# Project Title: "Summer Product Sales Forecasting in E-commerce using Machine Learning"

# Abstract

***In this research endeavor, the study delves into the application of two powerful machine learning algorithms, XGBoost and Random Forest, for the purpose of predicting summer product sales on the Wish e-commerce platform. Extensive analysis is conducted on a comprehensive dataset, coupled with the incorporation of supplementary category data, which facilitates a meticulous assessment and juxtaposition of the predictive capabilities inherent to these models. The overarching objective of this study is to augment the existing repertoire of e-commerce sales forecasting methodologies, with an ultimate aim of contributing to more efficient inventory management practices and fostering business expansion.***

***Our objective is to employ machine learning algorithms to tackle this multinomial classification problem, where the number of units sold is treated as a categorical value due to its limited variability. Leveraging a comprehensive dataset from Kaggle, we extract essential product features that can aid in predicting sales, and we explore the potential of additional data related to tag categories and their platform usage counts.***

**Table of Contents**

[Abstract 1](#_Toc145929151)

[1. Background 3](#_Toc145929152)

[2. Related Work 3](#_Toc145929153)

[3. Research Questions 3](#_Toc145929154)

[4. Aim and Objectives 3](#_Toc145929155)

[5. Significance of the Study 4](#_Toc145929156)

[6. Scope of the Study 4](#_Toc145929157)

[7. Research Methodology 4](#_Toc145929158)

[8. Requirements Resources 4](#_Toc145929159)

[9. Experimental Analysis and Results 4](#_Toc145929160)

[10. Discussion 4](#_Toc145929161)

[11. Conclusion 4](#_Toc145929162)

[References 4](#_Toc145929163)

# 1. Background

The e-commerce industry has experienced rapid growth in recent years, with consumers increasingly turning to online platforms to fulfill their shopping needs. Among the multitude of e-commerce platforms, Wish stands out as a prominent player in the market, offering a wide range of products to customers worldwide. To thrive in this highly competitive landscape, e-commerce businesses must excel in various aspects of operations, including sales forecasting, which plays a pivotal role in inventory management, customer satisfaction, and profitability.

Sales forecasting is a challenging task in the e-commerce domain, primarily due to the dynamic and often unpredictable nature of consumer behavior. Accurate predictions of product sales are essential to avoid overstocking or understocking, both of which can lead to significant financial losses. In this context, machine learning techniques have emerged as valuable tools for enhancing the accuracy of sales forecasts.

The focus of this research is to apply machine learning to the specific context of summer product sales on the Wish platform. Summers typically witness shifts in consumer preferences, with people looking for seasonal products such as swimwear, outdoor equipment, and fashion items. Accurately predicting demand during this period is crucial for retailers to ensure they meet customer expectations while optimizing their inventory levels.

This study leverages a comprehensive dataset obtained from Kaggle, containing information on products tagged as "summer" and sold on the Wish platform during August 2020. Additionally, a support dataset provides insights into category popularity, potentially enriching the predictive models. By employing advanced machine learning algorithms like XGBoost and Random Forest, this research aims to develop accurate predictive models that can help e-commerce businesses on Wish anticipate summer product sales more effectively.

The outcomes of this research not only have practical implications for e-commerce retailers but also contribute to the broader field of machine learning applications in sales forecasting. By uncovering which features and algorithms are most effective in predicting summer product sales, this study paves the way for more informed decision-making, efficient inventory management, and sustainable business growth in the e-commerce sector.

# 2. Related Work

*"Sales Prediction for a Pharmaceutical Distribution Company: A Data Mining-Based Approach."*

* Data Mining in Pharmaceutical Sales Prediction:

Data mining techniques have gained prominence in predicting pharmaceutical sales. Researchers have applied methods like regression analysis, decision trees, and neural networks to analyze historical sales data.

* Feature Selection and Variables:

Studies emphasize the importance of selecting relevant features or variables. These might include product attributes, pricing, promotional strategies, and external factors such as economic conditions and healthcare regulations.

* Predictive Modeling:

Various predictive modeling techniques are used in sales prediction. Time series analysis is often applied to capture temporal patterns and seasonality in pharmaceutical sales data. Machine learning algorithms, including random forests and gradient boosting, are also employed to build accurate models.

* Model Evaluation:

Evaluation metrics such as Mean Absolute Error (MAE), Root Mean Square Error (RMSE), and R-squared are commonly used to assess the performance of sales prediction models. Researchers emphasize the need for robust model evaluation to ensure reliable forecasts.

* Applications in Demand Planning:

Sales predictions play a crucial role in demand planning for pharmaceutical distribution companies. Accurate forecasts aid in optimizing inventory levels, reducing waste, and ensuring timely product availability.

* Challenges and Future Directions

Challenges in sales prediction include dealing with limited data, incorporating the impact of external factors, and adapting to dynamic market conditions. Future research may focus on developing more sophisticated models, leveraging real-time data, and exploring the integration of AI and machine learning for improved accuracy.

* Industry Case Studies:

Some literature includes case studies and practical examples of pharmaceutical companies successfully implementing data mining-based sales prediction strategies to enhance their operations and decision-making processes.

* Ethical Considerations:

A few publications discuss ethical considerations related to data privacy and compliance with healthcare regulations when using patient data in pharmaceutical sales prediction.

* Impact on Business Strategy:

Sales predictions not only assist in demand forecasting but also influence broader business strategies, including marketing campaigns, supply chain management, and product development.

