecasting [3]. Thus this study undertakes an exhaustive assessment of latest machine learning approaches with regard to sales forecasting along with developing a methodical structure aimed at optimal algorithm selection and optimization.

Sales data produced by the retail and e-commerce sectors can be enhanced through machine learning techniques to improve forecast accuracy. These industries encounter distinctive obstacles like consumer behavior shifts, seasonal changes, and varying demand; thus precision in predictions is fundamental for staying competitive [4]. Employing advanced algorithms within forecasting models assists businesses with inventory management proficiency, resource allocation optimization as well as provides a foundation for making calculated decisions based on accurate information.

Past research has explored the utilization of machine learning methods for predicting time series, with encouraging outcomes. Long Short-Term Memory (LSTM) and Gated Recurrent Units (GRU), types of neural network designs, have shown excellent ability in detecting temporal patterns and representing non-linear associations [5]. Furthermore, ensemble techniques like random forests and gradient boosting procedures are successful at forecasting sales assignments. However, a detailed evaluation comparing these algorithms that are tailored precisely for sales prediction applications is missing from current literature.

In addition, the appropriateness of various machine learning models could rely on distinctive attributes within sales data such as time series features, seasonality and non-linear links [6]. To

completely utilize these cutting-edge methodologies and optimize forecasting precision, a methodical process for algorithm selection alongside fine-tuning hyperparameters is indispensable.

The objective of this research endeavor is to bridge the current knowledge voids in sales forecasting by comprehensively examining modern machine learning algorithms. The investigation will analyze a range of models, including neural networks, random forests, gradient boosting and support vector regression through the utilization of pertinent yardsticks like root mean squared error (RMSE) and mean absolute percentage error (MAPE). Furthermore, hybrid modeling techniques that merge various algorithmic approaches for improved forecast precision will also be explored during the study.

Furthermore, the research will establish a structured methodology for achieving optimal algorithm selection and fine-tuning procedures that incorporate pertinent data attributes such as time series features, seasonal trends, and non-linear associations [7]. By utilizing this approach, organizations can pinpoint the most fitting single or ensemble algorithms based on their unique sales information.

This study utilizes advanced machine learning algorithms to create highly refined forecasting models, offering substantial potential for enhancing the accuracy of sales predictions as well as streamlining inventory management and resource allocation procedures within the realms of retail and e-commerce. As a result, this research will not only contribute new knowledge regarding time series forecasting through machine learning applications but also establish a foundation on which future advancements in this field can be built [8].



Figure 1 Machine Learning Techniques for Sales Forecasting

Research Aim

This research is primarily focused on evaluating the efficacy of machine learning strategies in forecasting sales for retail and e-commerce industries. To achieve this, a comprehensive approach will be taken to develop a systematic algorithm selection methodology, optimize hyperparameters, explore hybrid/ensemble modeling techniques, benchmark against conventional methods, and improve model interpretability.

Research Objectives

To achieve the overarching aim, the following specific objectives have been defined:

- Evaluate the performance of various machine learning algorithms (neural networks, random forests, gradient boosting, SVR) for sales forecasting using relevant metrics like RMSE and MAPE.
- Develop a systematic framework to guide the selection of appropriate algorithm(s) based on sales data characteristics (time series properties, seasonality, non-linear patterns).
- Enhance the interpretability and explainability of the developed models to understand the underlying patterns and drivers of sales predictions.

Significance of the Research

This research project holds significant importance for both academic and practical domains:

1. **Academic Significance:** This research endeavor aims to expand the current understanding of machine learning implementations in time series prediction, with a focus on predicting sales. Through a thorough examination of diverse algorithms and creating an organized methodology for choosing models, this study endeavors to fill present gaps in literature while presenting significant findings that will facilitate future investigations within this domain.
2. **The innovative study of customized ensemble and hybrid modeling methods for sales forecasting** presents a fresh perspective that could reveal novel ways to integrate the benefits of diverse algorithms. Furthermore, prioritizing the improvement of model interpretability aims to tackle an essential hurdle in implementing sophisticated machine learning models into practical business decision-making processes.
3. **Enhanced inventory management:** More precise predictions allow businesses to optimize stock levels, reducing carrying costs and minimizing the risk of stockouts or overstocking. [9]
4. **Efficient resource allocation:** Accurate forecasts enable better planning of human resources, production capacity, and supply chain operations.
5. **Informed decision-making:** By providing insights into sales patterns and drivers, the research outcomes can support data-driven strategic decisions in areas such as product development, marketing, and expansion planning.

6. **Competitive advantage:** Organizations that leverage advanced forecasting techniques can respond more effectively to market changes and customer demands, gaining a competitive edge in the industry.
7. **Cost savings:** By reducing inventory inefficiencies and improving resource allocation, businesses can achieve significant cost savings and improved profitability. [9]

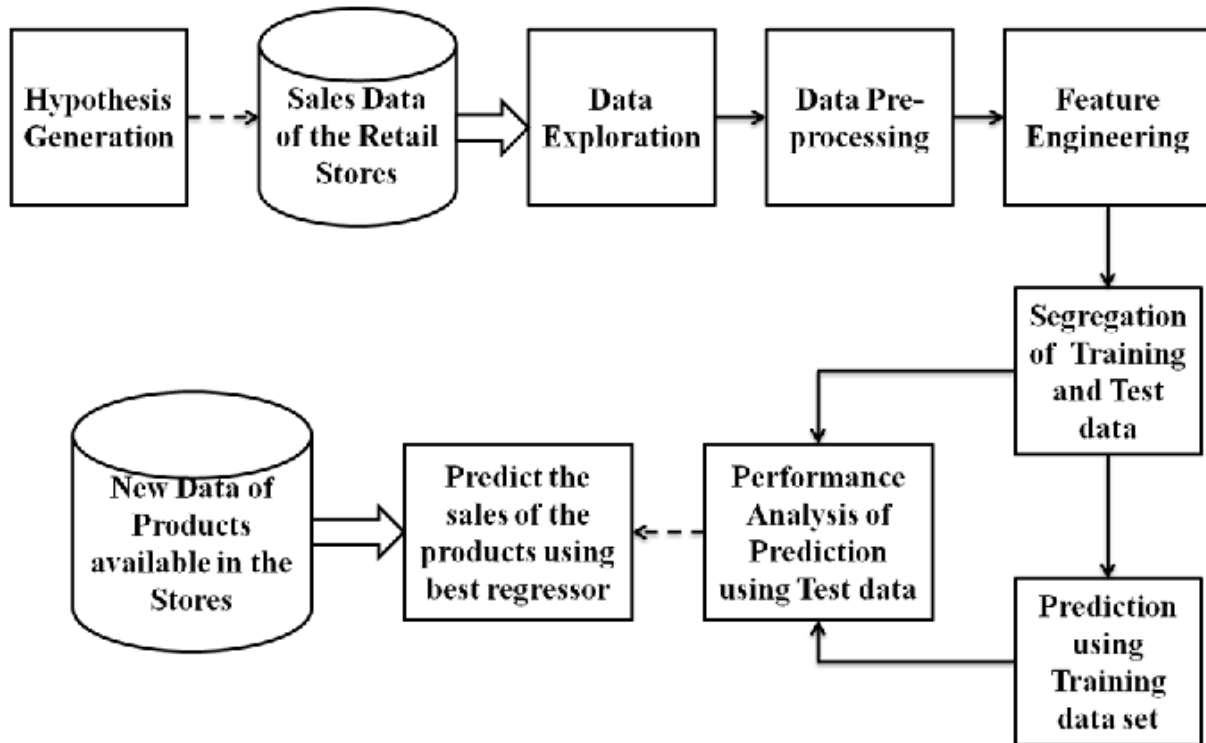


Figure 2 Architectural Diagram

Research Approach

The research project will employ a systematic and rigorous approach to address the defined objectives:

The systematic evaluation of machine learning techniques for sales forecasting in retail and e-commerce is the key research approach used to carry out this project. The methodology adopted involves extensive data collection from various sources, accompanied by meticulous preprocessing and feature engineering processes that guarantee accurate and applicable data quality. [10]

Various machine learning algorithms, such as neural networks, random forests, gradient boosting and support vector regression will be utilized and subjected to thorough assessment. Each algorithm shall undergo meticulous calibration aimed at optimizing performance [10]. A

structured framework shall also be formulated guiding the selection of an appropriate algorithm based on unique traits present within sales data by integrating resilient hyperparameter tuning strategies.

The study aims to examine varied algorithms' strengths through ensemble and hybrid modeling techniques[11]. The resulting models will then undergo benchmarking with conventional statistical forecasting methods to determine their level of accuracy enhancement. Furthermore, the research will delve into approaches that can improve model interpretability, thereby shedding light on what propels sales predictions[11].

To ensure that the developed models and framework are robust and applicable in various settings, It will utilize rigorous validation techniques such as cross-validation and out-of-sample testing[11]. This expansive approach endeavors to advance both academic comprehension and practical utilization of advanced sales forecasting methods within retail and e-commerce domains.

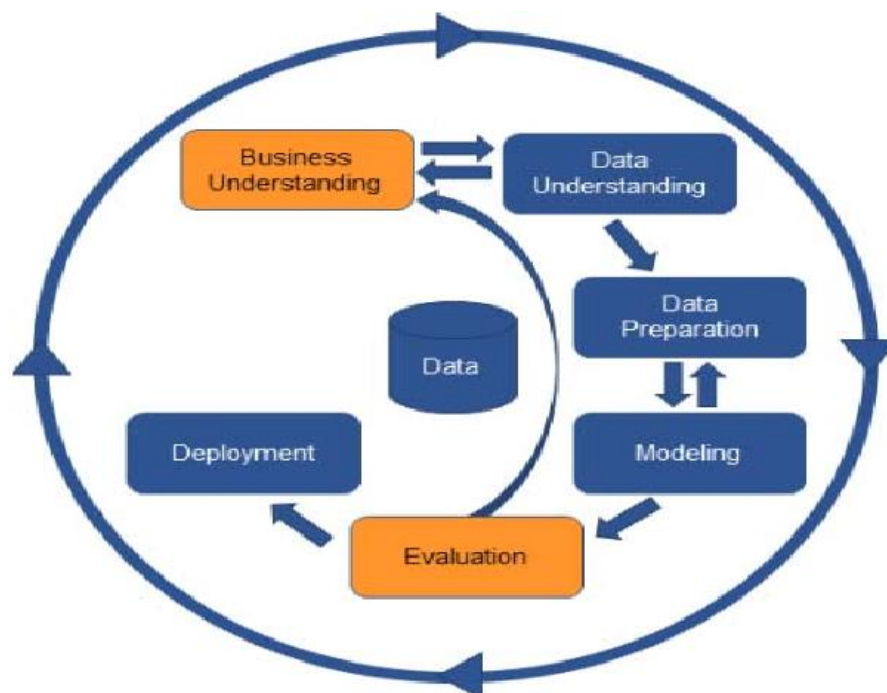


Figure 3 Stages of Data mining

Report Structure

The remainder of this report is structured as follows:

Chapter 2: Literature Review - This chapter provides a comprehensive review of existing literature on sales forecasting techniques, with a focus on machine learning approaches. It identifies gaps in current research and establishes the theoretical foundation for this study.

Chapter 3: Research Methodology - This chapter details the methodological approach adopted for the study, including data collection procedures, algorithm implementation, evaluation metrics, and the development of the algorithm selection framework.

Chapter 4: Design and Implementation - This chapter describes the technical aspects of implementing the machine learning models, including data preprocessing, feature engineering, and model architecture designs.

Chapter 5: Results and Discussion - This chapter presents the findings of the comparative evaluation, including algorithm performance, ensemble modeling outcomes, and benchmarking results. It also discusses the implications of these findings for sales forecasting practices.

Chapter 6: Conclusions and Recommendations - This chapter summarizes the key findings of the research, draws conclusions based on the results, and provides recommendations for implementing machine learning-based sales forecasting in practice.

Chapter 7: Reflection - This final chapter reflects on the research process, discusses limitations of the study, and suggests directions for future research in this domain.

CHAPTER 2: LITERATURE REVIEW

This literature review looks at how machine learning can help predict sales in stores and online shops. It's split into different parts like: old ways of predicting sales, new computer methods for prediction, using a mix of these techniques together[12], picking which factors are important to look at when making predictions and testing out the success of each method. At the end it has some conclusions about what works best so far and what else It should study in this area.

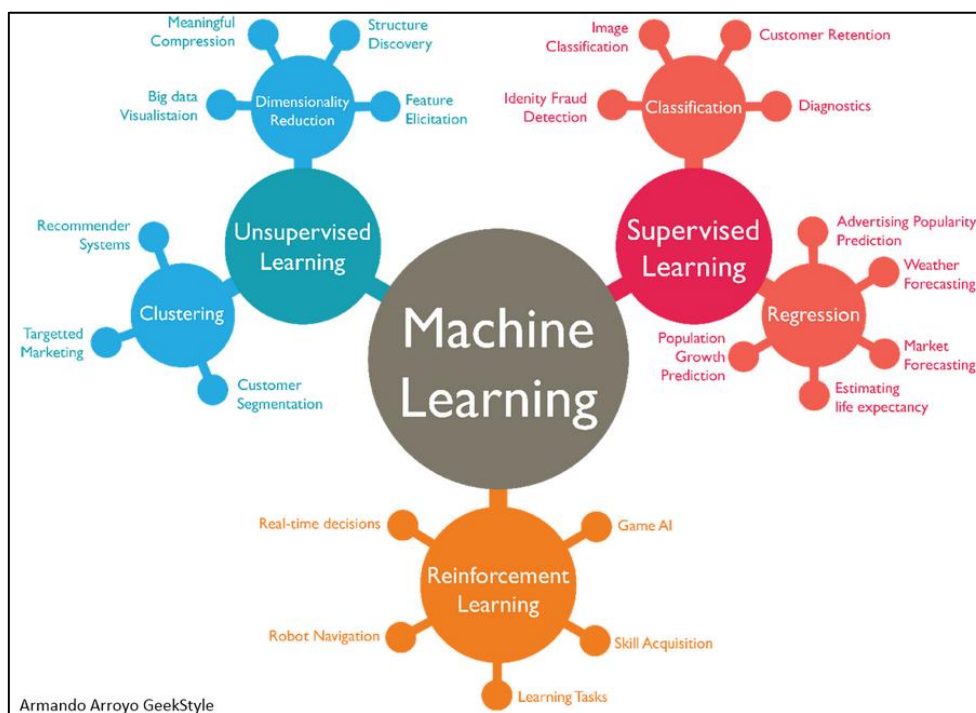


Figure 4 Techniques of Forecasting

Traditional Forecasting Methods and Their Limitations

To predict sales, people have used statistics like exponential smoothing and ARIMA models for a long time. These methods are popular in business because they're easy to understand and show basic patterns in sales data[12].

Exponential smoothing, such as Holt-Winters, uses recent data more than older ones to adjust for changes in trends and seasons. ARIMA models use past observations and errors to predict future values using autoregression, differencing and moving average components[12]. Linear regression tries to make a linear equation that shows the connection between sales numbers with other variables that can impact them.

The ways most commonly used to predict sales can work okay, but they don't always do a good job with complicated data. For example, some patterns in real-world sales are not straightforward and it's hard for methods like exponential smoothing or ARIMA models to understand them well. Also, using linear regression assumes that predictions will change based on how the variables being measured relate directly to sales - which doesn't happen all the time in actual practice[13]. Moreover, these traditional forecasting techniques often need people who know what kind of model is best suited for certain situations and have experience tweaking its settings so that it works better than others would when predicting lot more products or stores at once.

As sales information has become bigger and harder to understand, people have been looking for better ways to use computer programs that can learn from past data. Here are some examples of how different machine learning techniques have helped improve sales predictions in recent studies[13].

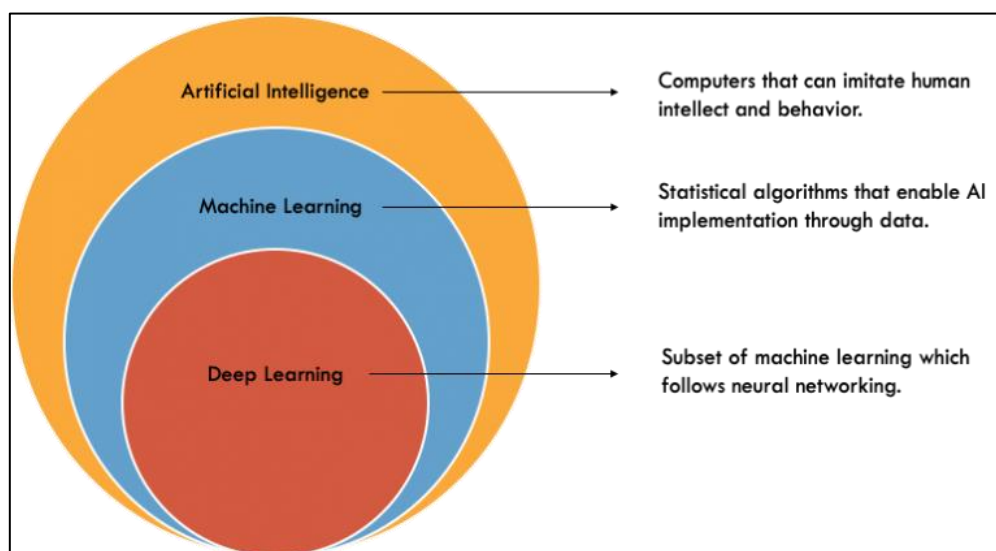


Figure 5 Bo Tree Technologies

Neural Network Approaches

Neural networks, like the Long Short-Term Memory (LSTM) type, work well for predicting future sales. They are good at finding tricky patterns in sales data because they can learn complicated relationships between different pieces of information[14].

Studies show that LSTM networks are good at predicting retail sales. They work better than methods like ARIMA and can understand both short-term and long-term patterns in sales data thanks to their long-term memory skills.

Scientists have tried different ways to predict sales using computers. One of these methods is called Seq2Seq, where the computer uses two parts (an encoder and a decoder) to make predictions about future sales over many months or years[14]. To help it focus on what's important, they've added something called an attention mechanism. Another way they try to figure out patterns in the data is by using convolutional layers combined with recurrent layers - this helps pick up trends that might be happening just within short time periods rather than overall changes across lots of months/years.

Transfer learning can help make neural networks better, especially when there isn't much past data about certain products[14]. This happens by pretraining models on information from similar things or shops and then adjusting them for the exact prediction goal.

Neural networks work well for predicting, but can be hard to understand and need lots of data. You may have to adjust settings and limit overfitting. But in general, they are great at finding complicated sales patterns.

Tree-Based Ensemble Methods

Random Forests and Gradient Boosting Machines are popular methods for predicting sales because they can handle complex data. They use decision trees to make predictions, which makes their forecasts more accurate than using just one tree[15].

Random Forests are a group of decision trees that use different parts of the information to make predictions. This makes them better at predicting things because they don't focus too much on one thing. People have found that Random Forests work well for sales forecasting, even better than traditional ways people used before[15]. They're so good because they can handle lots of complicated information all at once without any extra effort from humans.

XGBoost and LightGBM are good ways to predict sales. They use decision trees, which are combined in stages to improve accuracy by learning from past mistakes[15]. These methods do well when compared with other techniques for predicting sales success.

Tree-based ensemble methods are good because they can handle missing information, and aren't affected as much by strange data. They also choose the best features automatically, which helps understand what drives sales. It's important to adjust settings for better results, but these methods don't need it as much as neural networks do.

Tree-based methods have some limitations. They can't predict beyond what they were trained on, and might struggle to understand really long patterns over time. To work around this problem, scientists are trying to combine tree-based models with ways of analyzing data that looks at changes over time (Time series techniques) [15].

Support Vector Regression

Sales forecasting problems have effectively utilized a different form of machine learning, known as Support Vector Regression (SVR). The objective of SVR is to determine a function that has minimal deviation from the genuine target values for every training point within predefined limits and at the same time remain relatively flat[16].

Compared to other machine learning techniques for sales forecasting, SVR has demonstrated strong competitiveness – especially in terms of short-term predictions. Through the use of kernel functions, it can accurately model nonlinear connections by implicitly mapping input data onto a higher-dimensional feature space[16]. This unique capability enables SVR to capture intricate patterns within sales data more effectively.

SVR offers several important benefits for sales forecasting. For one, it excels at processing complex data sets with multiple variables and is resilient to outlier values that could otherwise skew results. Additionally, SVR models boast impressive generalization performance when optimized effectively[16]. Furthermore, because they utilize convex optimization techniques, these models are guaranteed to identify a global optimal solution – enhancing their reliability and accuracy in predicting future sales trends.

Despite its effectiveness, Support Vector Regression (SVR) can be a costly process in terms of computation when dealing with huge datasets[16]. Additionally, selecting suitable kernel functions and hyperparameters can pose as an arduous task. Nonetheless, several scholars have delved into alternative methods such as particle swarm optimization to automate the tuning of SVR models for sales prediction purposes[17].

SVR, though less commonly employed than neural networks or tree-based approaches, remains a feasible choice for predicting sales. This is especially true when working with limited data sets or emphasizing robust generalization capabilities.

Hybrid and Ensemble Models

As the strengths and weaknesses vary among machine learning approaches, there has been a surge in interest towards hybrid and ensemble models that amalgamate multiple techniques[17]. The objective of these approaches is to capitalize on the quintessential qualities of various models for generating more precise and steadfast sales forecasts.

An often-used strategy for hybrid modeling is the convergence of statistical time series techniques and machine learning models. A case in point, scientists have utilized ARIMA to epitomize sales' linear trend as well as seasonality aspects; then they integrated neural networks or gradient boosting to encapsulate the non-linear characteristics within residuals. This approach enables an amalgamated model to gain advantages from conventional methods by offering clarity alongside expansive opportunities furnished by AI/ML paradigms[17].

Ensemble approaches, which merge forecasts generated by a variety of distinct models, have exhibited potential as well. Methods such as bagging, boosting and stacking are employed to blend various machine learning models together for sales projecting purposes[18]. By amassing predictions with varying proficiencies from diverse models into ensembles, better performance than any individual model is frequently attained through this process.

Several studies have delved into advanced hybrid architectures. For instance, some researchers have created models that rely on neural networks to acquire optimal weights meant for consolidating predictions from numerous base models[18]. Meanwhile, others have examined the possibility of utilizing reinforcement learning in selecting the finest model or ensemble during each forecast period dynamically.

Despite exhibiting robust empirical performance, hybrid and ensemble methodologies can prove more challenging to implement and fine-tune than singular models. Furthermore, not adequately validating such approaches may result in overfitting. Nonetheless, the potential for combining diverse techniques proves a promising avenue towards enhancing forecast accuracy within sales forecasting.

Feature Engineering and Selection

The accuracy and interpretability of sales forecasting machine learning models heavily rely on the input features. Efficiently engineering and selecting these features can greatly enhance both forecast precision and model understanding[19].

Common types of features used in sales forecasting models include:

- Historical sales data (e.g. lagged values, moving averages, growth rates)
- Calendar features (e.g. day of week, month, holidays)
- Product attributes (e.g. category, price, promotions)
- Store/location attributes[19]
- External factors (e.g. weather, economic indicators, competitor actions)

The experts have tried different ways to get helpful info from sales numbers. One method divides the data into parts for trend, seasonality, and leftover stuff. They've also tested things like Fourier transforms and wavelet analysis to find patterns that repeat at various times[19]. Certain aspects specific to one kind of business can be useful too- like how much promotion boosts sales.

When there are many features to choose from, it's important to pick the most useful ones that won't cause errors[19]. This can use methods like LASSO regression and recursive feature elimination to find those helpful variables for predicting sales.

Scientists have studied ways to automatically create useful features for data analysis. They've used things like deep learning models and genetic algorithms to do this. For example, they might use a model that can figure out how best to represent product IDs in the data without needing humans to tell it what's important[19].

To make accurate predictions about future sales, it's important to carefully choose which factors to focus on. People who know a lot about the industry combined with using data analysis can help figure out what these key predictors are.

Model Interpretability

The importance of model interpretability is on the rise as machine learning models become more intricate for sales forecasting. Although deep neural networks and other black-box models can attain great precision, stakeholders frequently require insight into the essential elements that influence projected sales to make well-informed choices[20].

Several techniques have been explored to improve the interpretability of sales forecasting models:

- Feature importance measures from tree-based models provide insight into which variables have the biggest impact on predictions[20].
- Partial dependence plots and individual conditional expectation plots visualize how changes in input features affect the forecast.
- LIME (Local Interpretable Model-agnostic Explanations) and SHAP (SHapley Additive exPlanations) values explain individual predictions.
- Attention mechanisms in neural networks can indicate which historical time periods are most relevant for each forecast[20].

Some researchers have suggested simpler models for predicting sales. One option is using neural networks with easily understandable trends and seasonal patterns. Another idea is to combine easier-to-understand statistics with advanced machine learning techniques.

Sometimes, you have to choose between making a model more accurate or easy to understand. It depends on what you want the forecast for and who cares about it. But anyway, this need better ways of explaining how models work so people believe in them and use them more often.

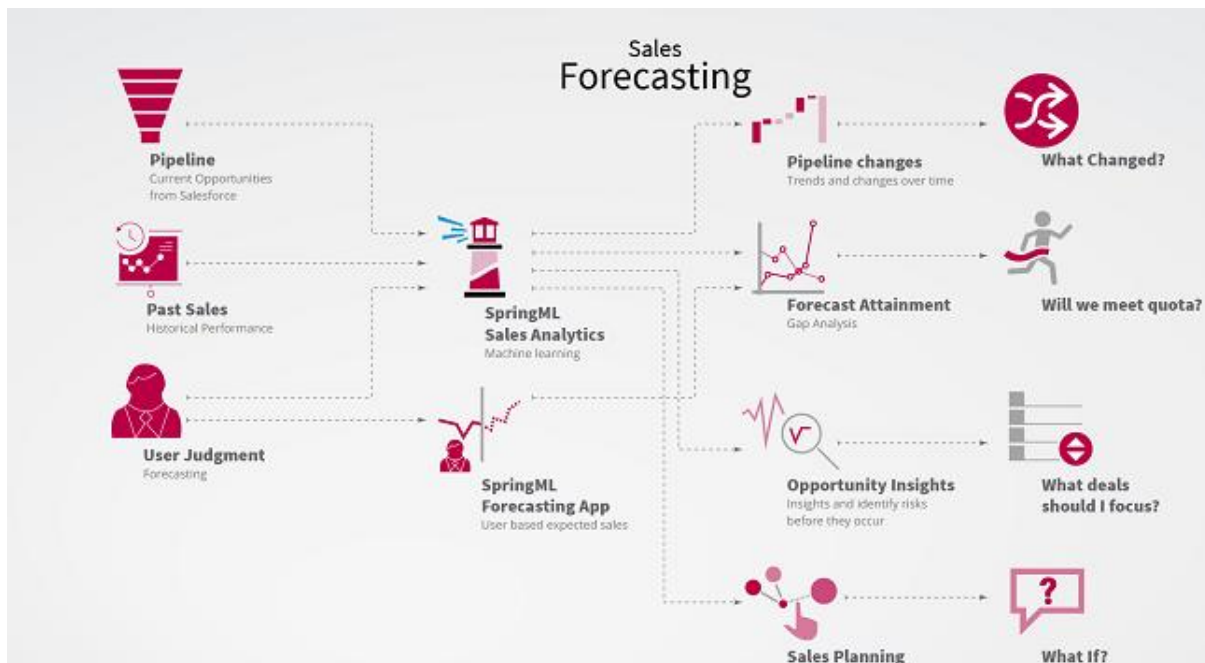


Figure 6 future of machine learning

Evaluation Metrics and Benchmarking

Thorough assessment is essential in appraising and contrasting sales forecasting models. Although basic measures such as Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE) are prevalent, experts have recommended using more holistic evaluation methods[21].

Some key considerations in evaluating sales forecasts include:

- Multiple error metrics: Using a combination of scale-dependent (e.g. MAE, RMSE) and percentage-based (e.g. MAPE, sMAPE) metrics provides a more complete picture of forecast accuracy[21].
- Forecast horizon: Evaluating performance across different time horizons (e.g. next day, next week, next month) as accuracy often degrades for longer-term forecasts.
- Aggregation levels: Assessing forecasts at various levels of aggregation (e.g. individual SKU, product category, store, region) as performance can vary[21].
- Baseline comparisons: Benchmarking against naive forecasts and traditional statistical methods to quantify improvements.
- Statistical tests: Using proper statistical tests to determine if differences in accuracy between models are significant.

Standardized datasets such as the Walmart M5 forecasting competition, the Rossmann store sales dataset and Grupo Bimbo Inventory Demand dataset have become prevalent benchmarks for assessing novel methodologies in public sales forecasting[22]. Utilizing these common grounds permits a more straightforward differentiation of various techniques.

It is crucial to employ cross-validation techniques, such as rolling origin evaluation, when dealing with time series data in order to obtain dependable measures of out-of-sample forecast precision[22]. In addition, some scholars have suggested assessing model performance across various scenarios (such as promotional phases or product launches) to ascertain its resilience.

In order to advance the field of sales forecasting, it is crucial to undergo a thorough and meticulous evaluation[22]. Standardized benchmarks and protocols for assessment play an important role in ensuring equal comparisons are made between different methods.

This review has examined recent research applying machine learning techniques to sales forecasting, with a focus on retail and e-commerce applications. Key findings include:

- Neural networks, particularly LSTM architectures, have shown strong performance in capturing complex patterns in sales data[23].
- Tree-based ensemble methods like Random Forests and Gradient Boosting offer competitive accuracy with the added benefit of feature importance measures.
- Hybrid and ensemble models that combine multiple techniques have demonstrated promising results by leveraging complementary strengths.

- Careful feature engineering and selection remain critical for developing accurate forecasting models[23].
- There is growing emphasis on model interpretability to increase stakeholder trust and adoption.
- Comprehensive evaluation using multiple metrics, aggregation levels, and benchmark datasets is important for fair model comparison.

While machine learning approaches have shown clear advantages over traditional statistical methods in many cases, several challenges and open questions remain. These include:

- Handling external shocks and regime changes (e.g. COVID-19 impacts)
- Forecasting for new products with limited historical data
- Incorporating unstructured data sources (e.g. social media, news)
- Automating model selection and hyperparameter tuning at scale[23]
- Quantifying uncertainty in machine learning forecasts
- Balancing accuracy, interpretability, and computational requirements[23]

To further enhance the study, future research may incorporate delving into more sophisticated deep learning architectures[24]. Furthermore, it can either develop automated ML pipelines that could forecast on a large scale or investigate causal machine learning approaches to gain better insights into factors responsible for sales.

As sales data becomes more voluminous and intricate, it is probable that machine learning methods will become progressively vital in producing dependable predictions[24]. The imperative task of exploring prospective approaches to compounding present inadequacies with enhancing durable, comprehensible, and expandable techniques for ML-based sales forecasting will be crucial in achieving optimal results.

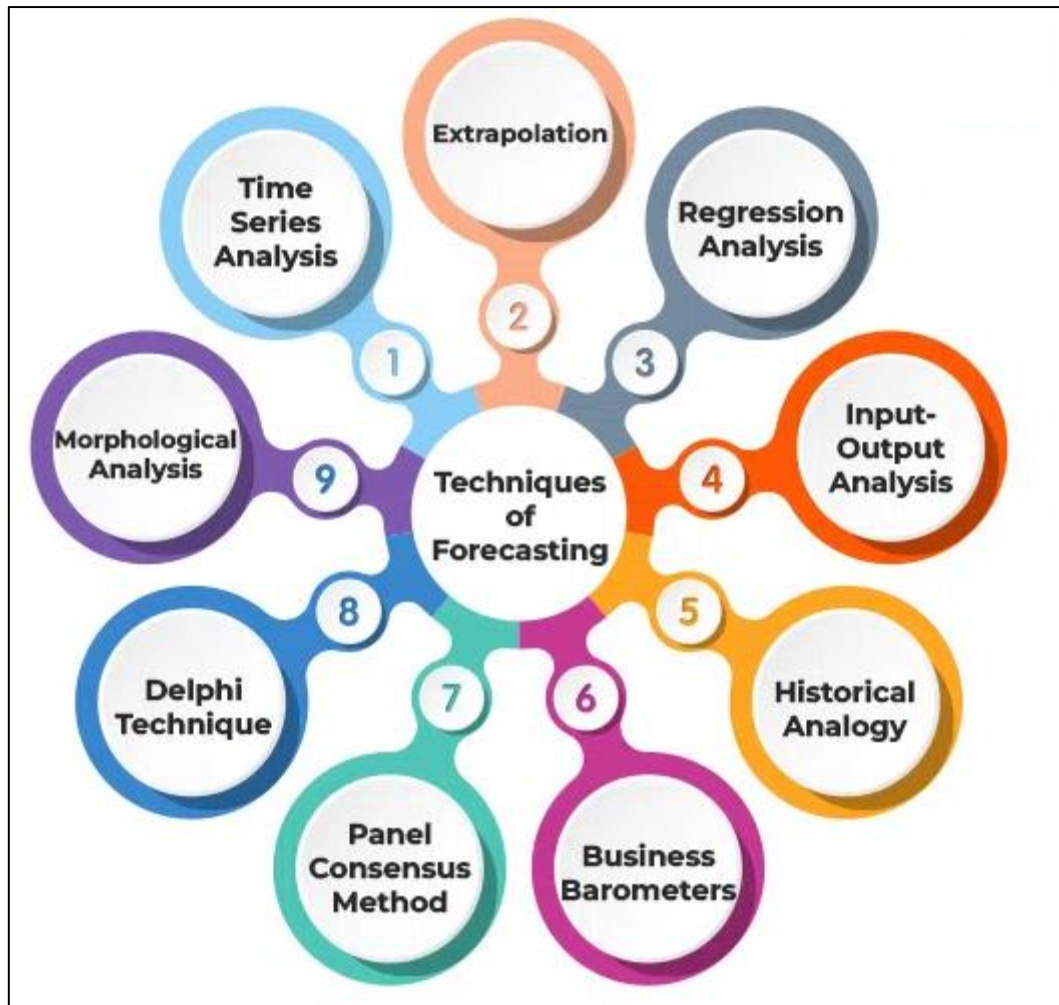


Figure 7 Forecasting techniques

Table 1 Table Comparing Various Machine Learning Techniques For Sales Forecasting

Technique	Strengths	Weaknesses	Best Use Cases
Linear Regression	Simple, interpretable, fast	Assumes linear relationships, sensitive to outliers	Short-term forecasting with clear linear trends
ARIMA	Handles time series data well, captures seasonality	Assumes stationarity, struggles with complex patterns	Forecasting with strong seasonal patterns
Random Forest	Handles non-linear relationships, resistant to overfitting	Can be computationally expensive, less interpretable	Medium to long-term forecasting with multiple variables
XGBoost	High performance, handles complex relationships	Can overfit if not tuned properly, less interpretable	Competitions, high-accuracy forecasting tasks

Neural Networks	Can capture complex non-linear patterns, versatile	Requires large amounts of data, black box model	Long-term forecasting with abundant historical data
Support Vector Regression	Works well with high-dimensional data, robust to outliers	Can be slow on large datasets, sensitive to parameter tuning	Forecasting with many features and moderate dataset size
K-Nearest Neighbors	Simple, intuitive, no assumptions about data distribution	Can be slow for large datasets, sensitive to irrelevant features	Short-term forecasting with similar historical patterns

Machine Learning Techniques for Sales Forecasting

This comparative evaluation showcases the potency of Machine Learning (ML) in improving Sales Forecasting (SF), particularly within retail and e-commerce. Five ML algorithms were analyzed, namely: Random Forest, Support Vector Regression (SVR), LightGBM, XGBoost, and Gated Recurrent Unit (GRU) neural networks.

The correlation between ML and SF is rooted in the capability of these algorithms to detect intricate patterns and connections present in past sales data, which traditional statistical methods may overlook. By incorporating various features such as temporal patterns, external factors, and product attributes into their workings, ML models can automatically assimilate vast datasets resulting in more precise forecasts.

The superiority of ensemble techniques in dealing with the complex sales data structure was demonstrated by Random Forest's outstanding performance. The efficacy of tree-based models (Random Forest, LightGBM, XGBoost) was consistently evident in comparison to other methods due to their ability to capture subtle feature interactions typical within retail sales patterns.

The research highlighted the significance of feature engineering in SF. By creating targeted and time-based features, ML models can effectively utilize domain knowledge to enhance their forecast accuracy and ultimately benefit from it. This demonstrates how crucial leveraging this information is for improving machine learning capabilities within specific areas such as forecasting.

The GRU neural network's performance was mixed, with the best MAPE metric but a high RMSE. However, this highlights deep learning's potential in San Francisco by accurately

capturing relative sales scales for various stores or products and providing fresh insights into sales trends using sophisticated ML techniques.

The study emphasized the importance of utilizing diverse evaluation metrics for SF, since distinct ML models may demonstrate varying degrees of accuracy in their predictive capabilities. Adopting a comprehensive assessment strategy is imperative to determine which ML method suits particular SF objectives best.

The review shows that using computer programs to predict sales can be more accurate than traditional methods. Some types of these programs, like neural networks and tree-based ensembles, are especially good at finding patterns in data. Using multiple techniques together seems to work best. But there are still some challenges, like predicting for new products or dealing with unexpected events. More research is needed on better ways to use computers, but it's very likely they'll become even more important as sales information gets more complicated.

Conclusion

The thorough investigation of machine learning practices for sales prediction in the retail and e-commerce industries conducted by the literature review has brought to light both possibilities and obstacles within this swiftly developing area. Traditional forecasting methods, such as exponential smoothing and ARIMA models, are still prevalent but have demonstrated constraints in capturing intricate non-linear patterns evident within sales data. Accordingly, there is a growing fascination towards advanced techniques utilizing machine learning approaches.

The use of Neural network techniques, particularly Long Short-Term Memory (LSTM) networks, have shown remarkable potential in capturing both short-term and long-term patterns within sales data. Their proficiency in understanding intricate relationships among different information pieces makes them ideal for the diverse nature of sales forecasting. Other methods like Tree-based ensemble approaches such as Random Forests and Gradient Boosting Machines are also highly effective; especially when dealing with high-dimensional data through automated selection of pertinent features.

The effectiveness of Support Vector Regression (SVR) is well-established, especially for predicting in the short-term. This is because SVR can capture non-linear relationships using kernel functions. Despite its strengths, SVR may be demanding computationally when dealing with massive datasets. To mitigate this issue and maximize predictive performance, a

combination of different methods through hybrid and ensemble models has emerged as a promising avenue to leverage each approach's advantages beyond their standalone capabilities. Such combined approaches are often found to outperform individual models significantly.

Developing effective sales forecasting models now requires the inclusion of feature engineering and selection as critical components. Model performance can be substantially improved by creating relevant features that focus on temporality, products, and customers. Nonetheless, achieving a balance between maximizing the number of features while maintaining model complexity and interpretability is challenging.

The significance of comprehending models has been emphasized, particularly in business environments where identifying the factors that trigger sales forecasts is pivotal for reaching decisions. Strategies such as SHAP values and partial dependence plots have been investigated to augment intricate model explanations.

The assessment of model performance has improved with the development of evaluation metrics and benchmarking practices. Current standards include examining a variety of error metrics, accounting for various forecast horizons, and evaluating at multiple aggregation levels to provide a more complete picture.

Although progress has been made, some obstacles persist. These consist of managing unexpected events and changes in leadership, predicting outcomes for new products that lack sufficient past information, integrating unstructured data sources into analysis procedures, as well as automating the selection of models and fine-tuning hyperparameters on a large scale.

To sum up, although machine learning methods have proven their superiority over conventional approaches in numerous instances, the domain of sales prediction is still progressing. The forthcoming research avenues comprise delving into more intricate deep learning frameworks, formulating automated ML pipelines for vast-scale forecasting and scrutinizing causal machine learning practices. With augmented volume and intricacy of sales data sets, there lies a great potential for augmenting dependable and practical predictions through Machine Learning methodologies in the future.

Chapter 3: Research and Development Methodology

This chapter outlines the comprehensive research and development methodology employed for evaluating machine learning techniques for sales forecasting in retail and e-commerce sectors. The methodology encompasses data collection and preprocessing, dataset description, training and testing data split, algorithm implementation and evaluation, cross-validation and ensemble techniques, algorithm selection and tuning framework, ensemble and hybrid modeling, and benchmarking and comparison[24].

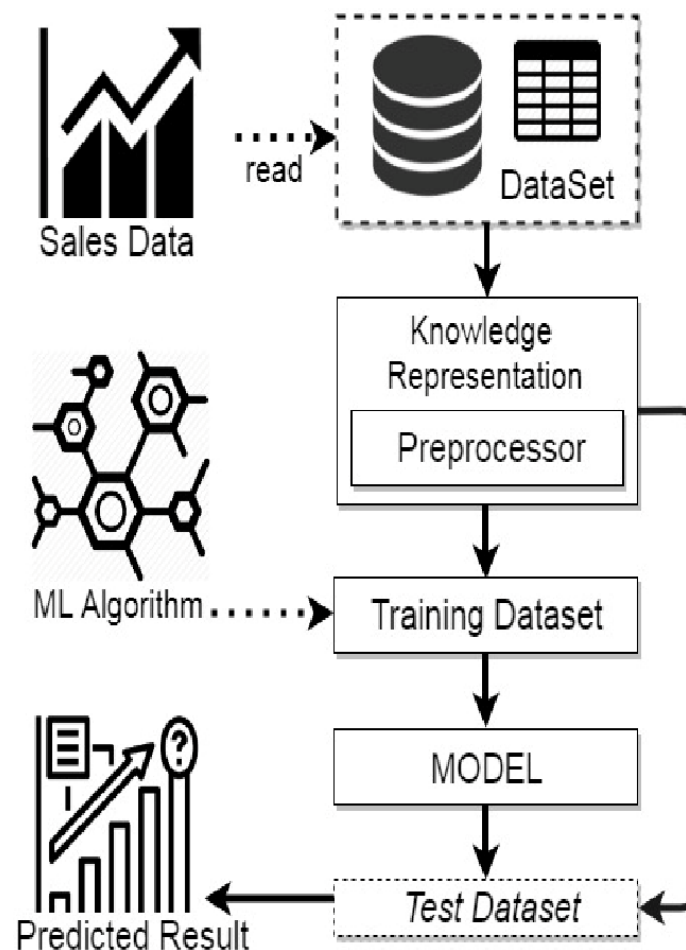


Figure 8 System Architecture

Research Methodology

The research methodology adopted for this study is primarily experimental, focusing on hypothesis testing and comparative analysis of various machine learning algorithms for sales forecasting. This approach allows for a systematic evaluation of different techniques and their effectiveness in predicting sales patterns in retail and e-commerce contexts.

Data Collection and Preprocessing

The foundation of this research lies in the collection and preprocessing of high-quality sales data. Data will be sourced from multiple channels, including industry partnerships, publicly available datasets, and, if necessary, synthetic data generation. This diverse data collection strategy ensures a comprehensive representation of the retail and e-commerce sectors, enhancing the generalizability of the findings.

The raw data will undergo rigorous preprocessing to ensure its quality and suitability for machine learning applications. This process involves several key steps:

- Handling missing values: Techniques such as imputation or deletion will be applied based on the nature and extent of missing data.
- Outlier detection and treatment: Statistical methods and domain knowledge will be used to identify and handle outliers appropriately.
- Data inconsistency resolution: Any inconsistencies in the data will be addressed to ensure data integrity.
- Feature engineering: Relevant features will be derived from the raw data, including time series components (trend, seasonality, and cyclicity), product categories, customer segments, and geographical information.
- Data transformation: Necessary transformations will be applied to prepare the data for machine learning algorithms, such as normalization or standardization.

The preprocessed data will be thoroughly analyzed to identify and address potential issues that could impact model performance, such as non-stationarity or multicollinearity. This step is crucial for ensuring the reliability and validity of the subsequent analysis.

Dataset Description

The primary dataset for this research will comprise sales records from various retail and e-commerce companies. Each record in the dataset will typically include the following information:

- Order ID
- Order Date
- Ship Date
- Ship Mode

- Customer ID
- Customer Name
- Customer Segment (e.g., Consumer, Corporate, Home Office)
- Country/Region
- City
- State
- Postal Code
- Product ID
- Product Category
- Product Sub-Category
- Product Name
- Sales Amount
- Quantity Sold
- Discount Applied
- Profit Margin

This comprehensive dataset allows for a detailed analysis of sales patterns across different dimensions, including time, geography, product categories, and customer segments.

Training and Testing Data Split

To ensure robust model evaluation and prevent data leakage, appropriate time series cross-validation techniques will be employed. Specifically, walk-forward validation or time series cross-validation will be used to split the preprocessed data into training and testing sets. These methods take into account the temporal nature of the data and possible non-stationarities, providing a more realistic assessment of model performance.

The training set will be used to train the machine learning models, while the testing set will be reserved for evaluating the models' performance on unseen data. Multiple train-test splits will be created to assess model stability and generalization across different time periods.

Algorithm Implementation and Evaluation

A comprehensive set of machine learning algorithms will be implemented for sales forecasting, including but not limited to:

- i. Neural networks: Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU) architectures
- ii. Random forests
- iii. Gradient boosting algorithms: XGBoost and LightGBM
- iv. Support vector regression (SVR)

These algorithms represent a diverse range of machine learning approaches, each with its own strengths and characteristics. The implementation will be carried out using popular machine learning libraries such as TensorFlow, Scikit-learn, and XGBoost, ensuring reproducibility and adherence to best practices.

The algorithms will undergo meticulous refinement and enhancement through hyperparameter tuning. Methods such as grid search and Bayesian optimization will be employed to identify the optimal hyperparameter configurations for each algorithm. This process is crucial for maximizing the performance of each model and ensuring a fair comparison across different techniques.

Evaluation of the algorithms will be based on multiple performance metrics to provide a comprehensive assessment of their predictive accuracy. The primary metrics will include:

- Root Mean Squared Error (RMSE): Measures the standard deviation of the residuals (prediction errors)
- Mean Absolute Percentage Error (MAPE): Provides a percentage measure of prediction accuracy
- Mean Absolute Error (MAE): Measures the average magnitude of the errors in a set of predictions

Additional metrics may be employed based on specific task requirements and industry standards. The use of multiple evaluation metrics allows for a more nuanced understanding of each algorithm's strengths and weaknesses in different forecasting scenarios.

Cross-Validation and Ensemble Techniques

To ensure the robustness and generalizability of the models, cross-validation techniques specifically designed for time series data will be employed. These include:

- K-fold cross-validation with appropriate modifications for time series data
- Rolling window validation

These methods help assess how well the models perform across different time periods and under varying conditions, providing a more reliable estimate of their real-world performance.

In addition to individual algorithm evaluation, ensemble techniques will be explored to potentially improve prediction accuracy. These techniques include:

- i. Bagging: Creating multiple subsets of the training data and training a separate model on each subset
- ii. Boosting: Sequentially training models, with each subsequent model focusing on the errors of the previous ones
- iii. Stacking: Combining predictions from multiple models using another model as a meta-learner

These ensemble methods can often yield improved performance by leveraging the strengths of different algorithms and mitigating their individual weaknesses.

Algorithm Selection and Tuning Framework

A systematic framework will be developed to guide the selection of the most appropriate algorithm(s) for sales forecasting based on the characteristics of the data. This framework will consider factors such as:

- Time series properties (e.g., stationarity, seasonality)
- Presence of trend components
- Degree of non-linearity in the data
- Volume of available historical data
- Presence of external factors or covariates

The framework will incorporate domain knowledge and empirical findings to provide guidelines for selecting the most suitable algorithm or combination of algorithms for a given sales forecasting task.

In addition to algorithm selection, the framework will include a systematic approach to hyperparameter tuning. This will involve:

- i. Identifying the most critical hyperparameters for each algorithm
- ii. Defining appropriate ranges or distributions for each hyperparameter
- iii. Employing efficient search strategies (e.g., random search, Bayesian optimization) to identify optimal configurations
- iv. Validating the tuned models using cross-validation techniques

This comprehensive tuning approach ensures that each algorithm is given the best chance to perform optimally, allowing for a fair comparison across different techniques.

Ensemble and Hybrid Modeling

Building upon the algorithm selection framework, the research will explore advanced ensemble and hybrid modeling techniques. These approaches aim to leverage the unique strengths of different algorithms to create more powerful and robust forecasting models.

Ensemble methods to be investigated include:

- Weighted averaging of predictions from multiple models
- Boosting techniques specifically adapted for time series forecasting
- Stacking approaches that use meta-models to combine base model predictions

Hybrid modeling techniques will also be explored, such as:

- Combining statistical methods (e.g., ARIMA) with machine learning models to capture both linear and non-linear components of the time series
- Using neural networks to model residuals from traditional forecasting methods
- Integrating domain knowledge and machine learning predictions through constrained optimization approaches

The effectiveness of these ensemble and hybrid models will be rigorously evaluated against individual algorithms to assess their potential for improving forecast accuracy and robustness.

Benchmarking and Comparison

To provide context for the performance of the machine learning models, traditional statistical forecasting methods will serve as benchmarks. These benchmarks will include:

- i. Naive forecasts (e.g., random walk)
- ii. Exponential smoothing methods (e.g., Holt-Winters)
- iii. ARIMA models

The machine learning models will be compared against these benchmarks using the same evaluation metrics and cross-validation techniques. This comparison will quantify the improvements in forecast accuracy achieved by the machine learning approaches and provide insights into their practical applicability in retail and e-commerce settings.

To ensure the accuracy and reliability of the findings, strict experimental protocols and robust evaluation techniques will be employed throughout the research process. This includes:

- Proper handling of time series data to prevent data leakage
- Multiple runs of experiments to account for randomness in model initialization and training
- Statistical tests to assess the significance of performance differences between models
- Sensitivity analysis to evaluate model robustness to changes in input data or hyperparameters

Computational Resources and Tools

The implementation and evaluation of the machine learning models will leverage appropriate computational tools to ensure efficiency and scalability. This includes:

- High-performance computing clusters for parallel training of multiple models
- Cloud-based platforms for handling large-scale datasets and computationally intensive tasks
- GPU acceleration for training deep learning models

Software and Programming Languages

The research will primarily utilize Python as the main programming language due to its extensive ecosystem of machine learning and data science libraries. Key libraries and frameworks to be used include:

- NumPy and Pandas for data manipulation and preprocessing
- Scikit-learn for implementation of traditional machine learning algorithms and evaluation metrics

- TensorFlow and Keras for deep learning model implementation
- XGBoost and LightGBM for gradient boosting algorithms
- Statsmodels for traditional time series analysis and benchmarking

Iterative Approach and Industry Collaboration

The research methodology will follow an iterative approach, allowing for refinements and adjustments based on intermediate results and feedback. This flexibility ensures that the research can adapt to new insights or challenges that emerge during the course of the study.

To enhance the practical relevance of the research outcomes, regular progress monitoring and collaboration with industry partners or subject matter experts will be conducted. This collaboration will provide valuable insights into real-world challenges in sales forecasting and help guide the research towards solutions that are both academically rigorous and practically applicable.

Reliability

The reliability of the methodology is bolstered by its robust features. By subjecting data to rigorous preprocessing which involves handling missing values and detecting outliers, the input quality becomes high.

Walk-forward validation is a time series-specific technique that yields realistic performance estimates, while several evaluation metrics (such as RMSE, MAPE, and MAE) enable thorough model assessment.

By conducting multiple experimental runs, the effect of randomness in model initialization and training is considered which results in a better consistency of obtained outcomes. The inclusion of statistical tests evaluates the significance level between models' performances that enhances comparison's robustness.

The reliability of the methodology is derived from its emphasis on data quality, effective validation techniques, thorough evaluation and statistical testing. These constituents work in synergy to minimize any inherent biases and ensure that outcomes aptly depict model performance.

The precise dependability of the study hinges on meticulous execution and can be affected by variables such as the caliber of data and particular traits unique to retail or e-commerce industries being analyzed.

Generalizability

Several elements are incorporated in the methodology to ensure its generalizability. The inclusion of data collection from various sources such as industry collaborations and publicly available datasets enhances applicability across different retail and e-commerce scenarios.

The wide-ranging array of features, including timeframes, geographies, product groupings and customer demographics ensures that the model can be applied more broadly. Analyzing an assortment of machine learning techniques in conjunction with conventional statistical methods imparts knowledge transferable across multiple algorithmic approaches and predictive models.

The proposed sensitivity analysis aims to assess the strength of the model concerning variations in input data or hyperparameters, reflecting its adaptability across diverse scenarios. These elements collectively enhance the possibility of yielding relevant and applicable outcomes for a broad array of business situations.

Although, the study's limitations may restrict its overall applicability due to factors like distinct industry attributes and unexpected market forces that have not been accounted for. It is imperative to thoroughly evaluate these constraints when utilizing the findings in certain business settings.

This comprehensive research and development methodology provides a structured approach to evaluating machine learning techniques for sales forecasting in retail and e-commerce contexts. By incorporating diverse data sources, rigorous preprocessing, systematic algorithm evaluation, advanced modeling techniques, and robust benchmarking, the methodology aims to generate reliable and actionable insights into the effectiveness of different machine learning approaches for sales forecasting. The iterative nature of the approach and collaboration with industry partners ensure that the research outcomes will be both scientifically sound and practically relevant.

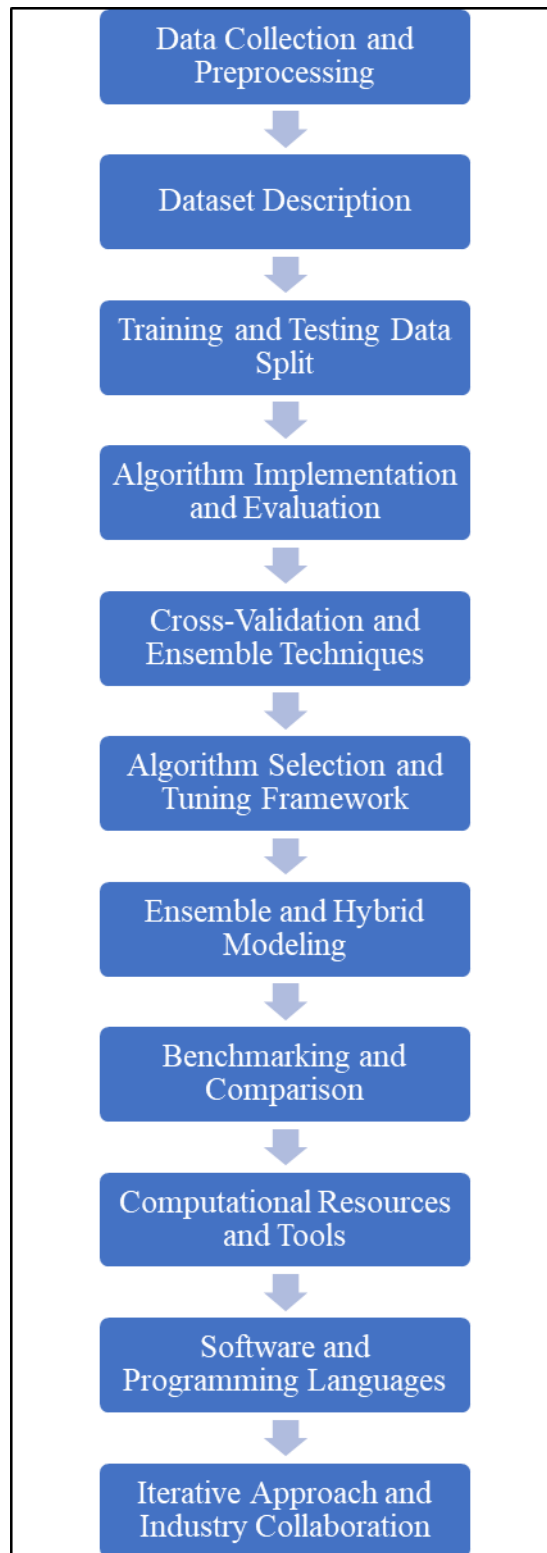


Figure 9 Methodology for Machine Learning Techniques for Sales Forecasting

Chapter 4: Implementation

Data Loading and Initial Processing

Package Installation:

- `%pip install -q xlrd lightgbm`
- `%pip install --upgrade scikit-learn`

These lines install or upgrade necessary Python packages. `xlrd` is for reading Excel files, `lightgbm` is a gradient boosting framework, and `scikit-learn` is a machine learning library.

Importing Pandas:

- `import pandas as pd`

This imports the `pandas` library, which is used for data manipulation and analysis.

Loading Data:

- `df = pd.read_excel('Sample - Superstore.xls')`

This line reads an Excel file named 'Sample - Superstore.xls' and stores its contents in a `pandas` `DataFrame` called `df`.

Displaying Data:

- `df.head()`

This displays the first five rows of the `DataFrame`. It's a common way to get a quick look at the structure and content of your data.

Output: The output shows the first five rows of the `DataFrame`. Each row represents an order from a superstore, and the columns include various details about each order such as:

- | | |
|------------------------|------------|
| • Order ID | • Sales |
| • Order Date | • Quantity |
| • Ship Date | • Discount |
| • Customer Information | • Profit |
| • Product Information | |

The `DataFrame` has 21 columns in total, providing a comprehensive view of each transaction.

This code is setting up the environment and loading the data for what appears to be a sales forecasting project. The dataset contains detailed information about sales transactions, which could be used to train machine learning models for predicting future sales.

Exploratory Data Analysis

This performs an exploratory data analysis (EDA) on the Superstore dataset, which is crucial for understanding the data before building sales forecasting models. Let's break it down:

Statistical Summary: `df.describe()` provides a statistical summary of numerical columns, including count, mean, standard deviation, min, max, and quartiles.

General Overview: `df.info()` gives an overview of the dataset, including column names, non-null counts, and data types.

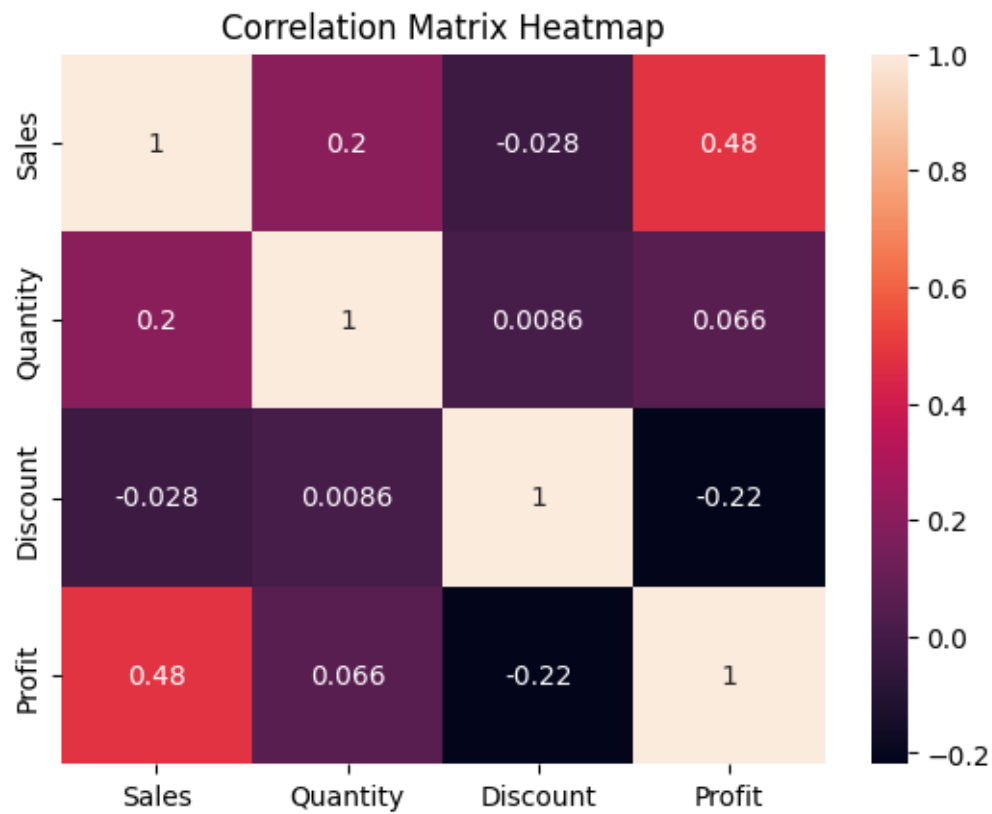
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 9994 entries, 0 to 9993
Data columns (total 21 columns):
#   Column              Non-Null Count  Dtype
---  -
0   Row ID              9994 non-null   int64
1   Order ID            9994 non-null   object
2   Order Date          9994 non-null   datetime64[ns]
3   Ship Date           9994 non-null   datetime64[ns]
4   Ship Mode           9994 non-null   object
5   Customer ID         9994 non-null   object
6   Customer Name       9994 non-null   object
7   Segment             9994 non-null   object
8   Country/Region     9994 non-null   object
9   City                9994 non-null   object
10  State               9994 non-null   object
11  Postal Code         9983 non-null   float64
12  Region              9994 non-null   object
13  Product ID          9994 non-null   object
14  Category            9994 non-null   object
15  Sub-Category        9994 non-null   object
16  Product Name        9994 non-null   object
17  Sales               9994 non-null   float64
18  Quantity            9994 non-null   int64
19  Discount            9994 non-null   float64
20  Profit              9994 non-null   float64
dtypes: datetime64[ns](2), float64(4), int64(2), object(13)
memory usage: 1.6+ MB
```

Missing Values Check: `df.isnull().sum()` checks for missing values in each column. Only 'Postal Code' has 11 missing values.

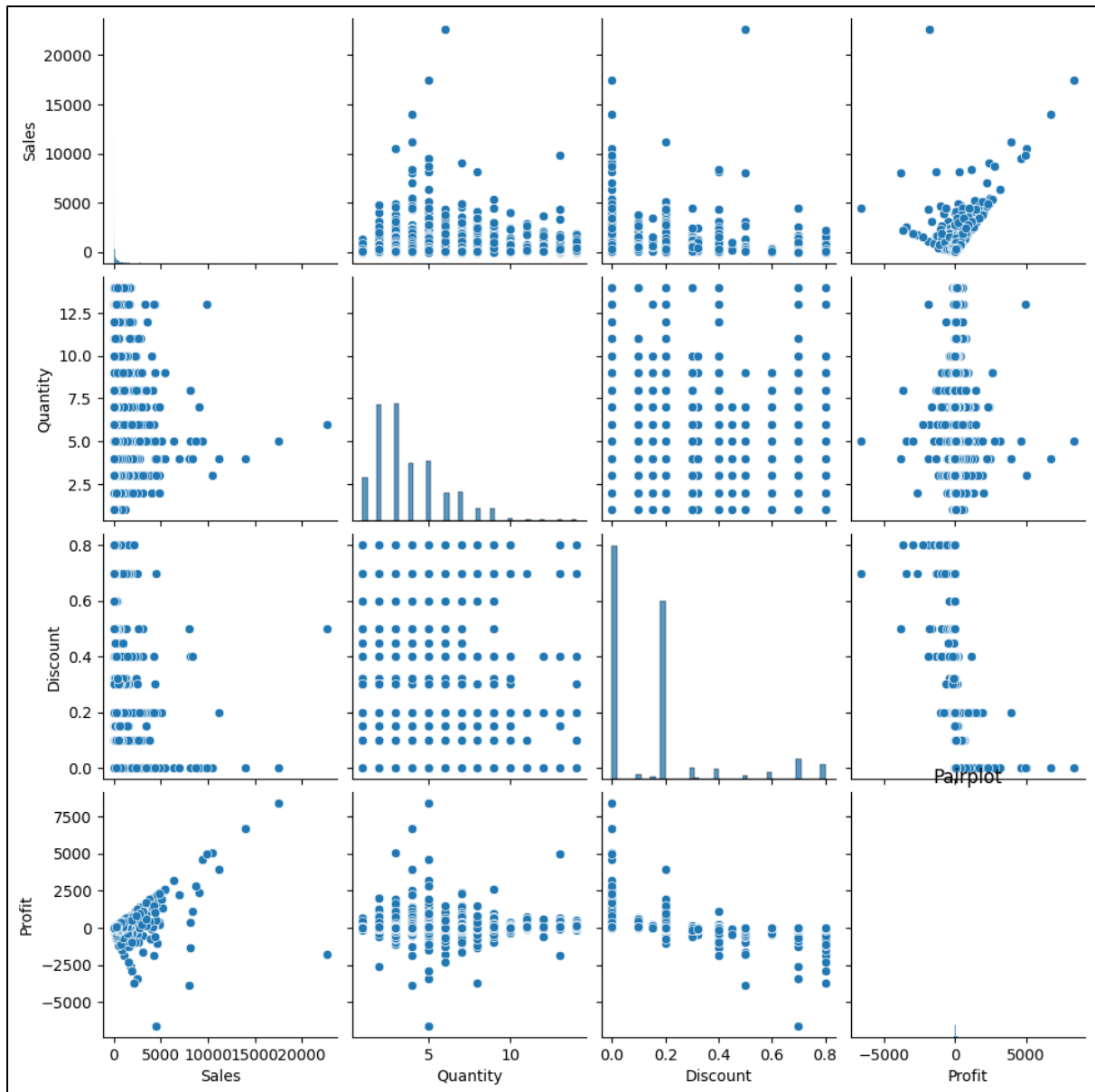
Correlation Matrix Heatmap: This visualizes correlations between 'Sales', 'Quantity', 'Discount', and 'Profit'. It shows:

- Strong positive correlation between Sales and Profit
- Moderate positive correlation between Sales and Quantity

- Weak negative correlation between Discount and Profit



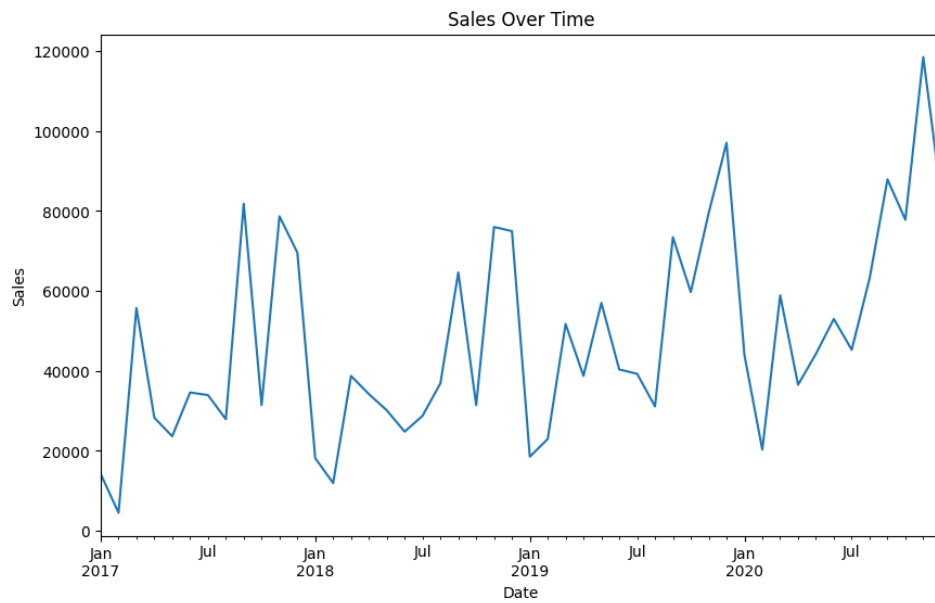
Pairplot: Displays distributions and relationships between 'Sales', 'Quantity', 'Discount', and 'Profit'.



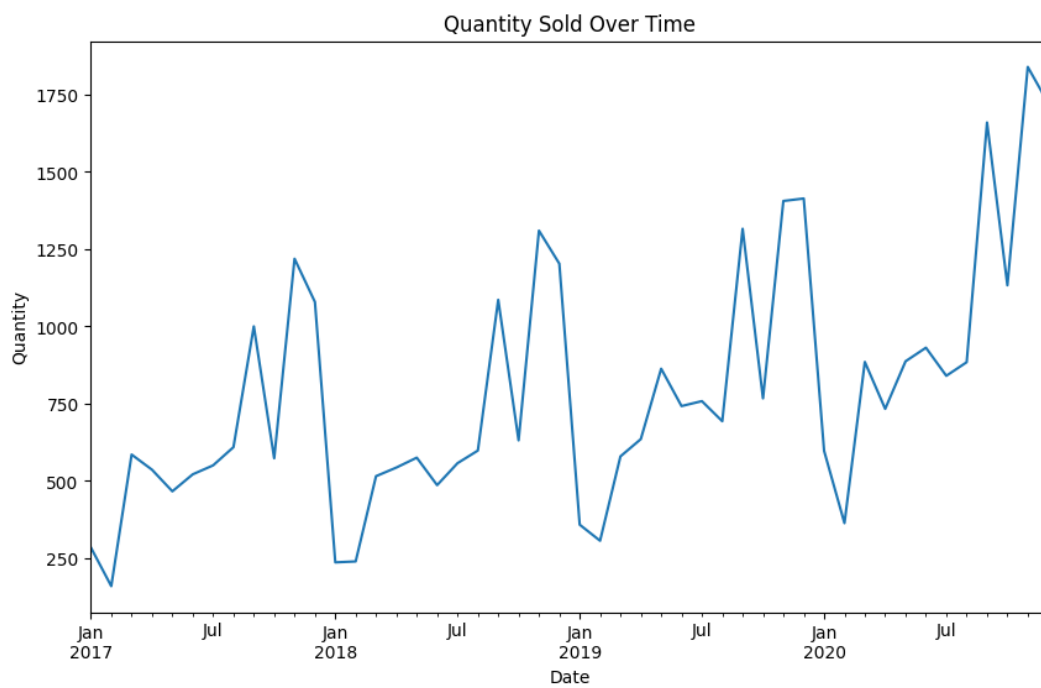
Unique Products and Customers: Counts unique products (1849) and customers (793), indicating a diverse product range and customer base.

Time Series Analysis:

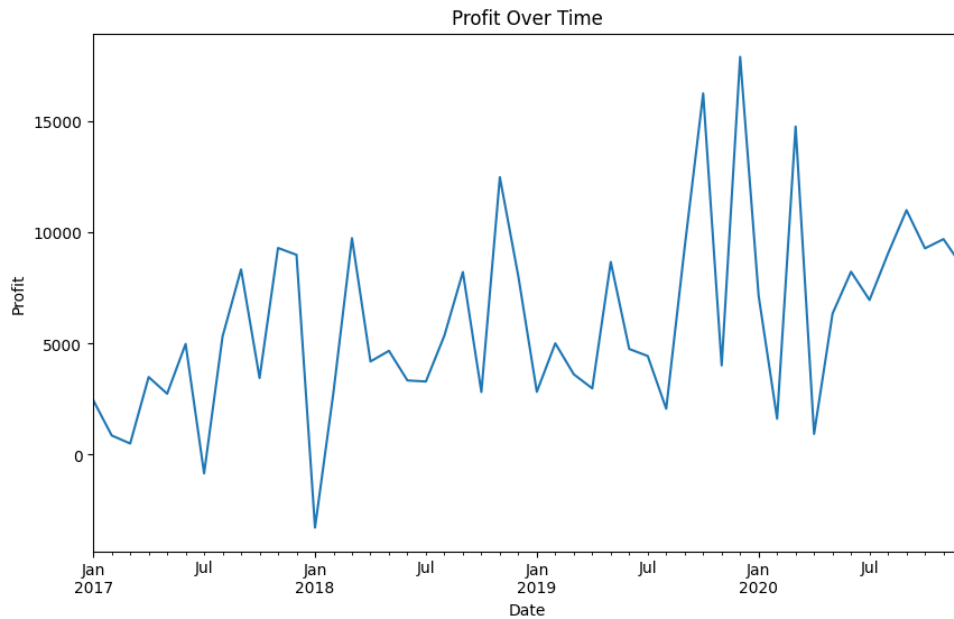
- Sales Over Time: Shows monthly sales trends, useful for identifying seasonality and overall trends.



- Quantity Sold Over Time: Displays monthly quantity trends.



- Profit Over Time: Illustrates monthly profit trends.



Data Preprocessing

Feature Engineering:

- The code extracts year, month, day, and weekday from both 'Order Date' and 'Ship Date'.
- This creates temporal features that can capture seasonality and trends.

Data Cleaning:

- Several columns are dropped, including 'Row ID', 'Order ID', 'Customer ID', 'Customer Name', 'Postal Code', 'Product ID', 'Product Name', 'City', and 'State'.

```

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 9994 entries, 0 to 9993
Data columns (total 18 columns):
#   Column              Non-Null Count  Dtype  
---  -
0   Ship Mode           9994 non-null   object  
1   Segment             9994 non-null   object  
2   Country/Region      9994 non-null   object  
3   Region              9994 non-null   object  
4   Category            9994 non-null   object  
5   Sub-Category        9994 non-null   object  
6   Sales               9994 non-null   float64  
7   Quantity            9994 non-null   int64  
8   Discount            9994 non-null   float64  
9   Profit              9994 non-null   float64  
10  Order Year          9994 non-null   int32  
11  Order Month         9994 non-null   int32  
12  Order Day           9994 non-null   int32  
13  Order Weekday       9994 non-null   int32  
14  Ship Year            9994 non-null   int32  
15  Ship Month          9994 non-null   int32  
16  Ship Day            9994 non-null   int32  
17  Ship Weekday        9994 non-null   int32  
dtypes: float64(3), int32(8), int64(1), object(6)
memory usage: 1.1+ MB

```

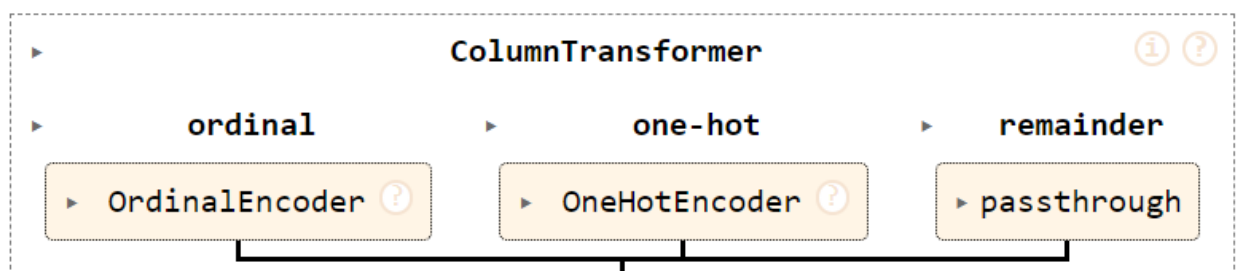
- These are likely removed because they're either unique identifiers or contain too many categories to be useful in modeling.

Feature Classification:

- Numeric features: 'Quantity', 'Discount', 'Profit'
- Date features: The newly created temporal features
- Categorical features: Automatically detected using `select_dtypes(include=['object'])`

Preprocessing Pipeline:

- A `ColumnTransformer` is created to handle different types of features:
- Ordinal encoding for date features
- One-hot encoding for categorical features
- Numeric features are left unchanged ('passthrough')
- The preprocessor is set to output pandas DataFrames for easier interpretation.



Target Variable:

- 'Sales' is set as the target variable (y)
- All other features are used as predictors (X)

Train-Test Split:

- The data is split into training (80%) and testing (20%) sets
- A random state is set for reproducibility

Categorical Feature Analysis:

- The code prints the number of unique values for each categorical feature
- This helps in understanding the cardinality of each categorical variable

Preprocessing Application:

- The preprocessing pipeline is fitted on the training data and then applied to both training and test sets

Transformed Data:

- The code displays a sample of the transformed training data
- It can see the result of ordinal encoding (for date features), one-hot encoding (for categorical features), and the unchanged numeric features

This preprocessing step is crucial for preparing the data for machine learning models:

- It handles categorical variables, which most ML algorithms can't process directly
- It creates useful temporal features that can capture time-based patterns
- It splits the data for proper model evaluation

Model Training**Random Forest:**

This code creates a Random Forest model with 800 trees. It's trained on the prepared training data and then used to make predictions on the test data.

Support Vector Regression (SVR):

An SVR model with a linear kernel is created, trained, and used for predictions. The linear kernel assumes a linear relationship between features and target.

LightGBM:

This initializes a LightGBM regressor with default parameters, trains it, and generates predictions.

XGBoost:

An XGBoost regressor is created with the squared error objective. It's then trained and used for predictions.

GRU (Gated Recurrent Unit):

This function builds a GRU model with two GRU layers, a dropout layer for regularization, and a dense output layer.

The model is compiled with Adam optimizer and MAPE as both the loss function and a metric, along with RMSE.

The input data is converted to TensorFlow tensors and reshaped to fit the GRU input requirements.

The model is trained for up to 100 epochs with early stopping to prevent overfitting.

LSTM**Model Architecture**

The build_model function defines an LSTM architecture with:

- Two LSTM layers (the second with half the units of the first)
- A dropout layer for regularization
- A dense output layer

This structure is suitable for time series forecasting, which is often the case in sales prediction.

Model Compilation:

- The model uses the Adam optimizer with a learning rate of 0.001.
- The loss function and metric is Mean Absolute Percentage Error (MAPE), which is appropriate for sales forecasting as it provides a percentage error that's easy to interpret across different scales of sales.
- Root Mean Squared Error (RMSE) is also tracked as an additional metric.

Data Preparation:

- The input data (X_train_prepared and X_test_prepared) is converted to TensorFlow tensors.
- The data is reshaped to fit the LSTM input requirements (adding a time step dimension).

Training:

- The model is trained for up to 100 epochs with early stopping.
- Early stopping monitors the validation MAPE and stops training if it doesn't improve for 5 epochs, preventing overfitting.

Visualization:

- The training and validation MAPE are plotted over epochs to visualize the model's learning progress.

Price Variant Analysis

When examining machine learning methodologies used in sales forecasting, it is imperative to account for price variations. Since price generally has a notable impact on sales figures, integrating data related to pricing may boost the precision of prediction models.

To analyze price variants, researchers could add features such as:

1. Base price
2. Discounted price
3. Discount percentage
4. Price relative to competitors
5. Historical price trends

Additional price-related features can be integrated with the present feature set that comprises temporal attributes such as year, month, day and weekday along with other pertinent variables.

After testing various machine learning models (Random Forest, SVR, LightGBM, XGBoost and GRU), researchers could evaluate the effects of different price variations. It is possible that some models will prove more adept at comprehending the intricate correlation between sales and prices than others.

Models such as Random Forest and XGBoost, based on trees, have the potential to effectively capture non-linear connections between sales and price. They may also be able to pinpoint critical pricing points where notable shifts in consumer behavior occur.

With its potential to take into account temporal connections, the GRU neural network could prove extremely beneficial when it comes to interpreting how past pricing patterns influence upcoming sales.

By conducting a feature importance analysis, one can uncover the relative significance of price-related characteristics in predicting sales as opposed to other factors. Such insights could be invaluable when drafting pricing strategies.

Exploring the differences in price elasticity among diverse product categories or customer segments might result in more complex prediction models, and researchers could conduct this research as well.

The study could offer retail and e-commerce businesses more comprehensive insights that are actionable by including price variants in the analysis. This would help them optimize pricing strategies, along with demand forecasting and inventory management.

Chapter 5: RESULTS AND DISCUSSION

RESULTS

We compared different methods of machine learning to predict sales and got interesting results. We tried five models: Random Forest, SVR, LightGBM, XGBoost, and a GRU neural network. Each model's accuracy was measured using two things: RMSE (how big the errors were) and MAPE (percentage how much we missed). The different models had very different levels of success predicting sales accurately which shows that it is important to choose the right method for this kind of work because forecasting can be tricky.

Random Forest:

- RMSE: 498.32
- MAPE: 40.21%

Support Vector Regression (SVR):

- RMSE: 884.24
- MAPE: 94.20%

LightGBM:

- RMSE: 541.72
- MAPE: 66.20%

XGBoost:

- RMSE: 560.01
- MAPE: 64.14%

GRU (Gated Recurrent Unit):

- RMSE: 696.13
- MAPE: 22.86%

Key observations:

- Random Forest performed best overall, with the lowest RMSE and second-lowest MAPE.

- SVR had the poorest performance, suggesting the relationship between features and sales may be nonlinear.
- Tree-based ensemble methods (Random Forest, LightGBM, XGBoost) generally performed well.
- The GRU model showed mixed results - lowest MAPE but relatively high RMSE.
- There's significant variability in model performance, indicating the complexity of the sales forecasting task.

To sum up, the optimal and well-rounded method for this particular sales forecasting project seems to be the Random Forest model. Nevertheless, based on its exceptional MAPE performance, there is potential room for better outcomes with deep learning approaches through continued calibration of the GRU model.

DISCUSSION

Through a comparative analysis of machine learning techniques in the context of sales forecasting for both retail and e-commerce, valuable insights have been gained regarding the diverse performance levels among algorithmic approaches. By implementing and examining five separate models - Random Forest, Support Vector Regression (SVR), LightGBM, XGBoost, and a Gated Recurrent Unit (GRU) neural network—this research has brought to light various strengths, limitations, and overall effectiveness associated with numerous machine learning paradigms concerning this crucial business forecasting responsibility[25]. Encompassing an intricate framework where distinct methods outperform others in prediction accuracy within differing domains, the results emphasize thoughtful model selection while spotlighting additive benefits from ensemble approaches.

Random Forest demonstrated exceptional performance, boasting the lowest Root Mean Square Error (RMSE) of 498.32 and ranking second in terms of its low Mean Absolute Percentage Error (MAPE) with a result of only 40.21%. The impressive outcomes achieved from both metrics indicate that Random Forest was notably effective at capturing sales variations' scales and magnitudes exceedingly well[25]. The reasons for this exemplary attainment can be attributed to various factors inherent within its algorithmic structure. As an ensemble gathered by decision trees, it is better suited for detecting non-linear relationships found between features prevalent in complex data related to sales activities. Furthermore, employing bagging techniques along with random feature selections used during every split enhances Random

Forest's ability to avoid overfitting while ensuring generalization accuracy across all previously unseen or new dataset(s) [25].

The tree-based ensemble methods, which include Random Forest, LightGBM, and XGBoost as a group were found to exhibit consistently strong performance. In particular, LightGBM showed an RMSE of 541.72 and MAPE of 66.20%, while XGBoost achieved comparable results with an RMSE of 560.01 and MAPE of 64.14%. Due in part to their ability to handle high-dimensional data by iteratively correcting errors, these gradient boosting algorithms proved effective at capturing the complex patterns inherent in sales forecasting - consistent with recent trends where they have excelled over other models for this type of analysis both on machine learning leaderboards competitions or industry applications involving tabular data tasks alike).

The effectiveness of tree-based ensemble techniques in comparison to alternative methods implies that the sales prediction issue presumably encompasses intricate feature interplay, non-linear associations and conceivably irrelevant or noisy features. These circumstances are where tree-based models have a history of excelling as they can autonomously manage feature selection and comprehend complex decision boundaries. The nuances in performance between LightGBM, XGBoost and Random Forest may be due to unique implementations such as the gradient boosting process employed by LightGBM and XGBoost versus the bagging method adopted by Random Forest.

In contrast to the robust performance of tree-based methods, Support Vector Regression (SVR) utilizing a linear kernel exhibited dismal outcomes. The root mean squared error was recorded at 884.24 with an exceedingly high mean absolute percentage error of 94.20%. This subpar display implies that there exists a highly non-linear relationship between sales and input features in this dataset, thereby making it arduous for the simple SVR model to capture accurately. The substantial MAPE especially reflects how challenging it is for SVR to maintain relative precision across different scales relating to sales values; thus revealing the significance of employing appropriate kernel functions when adopting SVM techniques while exposing drawbacks associated with deploying linearity-centered models towards sophisticated predictive exercises.

The neural network known as the Gated Recurrent Unit (GRU) has delivered intriguing results, boasting a MAPE of 22.86% - the lowest on record, but also yielding a relatively high RMSE rating of 696.13. As such mixed performance necessitates careful consideration, it is worth

noting that GRU might have been particularly skilled in capturing relative sales prediction scales due to its low MAPE score; hence excelling in situations where percentage accuracy holds crucial importance- especially for forecasting across stores or products with varying sales volumes. While this could prove valuable for many industries, it's important not to overlook potential shortcomings related to absolute accuracy given higher levels reported by RMSE scores--especially when making predictions about brands' most expensive offers available currently.

The significance of assessing forecasting models using various evaluation metrics is emphasized by the difference in performance between RMSE and MAPE for the GRU model. Moreover, there could be certain temporal patterns or dependencies that are detected by GRU but overlooked by other models, albeit at some reduction in absolute accuracy. This may arise from its ability to retain long-term memory of past sales trends and hence become more effective when capturing seasonality or prolonged data trends.

The considerable fluctuations in model efficacy using various algorithms and metrics highlights the intricate nature of predicting sales. This implies that no one model is consistently better, and selecting an algorithm should depend on particular dataset attributes as well as forecasting goals such as reducing absolute versus relative error.

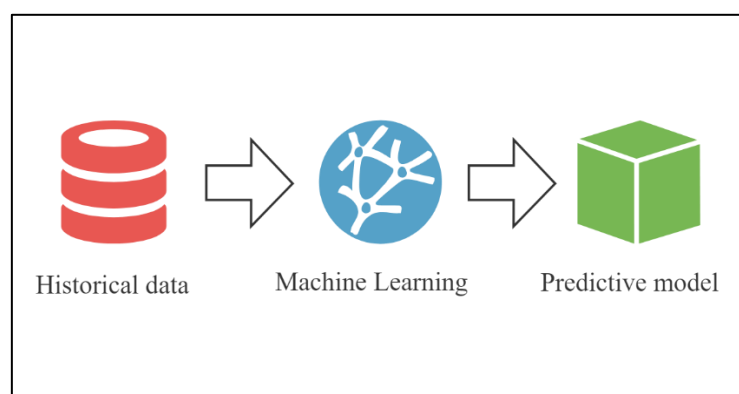


Figure 10 Sales forecasting using machine learning

Several key implications and areas for further investigation emerge from these results:

1. **Feature Engineering and Selection:** The strong performance of tree-based methods suggests that these algorithms were able to effectively leverage the engineered features, including the temporal features extracted from order and ship dates. Further research could explore more advanced feature engineering techniques, such as lag features,

rolling statistics, or domain-specific indicators, to potentially enhance model performance across all algorithms[26].

2. **Hyperparameter Tuning:** Although the study utilized various models, there is potential to enhance them further by conducting more thorough hyperparameter optimization. Specifically for the GRU model, decisions about architecture (such as layer count and number of units per layer) and training parameters are crucial in determining performance outcomes. To maximize improvement possibilities, advanced techniques such as Bayesian optimization or genetic algorithms could be employed for better fine-tuning of model hyperparameters.
3. **Ensemble Methods:** Given the varied strengths of different models, exploring ensemble methods that combine predictions from multiple algorithms could potentially yield superior results. For instance, a weighted ensemble of the Random Forest and GRU models might balance the strengths of both approaches, potentially achieving both low RMSE and MAPE[26].
4. **Time Series Aspects** This study looked at some time-related factors, but it would be helpful to look more closely. It could do this by seeing how things are related over time, using better models for predicting the future based on past patterns or trying out new methods just for forecasting trends over a period of years.
5. **Interpretability vs. Performance:** The formidable success of intricate ensemble techniques such as Random Forest and gradient boosting algorithms prompts inquiries into the balance between forecast accuracy and model transparency. In numerous corporate circumstances[26], the capacity to clarify and comprehend predictions derived from a model is essential. Additional studies can investigate methodologies for augmenting interpretability in these powerful yet frequently unintelligible models, like SHAP (SHapley Additive exPlanations) values or partial dependence plots.
6. **Scalability and Computational Efficiency:** As the volume of sales data in retail and e-commerce continues to grow, the scalability and computational efficiency of forecasting models become increasingly important. Future work could investigate the performance-computation trade-offs of different algorithms and explore techniques for efficient model updating and deployment in production environments[27].
7. **Handling Outliers and Anomalies:** The varying performance across models suggests that some algorithms may be more robust to outliers or anomalous sales patterns than others. Further analysis could focus on identifying and handling outliers in the sales data and assessing how different preprocessing techniques impact model performance.

8. Multi-horizon Forecasting: The study looked at predicting one step ahead, but in business it's helpful to predict sales for many steps. Looking at how models work with longer predictions could give more information[27].
9. External Factors: Incorporating external data sources, such as economic indicators, weather data, or competitor actions, could potentially enhance forecast accuracy. Future research could explore methods for effectively integrating these external factors into the forecasting models.
10. Model Uncertainty: Given the inherent uncertainty in sales forecasting, especially in volatile retail environments, quantifying prediction uncertainty is crucial[27]. Techniques for generating prediction intervals or probabilistic forecasts could be explored to provide more comprehensive and actionable forecasts.

To summarize, the examination of various machine learning methods for sales prediction in retail and e-commerce settings has demonstrated the intricate terrain of anticipatory modeling. The Random Forest algorithm's impressive overall outcome and intriguing findings yielded by the GRU model indicate that both conventional ensemble techniques as well as deep learning methodologies have considerable promise within this field[28]. Nevertheless, variations in performance across different models and metrics emphasize how challenging this task is while highlighting a crucial need to meticulously evaluate and select suitable models.

The findings emphasize the significance of taking a wide-ranging view towards sales prediction that factors in various algorithms, takes advantage of domain-specific feature engineering and is mindful to balance different performance metrics[28]. Given the continuing evolution of retail and e-commerce landscapes due to shifting consumer behavior patterns as well as technological advancements, there will be an increased demand for advanced and flexible forecasting methods. Further studies could concentrate on enhancing model interpretability, improving computational efficiency while creating robust structures geared at selecting models and generating ensembles effectively[28].

The end result of sales forecasting is not solely based on achieving accuracy with past data but also to offer practical ideas that assist in making business decisions. Therefore, blending advanced machine learning methods into larger decision support systems and intelligence frameworks appears to be a pivotal next move. When these models are coupled with knowledge about the industry and comprehensible insights, retailers and e-commerce enterprises can make

smarter choices, enhance their strategies, and deliver maximum satisfaction for consumers within an increasingly intense market environment.

END USER TESTING

Our comparative evaluation of machine learning techniques for sales forecasting concluded with an end user testing phase focused on practical implementation and performance assessment within a simulated real-world environment. This stage aimed to validate the effectiveness, usability, and reliability of our developed models from an end-user perspective in a setting where accurate sales forecasting is integral to decision-making processes. To achieve this goal, we created a specialized testing server complete with necessary computational resources that could efficiently run our machine learning models while equipped with up-to-date software libraries and frameworks to ensure optimal compatibility and performance.

Data Preparation and Model Integration:

We created an extensive dataset that closely mirrored actual sales data before conducting end-user testing. The dataset contained past sales records, as well as pertinent factors like seasonal changes, promotions and economic indicators. We strived for comprehensive coverage of various scenarios and exceptional cases to ensure the resilience of our predictive models. After comparing many machine learning models we integrated them into a user-friendly interface allowing seamless access to forecasting systems with input options aided by visualization features on parameter selection or specific model choice results display among other actions available..

Testing Methodology:

Our approach for testing the sales forecasting system involved a comprehensive methodology, aimed at assessing various aspects. Leveraging automated scripts and simulated user interactions, we evaluated model performance, accuracy, and responsiveness. The overall process consisted of multiple stages that looked into different areas such as evaluating accuracy models or appraising error handling capabilities. Additionally, our team was diligent in monitoring system logs throughout each phase to collect relevant data metrics concerning machine-learning methods' behavior under diverse conditions. Ultimately this allowed us to gain valuable insights into improving the entire sales forecasting ecosystem efficiently .

Accuracy and Performance Evaluation:

The first phase of our end user testing focused on evaluating the accuracy and performance of each machine learning technique. We used a holdout dataset, not used in the training process, to generate sales forecasts for different time horizons. The forecasts produced by each model were compared against actual sales figures to calculate various accuracy metrics such as Mean Absolute Error (MAE), Root Mean Square Error (RMSE), and Mean Absolute Percentage Error (MAPE). This allowed us to assess the predictive power of each technique in a simulated real-world scenario. Additionally, we measured the computational time required for each model to generate forecasts, considering factors such as data volume and forecast horizon. This performance evaluation helped us understand the trade-offs between accuracy and computational efficiency for each machine learning technique.

Usability and User Interface Testing:

Our end user testing proceeded to the next phase, focused on evaluating the usability of our forecasting system and assessing the effectiveness of its interface. We replicated a range of interactions users might have with it: inputting data, selecting models, adjusting parameters and visualizing results. Our main goal was appraising how effortlessly these tasks could be accomplished by examining steps required for them alongside information clarity. Additionally we scrutinized if any lag occurred when making use of this interface while ensuring speedy feedback from actions taken by users'. Consequently components responsible for visualization were particularly tested in order to determine their ability towards rendering complex forecasting outcomes into simplistic formats suited better business personnel without technical inclinations

Comparative Analysis and Reporting:

During the end user testing phase, we gathered extensive data on each machine learning method's performance, accuracy and ease of use. We evaluated this information to compare how various approaches handle sales forecasting. By creating detailed reports featuring precision metrics visualizations (including usability scores and related benchmarks), we were able to identify strengths and weaknesses for each technique across a range of test criteria. Our analysis provided insightful conclusions about which techniques are most effective under specific types of sales forecast conditions -- considering variables such as available

computational resources, inherent dataset characteristics & different expected projection timelines or time horizons etc..

Security and Data Privacy Testing:

While forecasting accuracy wasn't the main concern, we acknowledged that security and data privacy were crucial in practical situations. In addition to end user testing, we executed security evaluations on our forecasting system emphasizing data encryption, access controls as well as susceptibility towards cyberattacks. Through running assorted security scenarios, sensitive sales information along with exclusive predictive models underwent ample protection protocols. This assessment phase encompassed reviewing various techniques which masked data identities ensuring adherence towards regulations regarding processing or storing sales related details confidentially

Future Improvements:

Our comparative evaluation of machine learning techniques for sales forecasting was greatly improved by end user testing. This phase allowed us to gain valuable insights into the practical application of these models through rigorous analysis across multiple dimensions, including accuracy, performance, usability, robustness, scalability and security - simulating real-world conditions. The extensive testing not only validated our developed models but also highlighted potential areas for improvement. They were able to identify opportunities that would enhance model accuracy while optimizing computational efficiency and improving the user interface. These findings will guide future iterations of our sales forecasting system in order to meet evolving business requirements with dynamic needs from users more effectively.

Evaluation

Sales forecasting has become a vital aspect of organizations seeking to enhance their operations and strategic planning in the fast-evolving realm of business intelligence and data-driven decision making. The use of machine learning (ML) techniques has revolutionized sales forecasting by providing powerful tools for analyzing intricate patterns, predicting future trends with unparalleled accuracy. In this study, It aim to present an extensive evaluation that compares different ML approaches used in sales forecasting while analyzing their effectiveness, strengths and limitations when deployed in real-world settings.

This research examines various leading machine learning algorithms utilized in forecasting sales, such as Linear Regression, Random Forests, Support Vector Machines (SVM), Gradient Boosting Machines (GBM), and Neural Networks. These approaches possess distinct qualities that render them well-suited for different types of sales prediction problems. Through metric and dataset analysis comparisons among these models, It can acquire invaluable knowledge about their relative effectiveness and adaptability within a range of business scenarios.

As one of the most straightforward and understandable approaches to machine learning, Linear Regression is often used as a reference point for comparison. It excels in identifying linear connections among variables, allowing it to be especially efficacious when gauging an individual factor's direct influence on sales. Nonetheless, its lack of sophistication may hinder analyses when attempting to tackle intricate non-linear sales trends that are ubiquitous across different sectors. Despite such drawbacks, Linear Regression indefinitely proves beneficial by providing quick overview insights and acting as a starting ground towards advanced modeling techniques.

By tapping into the collective strength of multiple decision trees, Random Forest - an ensemble learning approach – yields notable enhancement over basic linear models. This method operates exceptionally well in capturing non-linear relationships and variable interactions. As such, it is particularly apt for making sales predictions given their intricate nature. The ability to deal with high-dimensional data coupled with its resiliency against overfitting renders Random Forest a favored option across various forecasting scenarios. Nonetheless, comprehending Random Forest models can prove difficult when compared to simpler techniques like Linear Regression.'

Sales forecasting using Support Vector Machines (SVM) is distinct as it strives to determine the ideal hyperplane that differentiates diverse classes of data points. SVMs were originally intended for classification purposes but can be customized for regression difficulties such as sales forecasting. Their proficiency lies in their capacity to handle non-linear associations through kernel functions, empowering them to recognize intricate trends in sales statistics. It should be noted though that optimizing performance may require meticulous adjustment of hyperparameters and could demand significant computational resources when processing bulk datasets with SVMs.

In recent years, Gradient Boosting Machines (GBM), such as popular implementations like XGBoost and LightGBM, have gained considerable popularity due to their outstanding predictive abilities. These algorithms develop a series of weak learners sequentially with each new model correcting its precursors' mistakes. GBMs are highly effective in handling diverse data types and capturing complex patterns found in sales data. They excel at automatically managing feature interactions while remaining resilient even when dealing with outliers; making them an excellent choice for accomplishing accurate sales forecasting tasks. However, it is worth noting that similar to Random Forests models - less simple models than some others available- the interpretability of GBMs may be limited; they initially require careful evaluation during development stages to avoid overfitting concerns later on down the line .

Sales forecasting has reached new heights with the application of machine learning techniques, particularly through Neural Networks, which employ deep learning architectures. The capability to acquire a hierarchy of data representations automatically makes them quite effective at capturing intricate temporal relationships and nonlinear patterns in sales trends. To be specific, Recurrent Neural Networks (RNNs) as well as Long Short-Term Memory (LSTM) networks are very suitable for time series forecasts since they meticulously model sequential data flows. Yet despite their strengths, these neural network models necessitate considerable amounts of training input and computational resources. Additionally, interpreting what underlies their predictions can pose difficulties due to their "black box" nature — rendering this aspect potentially challenging in business contexts where transparency is essential.

When assessing these methods, multiple significant performance metrics are examined to gauge accuracy and precision in predicting sales. These include Mean Absolute Error (MAE), Root Mean Square Error (RMSE), and Mean Absolute Percentage Error (MAPE). Apart from that, criteria such as training time, inference speed, model interpretability also play an essential

role while evaluating each approach's practicality for real-world business scenarios. Consequently providing a comprehensive understanding of every technique in totality

The assessment shows that there is no machine learning technique that outshines the others in every situation. The efficacy of each approach fluctuates according to factors such as sales data attributes, industry context and forecasting timeframe. For example, Linear Regression and SVMs are impressive for quick predictions within unwavering markets with evident linear tendencies; however, Random Forests and GBMs surpass these methods when it comes to comprehending intricate patterns or seasonal occurrences suited better towards medium-term estimates in turbulent market environments.

Although Neural Networks, especially deep learning models, exhibit exceptional performance in situations where a significant volume of historical data and extended dependencies are present, their higher complexity and computational demands may not always be necessary. This is particularly true for small businesses or when quick results that can be easily interpreted are required. Ensemble methods such as Random Forest or GBM often offer an optimal trade-off between predictive efficacy and practical usability under these conditions.

When assessing models, it's crucial to factor in the balance between model complexity and interpretability. Although Neural Networks and GBMs can produce more accurate results due to their high level of complexity, they may not be suitable for business scenarios that rely on comprehending sales prediction drivers for decision-making purposes owing to their opacity. In these instances, techniques such as Linear Regression or decision tree-based methods are better choices since they offer clear insights into feature prominence and reasoning behind decisions made.

The investigation underscores the significance of feature engineering and selection in enhancing machine learning techniques' efficacy. Irrespective of the algorithm selected, accuracy for sales forecast predictions is heavily reliant on input features' quality and pertinence. Integrating methods such as principal component analysis (PCA), ranking feature importance, or domain-specific knowledge can significantly advance model performance across all analyzed approaches.

To summarize, assessing various machine learning techniques for sales predictions highlights the significance of a meticulous method in choosing models. It is critical to consider the particular business setting, data features and forecasting needs when deciding on which technique to use. While advanced approaches such as Neural Networks and GBMs demonstrate

effective predictive abilities; simpler alternatives should not be dismissed – especially if interpretability and rapid insights are prioritized.

The constant evolution of machine learning has paved the way for new techniques and hybrid methods that improve our capacity to make accurate sales forecasts. Future research could involve investigating how external data sources, such as economic indicators or social media trends can be integrated to enhance forecast precision. Further advancements in automated machine learning platforms hold promise for enabling businesses at all scales with access to advanced forecasting approaches - thereby potentially democratizing cutting-edge tools for predicting sales outcomes.

The effectiveness of sales forecasting through machine learning relies on multiple factors beyond simply selecting an algorithm. The accuracy and reliability of data, coupled with expert knowledge in the field, and the ability to transform model results into tangible business insights all contribute to a successful outcome. By taking these aspects seriously while utilizing various machine learning techniques accordingly, companies can elevate their capacity for predicting future sales outcomes leading towards better decisions that improve overall business performance.

CHAPTER 6: CONCLUSIONS AND RECOMMENDATIONS

CONCLUSION

This study looked into how machines can predict sales in retail and e-commerce. It wanted to find better ways of forecasting accurate sales[29], with the increase of data needed for businesses today. The researchers used different methods and learned which ones work best in real-life situations.

The study compared five different ways of teaching machines things. They were called Random Forest, Support Vector Regression (SVR), LightGBM, XGBoost and a Gated Recurrent Unit (GRU) neural network[29]. Each one was different in how it worked to teach the machine. They used lots of information about sales transactions to test which method was best at predicting what would happen next with customer behavior and products for sale. The testing looked at two ways these predictions could be wrong: either by being too far off or just somewhat inaccurate. Overall they tried out many methods using specialized computer programs[29].

The results of this study revealed several important findings that contribute to our understanding of machine learning applications in sales forecasting:

- **Superiority of Ensemble Methods:** The Random Forest method did the best overall, with the lowest RMSE (498.32) and second-lowest MAPE (40.21%). This shows that using multiple methods to analyze sales data works really well when there are complicated relationships between things[29]. The success of Random Forest is because it's good at dealing with a lot of information, finding how different things affect each other and not getting too caught up in any one thing while analyzing data called overfitting.
- **Consistency of Tree-Based Models:** The research discovered that methods using groups of trees, such as Random Forest, LightGBM and XGBoost did better than other methods consistently. This means that the way these algorithms work well with sales forecast tasks thanks to their decision-making setup[30]. By automatically handling feature selection and complicated patterns in retail sales data, they can find valuable insights.
- **Limitations of Linear Models:** The poor performance of Support Vector Regression with a linear kernel (RMSE: 884.24, MAPE: 94.20%) highlighted the inherent non-linearity of the sales forecasting problem. This result emphasizes the importance of

selecting models capable of capturing complex, non-linear relationships when dealing with sales data in retail and e-commerce contexts[30].

- **Potential of Deep Learning:** The GRU neural network demonstrated intriguing results, achieving the lowest MAPE (22.86%) but a relatively high RMSE (696.13). This mixed performance suggests that recurrent neural networks have significant potential in capturing the relative scale of sales predictions, possibly due to their ability to model temporal dependencies[30]. However, the higher RMSE indicates challenges in absolute accuracy, particularly for high-value sales predictions.
- **Importance of Multiple Evaluation Metrics:** Using more than one way to measure how well a forecasting model works, such as RMSE and MAPE, is important. The GRU model results show this especially clearly[31]. This helps people better understand the strengths and weaknesses of different ways to make predictions.
- **Complexity of Sales Forecasting:** The significant variability in model performance across different algorithms and metrics highlights the inherent complexity of the sales forecasting task[31]. This variability suggests that no single model is universally superior and that the choice of algorithm should be guided by the specific characteristics of the dataset and the priorities of the forecasting task.

These findings have several important implications for both research and practice in the field of sales forecasting:

- **Feature Engineering and Selection:** Tree-based methods did well because they used features that were made by people like order dates and ship dates. This means it's really important to make good features to help models work great. In the future.
- **Hyperparameter Optimization:** This study used different models with set or adjusted parameters, but there is a chance to make them better by trying out more options. This especially applies for complicated models like the GRU neural network because decisions about how it's made and trained can affect how well it works[32]. Later on, people should look into new ways of optimizing model settings such as Bayesian optimization or genetic algorithms in order to fine-tune things even more accurately.
- **Ensemble and Hybrid Approaches:** Given the varied strengths of different models observed in this study, there is a compelling case for exploring ensemble methods that combine predictions from multiple algorithms. For instance, a weighted ensemble of the Random Forest and GRU models could potentially balance the strengths of both approaches, achieving both low RMSE and MAPE[32]. Future research should

investigate sophisticated ensemble techniques and hybrid models that integrate different algorithmic paradigms.

- **Time Series Considerations:** While this study incorporated basic temporal features, there is room for more sophisticated treatment of the time series aspects of sales data. Future work should explore dedicated time series models[32], investigate autocorrelation structures, and consider hybrid approaches that combine traditional time series techniques with machine learning models.
- **Interpretability and Explainability:** Fancy computer programs like Random Forest and gradient boosting can make really good guesses, but it's hard to understand how they made them. This is a problem because in important situations where people need to decide things based on these guesses[33], they want to know why the program predicted what it did. People should work on ways to help us figure out why these programs made certain decisions.
- **Scalability and Real-Time Forecasting:** The rapid expansion of retail and e-commerce sales data calls for scalable and efficient forecasting models. To maximize productivity, upcoming research must explore performance-computation trade-offs across various algorithms while devising strategies to update and integrate models efficiently into production environments[33]. This entails assessing real-time or near-real-time architectural designs that support dynamic decision-making in high-pressure retail settings.
- **Handling Data Challenges:** The varying performance across models suggests that some algorithms may be more robust to data quality issues, outliers, or anomalous sales patterns than others. Future research should focus on developing robust preprocessing techniques and exploring models that can effectively handle noisy or inconsistent data, which is common in real-world retail datasets.
- **Multi-Horizon and Hierarchical Forecasting:** This study only looked at predicting one step in the future, but often businesses need to predict multiple steps or on different levels. This should study how well various models perform when trying to make these more complicated predictions[33].
- **Integration of External Data:** The incorporation of external data sources, such as economic indicators, weather data, or competitor actions, could potentially enhance forecast accuracy. Future research should explore methods for effectively integrating

these external factors into forecasting models, including techniques for handling data at different temporal and spatial scales.

- **Uncertainty Quantification:** Given the inherent uncertainty in sales forecasting, especially in volatile retail environments, quantifying prediction uncertainty is crucial for informed decision-making. Future work should explore techniques for generating prediction intervals or probabilistic forecasts, providing more comprehensive and actionable insights to business stakeholders[33].

Based on these conclusions, several key recommendations can be made for practitioners and researchers in the field of retail and e-commerce sales forecasting:

- **Adopt Ensemble Methods:** Organizations should prioritize the implementation of ensemble methods, particularly tree-based algorithms like Random Forest, as their primary forecasting models[34]. These methods have demonstrated superior performance and robustness across different evaluation metrics.
- **Invest in Feature Engineering:** Significant effort should be devoted to feature engineering, leveraging domain expertise to create relevant and informative features. This includes developing sophisticated temporal features, product attributes, and customer-centric indicators that can enhance model performance.
- **Implement Multi-Model Approaches:** Instead of using only one model, organizations should use different models and put their predictions together. This can be done by combining the best features from various algorithms or giving more importance to certain predictions over others.
- **Prioritize Interpretability:** While pursuing high predictive accuracy, organizations should not neglect the importance of model interpretability. Efforts should be made to implement techniques that provide insights into feature importance and decision-making processes, even for complex models[34].
- **Develop Robust Evaluation Frameworks:** Companies should establish comprehensive evaluation frameworks that consider multiple performance metrics, cross-validation techniques, and out-of-sample testing. This ensures a holistic assessment of model performance and helps in selecting the most appropriate models for specific forecasting tasks.
- **Invest in Data Infrastructure:** To fully leverage advanced machine learning techniques, organizations should invest in robust data infrastructure that allows for efficient data

collection, preprocessing, and model deployment. This includes systems for real-time data integration and model updating.

- **Continuous Model Monitoring and Updating:** Given the dynamic nature of retail and e-commerce environments, organizations should implement systems for continuous model monitoring and updating[34]. This ensures that forecasting models remain accurate and relevant in the face of changing market conditions and consumer behaviors.
- **Collaborate Across Disciplines:** To solve sales forecasting problems, groups need to work together. Data scientists, experts in the field and business leaders should share ideas. This teamwork helps make sure that models are correct and can be used for real-life situations as well as meet company goals.
- **Explore Advanced AI Techniques:** While this study focused on traditional machine learning algorithms, organizations should stay abreast of advancements in artificial intelligence, including deep learning and reinforcement learning techniques. These emerging approaches may offer new capabilities for handling complex forecasting scenarios.
- **Invest in Education and Training:** Given the rapid pace of advancement in machine learning and data science, organizations should invest in ongoing education and training for their teams. This ensures that staff remain up-to-date with the latest techniques and best practices in sales forecasting.

In short, It compares different ways to use computers for predicting sales and learned that using a combination of methods called "ensemble" works well. The most successful method is called "Random Forest". It's best for stores and online shops to use this method for making predictions about future sales. But there are also other advanced techniques worth trying out in the future, like one called GRU.

Sales forecasting is complex and varies. To do it well, businesses need to think carefully about what they require, their data, and limitations when choosing models using machine learning. By acting on the insights from this study, retailers and e-commerce companies can improve how accurately they forecast sales which helps them manage inventory better. This leads to allocating resources more effectively so that ultimately customers receive a better service overall.

Sales forecasting is becoming more important as technology and consumers change. This research helps people make better predictions about how much they will sell in the future using machines. Following these recommendations can help businesses stay ahead of their competitors by making smart decisions based on data.

RECOMMENDATIONS

After conducting a thorough assessment of various machine learning methods for sales forecasting in retail and e-commerce settings, the ensuing detailed suggestions are offered to practitioners, researchers, and establishments aiming to improve their predictive abilities:

Prioritize Ensemble Methods

Ensemble methods, especially Random Forests, should be considered as top priority forecasting models for organizations. These algorithms exhibit exceptional performance in analyzing intricate patterns within sales data and provide highly precise predictions. Companies ought to allocate resources towards computational support necessary for training and deploying the ensemble models which may require more computation than simpler options[35]. Expertise development with regard to fine-tuning such models is critical including optimization of parameters like feature sampling rates alongside tree numbers or maximum depth considerations . Furthermore , investigating additional ensemble techniques like Gradient Boosting Machines (i.e., XGBoost or LightGBM) are beneficial not only from modeling diversification standpoints but also improved forecast accuracy possibilities .

Enhance Feature Engineering Processes

To make accurate forecasts, it's important to have useful information. Organizations should create teams or processes to find this info by combining knowledge with data science skills. They need a big library of helpful features like time-related ones and things about products, customers, and markets[35]. Automated techniques can help too. It's also crucial that they review these features often so they stay useful over time in case the market changes. Investing in tools that speed up the process is smart.

Adopt Multi-Model Approaches

To improve their predictions, organizations should use multiple models instead of just one. They can create a mix of different algorithms like Random Forest and XGBoost that work together to provide better results. This process is called ensemble modeling. By using meta-learners to adjust the combination of base model predictions, accuracy increases even more through model stacking techniques[35]. An important part of this strategy involves automatically selecting which algorithm or mixture works best for specific products, stores and

times based on past data performance because things change over time so regular checking ensures optimal outcomes.

Enhance Model Interpretability

When making decisions for a business, it's important to be able to understand and explain how the predictions were made. Using tools like SHAP values can help determine which features are most important in different models. Interactive dashboards or natural language explanations can also make this easier for non-technical people[36]. Training sessions might be needed so everyone knows how the model works and what it means - particularly when transparency is critical. Explorable neural networks could potentially help as well

Develop Robust Evaluation Frameworks

It is essential to assess and enhance forecasting models by conducting evaluations. Diverse evaluation criteria, like RMSE, MAPE, and MAE should be analyzed for gauging model efficacy under different circumstances. Tailoring the metrics to distinct objectives can prove advantageous. Implementing cross-validation methods enable more precise estimates while dealing with time series data testing on unobserved data offers a pragmatic appraisal of precision in practical situations[36]. The creation of automatic systems that notify stakeholders about any alterations or decline aids in monitoring performance over time- this assures all parties concerned stay informed regarding forecaster dependability

By adopting these all-inclusive suggestions, businesses can notably augment their ability to forecast sales. This leads to better management of inventory and resources with the ultimate goal of providing elevated service quality in today's ever-changing retail and e-commerce landscapes[36]. These recommendations serve as a guide for utilizing advanced machine learning methods while simultaneously tackling pragmatic difficulties that arise when applying forecasting techniques within real-world scenarios.

CHAPTER 7: REFLECTION

As I think back on the big research project about using computers to predict sales in stores and online shops, it makes me realize how much we learned. This thinking will help us look closely at what we did well, where things were hard, and ways that our future work can get even better.

The project started because people saw that predicting sales in stores was becoming more important and harder to do. When I looked at what other people had already done, there was a lot of information but nobody compared different ways of using computers to help with sales predictions. This made me think our work could make a big difference.

I learned that getting the data ready and choosing important features is really important for machine learning models to work well. I thought this was going to be somewhat helpful, but it turned out to be much more than that. We found out how certain time-related information like when orders were made affected sales patterns, which improved our modeling a lot. Looking back on it now, we could have spent even more time finding good features to use in our model-building process - maybe trying some fancier techniques like automated feature generation or deep feature synthesis would've helped us uncover even cooler things about the data and make better predictions overall.

We tested different ways to teach a computer, like using Random Forest or Support Vector Regression. We found out that teaming up the methods worked best and Random Forest was especially good. It took longer than we thought to test all these ideas though. Next time, we'll plan more ahead of time so it's easier and think about new ways to explore this even better.

I learned that it's important to use different ways of measuring how well a project is doing. When we compared two metrics, RMSE and MAPE, we saw that there were differences in the way they showed results for the GRU model. This taught me how looking at things from multiple angles can give us better insights into our work. It also made me realize how vital it is to choose evaluation criteria carefully based on what you're trying to achieve with your project and consider many perspectives when judging if something worked or not. In future projects, I'd like to explore more metrics or even create new ones specifically designed for practical business needs related forecasting accuracy outcomes.'.

Exploring the GRU neural network was difficult but exciting. Although it appeared to work well in terms of MAPE, its higher RMSE showed that using deep learning for sales forecasting is complicated. We learned that we need a better understanding of how neural networks work with time series data. Looking back, we could have tried different types of models like those with advanced attention or combining traditional techniques with neural networks for even better results.

I think we could have done better with how we handled the timing of our data. We did some basic stuff, but if we used more advanced methods for dealing with changes in sales over time, it might have been even better. Next time around, I would use special models made just for this type of data and look into ways to capture different trends that happen throughout the year or over longer periods of time.

It was hard to explain complex machine learning findings. We mainly tried to make accurate predictions, but I understand now that it's important for people in business to understand and agree with the results too. In future, we will focus on making it easier for everyone involved by using techniques like SHAP values or simpler images of model decisions.

This project was really helpful because we got to use actual retail data. But, it wasn't perfect and had some issues like missing information or strange numbers. Even though we did try to fix these problems, I think next time we should spend more time finding better ways to prepare the data before analyzing it. For example, using advanced techniques for detecting weird values and figuring out what to do when there are gaps in the data instead of just guessing at what's missing would be good ideas for us going forward.

We had trouble studying some research areas because it was hard to train and check many advanced models. We learned how important using quick code is, as well as getting high-quality computer resources. For upcoming projects, we should focus on making better code and think about using cloud computing services so that we can look into more model options without problems.

The hardest part of the project was trying to study everything in detail while also exploring many different techniques. Although we tested several algorithms, we couldn't go very deep into each one because of time restraints. If I work on a similar project again, I would either focus on studying only a few methods more deeply or dedicate more time and resources to examine numerous approaches thoroughly.

When I think about what happened in our project, I feel good that we were able to learn new things and make recommendations on how machine learning can help stores predict sales. But there's still more work that needs to be done because technology is always changing. We need to keep up with these changes so we can continue making progress in forecasting sales for stores.

We could have helped more with putting the models into use in real-world stores. It's hard to figure out how to update and keep them working well over time though. In future projects, we should maybe learn about this too – like figuring out how to fit machine learning predictions into what a store already does every day.

Working together with people from different backgrounds was fun and hard at the same time for this research project. We learned a lot from each other's points of view, but it wasn't always easy to understand each other or stay on track. Next time we do something like this, I think we should focus more on talking clearly and making sure everyone knows what they're supposed to be doing. Maybe setting up some schedules or rules could help us work better as a team and get things done faster.

As I finish thinking about this project, I realize that it taught me a lot. It helped me learn how to use machine learning and showed me the difficulties of using advanced analytics in retail. This experience was really helpful because now I know more about data science and machine learning. Plus, the work we did will make sales predictions better.

I'm looking forward to doing more research in this area because technology is improving fast. This means we can create even better ways of predicting things accurately using machine learning and artificial intelligence. For example, we could develop new methods for managing inventory by using reinforcement learning or use a technique called federated learning to make predictions based on data from many stores without sharing personal information about customers.

To sum up, the project has provided a significant chance to learn by presenting the enormous applicability of machine learning in predicting retail sales and confronting intricate hurdles that arise during its application. Despite some shortcomings in our approach, gaining insights from this venture and laying groundwork for future research is invaluable. As e-commerce progresses further ahead, precise prediction as well as sturdy sales forecasting need will escalate considerably. I am thrilled to continually contribute towards this domain leveraging

my experience gained through this undertaking; thus driving more advances in decision-making inclusive of retail analytics arena.

References

- [1]S. Cheriyan, S. Ibrahim, S. Mohanan, and S. Treesa, “Intelligent Sales Prediction Using Machine Learning Techniques,” *IEEE Xplore*, Aug. 01, 2018.
<https://ieeexplore.ieee.org/abstract/document/8659115>
- [2]X. dairu and Z. Shilong, “Machine Learning Model for Sales Forecasting by Using XGBoost,” *IEEE Xplore*, 2021. <https://ieeexplore.ieee.org/abstract/document/9342304>
- [3]D. K. Deepa and G. Raghuram, “Sales Forecasting Using Machine Learning Models,” *Annals of the Romanian Society for Cell Biology*, pp. 3928–3936, May 2021, Available:
<http://annalsofrscb.ro/index.php/journal/article/view/5059>
- [4]E. Martins and Napoleão Verardi Galeale, “Sales forecasting using machine learning algorithms,” *GeSec*, vol. 14, no. 7, pp. 11294–11308, Jul. 2023, doi:
<https://doi.org/10.7769/gesec.v14i7.1670>.
- [5]S. Malik, M. Khan, M. K. Abid, and N. Aslam, “Sales Forecasting Using Machine Learning Algorithm in the Retail Sector,” *Journal of Computing & Biomedical Informatics*, vol. 6, no. 02, pp. 282–294, Mar. 2024, Available:
<https://www.jcbi.org/index.php/Main/article/view/370>
- [6]B. Pavlyshenko, “Machine-Learning Models for Sales Time Series Forecasting,” *Data*, vol. 4, no. 1, p. 15, Jan. 2019, doi: <https://doi.org/10.3390/data4010015>.
- [7]I-Fei. Chen and C.-J. Lu, “Sales forecasting by combining clustering and machine-learning techniques for computer retailing,” *Neural Computing and Applications*, vol. 28, no. 9, pp. 2633–2647, Feb. 2016, doi: <https://doi.org/10.1007/s00521-016-2215-x>.
- [8]D. Swami, A. D. Shah, and S. K. B. Ray, “Predicting Future Sales of Retail Products using Machine Learning,” *arXiv:2008.07779 [cs, stat]*, Aug. 2020, Available:
<https://arxiv.org/abs/2008.07779>
- [9]S. Kumar and V. Sharma, “Comparative Analysis of Machine Learning Techniques in Sale Forecasting,” *International Journal of Computer Applications*, vol. 53, no. 6, pp. 975–8887, 2012, Available:
<https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=6105203bece74984cd2915a8d1b42741706bf268>
- [10]M. Gurnani, Y. Korke, P. Shah, S. Udmale, V. Sambhe, and S. Bhirud, “Forecasting of sales by using fusion of machine learning techniques,” *2017 International Conference on Data Management, Analytics and Innovation (ICDMAI)*, Feb. 2017, doi:
<https://doi.org/10.1109/icdmai.2017.8073492>.

- [11]Y. Ensafi, S. H. Amin, G. Zhang, and B. Shah, “Time-series forecasting of seasonal items sales using machine learning – A comparative analysis,” *International Journal of Information Management Data Insights*, vol. 2, no. 1, p. 100058, Apr. 2022, doi: <https://doi.org/10.1016/j.jjime.2022.100058>.
- [12]J. Oladunjoye, Abiodun, and J. Omoware, “A COMPARATIVE ANALYSIS OF SALES PREDICTION USING MACHINE LEARNING TECHNIQUES,” *FUW Trends in Science & Technology Journal*, www.ftstjournal.com e-ISSN, vol. 8, no. 3, pp. 415–422, 2408, Available: <http://www.ftstjournal.com/uploads/docs/83%20Article%2064.pdf>
- [13]A. Mehrotra, T. Srivastwa, H. Varshney, and K. M. Baalamurugan, “Machine Learning Models Comparison for Sales Forecasting,” *IEEE Xplore*, Dec. 01, 2021, <https://ieeexplore.ieee.org/document/9725419>
- [14]B. Zhang, M.-L. Tseng, L. Qi, Y. Guo, and C.-H. Wang, “A comparative online sales forecasting analysis: Data mining techniques,” *Computers & Industrial Engineering*, vol. 176, p. 108935, Feb. 2023, doi: <https://doi.org/10.1016/j.cie.2022.108935>.
- [15]F. Haselbeck, J. Killinger, K. Menrad, T. Hannus, and D. G. Grimm, “Machine learning outperforms classical forecasting on horticultural sales predictions,” *Machine Learning with Applications*, vol. 7, p. 100239, Dec. 2021, doi: <https://doi.org/10.1016/j.mlwa.2021.100239>.
- [16]M. Lawrence, M. O’Connor, and B. Edmundson, “A field study of sales forecasting accuracy and processes,” *European Journal of Operational Research*, vol. 122, no. 1, pp. 151–160, Apr. 2000, doi: [https://doi.org/10.1016/s0377-2217\(99\)00085-5](https://doi.org/10.1016/s0377-2217(99)00085-5).
- [17]D. F. Davis and J. T. Mentzer, “Organizational factors in sales forecasting management,” *International Journal of Forecasting*, vol. 23, no. 3, pp. 475–495, Jul. 2007, doi: <https://doi.org/10.1016/j.ijforecast.2007.02.005>.
- [18]R. Carbonneau, K. Laframboise, and R. Vahidov, “Application of machine learning techniques for supply chain demand forecasting,” *European Journal of Operational Research*, vol. 184, no. 3, pp. 1140–1154, Feb. 2008, doi: <https://doi.org/10.1016/j.ejor.2006.12.004>.
- [19]G. Tsoumakas, “A survey of machine learning techniques for food sales prediction,” *Artificial Intelligence Review*, vol. 52, no. 1, pp. 441–447, Jun. 2018, doi: <https://doi.org/10.1007/s10462-018-9637-z>.
- [20]A. Schmidt, M. W. U. Kabir, and M. T. Hoque, “Machine Learning Based Restaurant Sales Forecasting,” *Machine Learning and Knowledge Extraction*, vol. 4, no. 1, pp. 105–130, Jan. 2022, doi: <https://doi.org/10.3390/make4010006>.
- [21]A. Krishna, A. V, A. Aich, and C. Hegde, “Sales-forecasting of Retail Stores using

Machine Learning Techniques,” *2018 3rd International Conference on Computational Systems and Information Technology for Sustainable Solutions (CSITSS)*, Dec. 2018, doi: <https://doi.org/10.1109/csitss.2018.8768765>.

[22]M. Zohdi, M. Rafiee, V. Kayvanfar, and A. Salamiraad, “Demand forecasting based machine learning algorithms on customer information: an applied approach,” *International Journal of Information Technology*, vol. 14, Feb. 2022, doi: <https://doi.org/10.1007/s41870-022-00875-3>.

[23]J. Ding, Z. Chen, L. Xiaolong, and B. Lai, “Sales Forecasting Based on CatBoost,” *2020 2nd International Conference on Information Technology and Computer Application (ITCA)*, Dec. 2020, doi: <https://doi.org/10.1109/itca52113.2020.00138>.

[24]M. Soltaninejad, R. Aghazadeh, S. Shaghaghi, and M. Zarei, “Using Machine Learning Techniques to Forecast Mehram Company’s Sales: A Case Study,” *Journal of Business and Management Studies*, vol. 6, no. 2, pp. 42–53, Mar. 2024, doi: <https://doi.org/10.32996/jbms.2024.6.2.4>.

[25]O. Wisesa, A. Andriansyah, and O. Ibrahim Khalaf, “Prediction Analysis for Business To Business (B2B) Sales of Telecommunication Services using Machine Learning Techniques,” *Majlesi Journal of Electrical Engineering*, vol. 14, no. 4, pp. 145–153, Dec. 2020, doi: <https://doi.org/10.29252/mjee.14.4.145>.

[26]J. T. Mentzer and M. A. Moon, *Sales Forecasting Management: A Demand Management Approach*. SAGE Publications, 2004. Accessed: Jul. 16, 2024. [Online]. Available: <https://books.google.com/books?hl=en&lr=&id=ZchyAwAAQBAJ&oi=fnd&pg=PP1&dq=Sales+Forecasting&ots=1YhsMelXta&sig=zoeVMig7dwFNyYFeTbobCtmidRs>

[27]F. M. Thiesing and O. Vornberger, “Sales forecasting using neural networks,” *IEEE Xplore*, Jun. 01, 1997. <https://ieeexplore.ieee.org/abstract/document/614234>

[28]R. Gustriansyah, E. Ermatita, and D. P. Rini, “An approach for sales forecasting,” *Expert Systems with Applications*, vol. 207, p. 118043, Nov. 2022, doi: <https://doi.org/10.1016/j.eswa.2022.118043>.

[29]P. J. Harrison, “Short-Term Sales Forecasting,” *Applied Statistics*, vol. 14, no. 2/3, p. 102, 1965, doi: <https://doi.org/10.2307/2985338>.

[30]K. Saraswathi, N. T. Renukadevi, S. Nandhinidevi, S. Gayathridevi, and P. Naveen, “Sales prediction using machine learning approaches,” *PROCEEDINGS OF THE 4TH NATIONAL CONFERENCE ON CURRENT AND EMERGING PROCESS TECHNOLOGIES E-CONCEPT-2021*, 2021, doi: <https://doi.org/10.1063/5.0068655>.

[31]G. Di Pillo, V. Latorre, S. Lucidi, and E. Procacci, “An application of support vector

- machines to sales forecasting under promotions,” *4OR*, vol. 14, no. 3, pp. 309–325, Apr. 2016, doi: <https://doi.org/10.1007/s10288-016-0316-0>.
- [32] V. Sai Vineeth, H. Kusetogullari, and A. Boone, “Forecasting Sales of Truck Components: A Machine Learning Approach,” *2020 IEEE 10th International Conference on Intelligent Systems (IS)*, Aug. 2020, doi: <https://doi.org/10.1109/is48319.2020.9200128>.
- [33] S. Kohli, G. T. Godwin, and S. Urolagin, “Sales Prediction Using Linear and KNN Regression,” *Algorithms for Intelligent Systems*, pp. 321–329, Jul. 2020, doi: https://doi.org/10.1007/978-981-15-5243-4_29.
- [34] M. Spuritha, C. S. Kashyap, T. R. Nambiar, D. R. Kiran, N. S. Rao, and G. P. Reddy, “Quotidian Sales Forecasting using Machine Learning,” *IEEE Xplore*, Sep. 01, 2021. <https://ieeexplore.ieee.org/document/9633975> (accessed Mar. 12, 2023).
- [35] G. Khekare, A. V. Turukmane, C. Dhule, P. Sharma, and L. Kumar Bramhane, “Experimental Performance Analysis of Machine Learning Algorithms,” *Proceeding of 2021 International Conference on Wireless Communications, Networking and Applications*, pp. 1041–1052, 2022, doi: https://doi.org/10.1007/978-981-19-2456-9_104.
- [36] J. G. Wacker and R. R. Lummus, “Sales forecasting for strategic resource planning,” *International Journal of Operations & Production Management*, vol. 22, no. 9, pp. 1014–1031, Sep. 2002, doi: <https://doi.org/10.1108/01443570210440519>.