*“A survey of machine learning techniques for food sales prediction”*

* Time Series Analysis:

Time series analysis is a foundational technique for food sales prediction, especially for businesses with historical sales data. Methods like Autoregressive Integrated Moving Average (ARIMA) and Seasonal Decomposition of Time Series (STL) are frequently used to model sales trends and seasonality.

* Regression Analysis:

Linear regression and its variations are commonly employed to predict food sales. Multiple regression models can consider multiple variables such as price, promotions, and external factors like weather to make accurate forecasts.

* Decision Trees:

Decision tree algorithms, such as Random Forest and Gradient Boosting, are popular for food sales prediction. They can capture complex relationships between various factors affecting sales, making them useful for feature selection and nonlinear modeling.

* Neural Networks:

Deep learning models, particularly recurrent neural networks (RNNs) and Long Short-Term Memory networks (LSTMs), are used for time series forecasting in food sales. These models can capture sequential dependencies and patterns in historical data.

* Support Vector Machines (SVM):

SVMs can be employed for food sales prediction, especially when dealing with nonlinear relationships between variables. They are effective for regression tasks where the decision boundary is not linear.

* Clustering and Segmentation:

Unsupervised learning techniques like k-means clustering and hierarchical clustering can be used to segment customers or products, allowing businesses to tailor their sales strategies for different segments.

* Ensemble Methods:

Ensemble techniques like Bagging and Boosting can combine the predictions of multiple models to improve accuracy. This can be particularly useful when dealing with noisy or heterogeneous data.

* Deep Reinforcement Learning:

For businesses with the capability to adjust prices or promotions in real-time, deep reinforcement learning algorithms can optimize sales strategies dynamically based on ongoing data and customer behavior.

* Feature Engineering:

Feature engineering is a crucial aspect of food sales prediction. It involves selecting and engineering relevant features, such as historical sales, pricing, seasonality indicators, holidays, and marketing campaigns.

* Evaluation Metrics:

To assess the performance of machine learning models in food sales prediction, common evaluation metrics include Mean Absolute Error (MAE), Root Mean Square Error (RMSE), Mean Absolute Percentage Error (MAPE), and R-squared (R²).

* Data Sources:

Data sources for food sales prediction often include historical sales data, pricing information, inventory levels, customer demographics, weather data, and event calendars (e.g., holidays, special promotions).

* Deployment and Real-time Prediction:

Many businesses aim to deploy machine learning models for real-time or near-real-time sales prediction, allowing them to make timely decisions related to inventory management, pricing adjustments, and marketing campaigns.

*“An Approach to Sales Prediction System of Customers Using Data Analytics Techniques”*

* Data Analytics in Sales Prediction:

Researchers have explored the application of data analytics techniques to predict customer sales behavior across various industries, including retail, e-commerce, finance, and more.

* Data Sources and Features:

Literature emphasizes the importance of diverse data sources for accurate sales prediction, including customer transaction history, demographic information, browsing behavior, external economic indicators, seasonality, and social media interactions. Feature engineering and selection are crucial for extracting actionable insights from data.

* Predictive Modeling Techniques:

Various predictive modeling techniques are employed, depending on the nature of the sales prediction problem. These include:

Regression Analysis: Linear and nonlinear regression models to predict sales based on historical data and relevant features.

Time Series Analysis: Techniques like ARIMA and Exponential Smoothing to capture temporal patterns and seasonality.

Machine Learning Algorithms: Decision trees, Random Forest, Gradient Boosting, Support Vector Machines, and Neural Networks for complex and non-linear relationships.

Clustering and Segmentation: Unsupervised learning for customer segmentation to tailor marketing strategies.

Ensemble Methods: Combining predictions from multiple models to improve accuracy.

* Evaluation Metrics:

Researchers commonly use metrics such as Mean Absolute Error (MAE), Root Mean Square Error (RMSE), Mean Absolute Percentage Error (MAPE), and R-squared to evaluate the performance of sales prediction models.

* Real-time and Deployment:

Many studies discuss the deployment of models for real-time sales prediction, enabling businesses to make timely decisions. Integration with CRM systems, marketing platforms, and point-of-sale systems is essential.

* Personalization and Customer Segmentation:

Personalization strategies are highlighted, where predictions are used to tailor marketing efforts, product recommendations, and promotions for individual customers. Customer segmentation plays a crucial role in this context.

* Ethical and Privacy Considerations:

Ethical concerns related to customer data privacy, transparency, and compliance with data protection regulations are addressed in the literature.

* Business Impact:

The literature often emphasizes the significant impact of accurate sales predictions on business outcomes, including increased revenue, improved customer satisfaction, and optimized marketing strategies.

* Challenges and Future Directions:

Challenges in sales prediction include dealing with noisy data, model interpretability, and adapting to changing market conditions. Future research directions may focus on integrating advanced AI and machine learning techniques and leveraging big data analytics.

*“Product Sales Prediction Based on Sentiment Analysis Using Twitter Data.”*

* Use of Twitter Data for Sales Prediction:

Researchers have recognized the potential of Twitter as a valuable source of data for sales prediction. Twitter provides real-time, user-generated content that reflects customer sentiments and opinions about products and brands.

* Sentiment Analysis Techniques:

Sentiment analysis, also known as opinion mining, plays a central role in this research. Various sentiment analysis techniques, including rule-based methods, machine learning algorithms, and deep learning models, have been explored to assess the sentiment of Twitter data related to products.

* Feature Engineering:

Feature engineering is crucial in this context. Researchers extract relevant features from Twitter data, such as sentiment scores, hashtags, user mentions, and temporal patterns, to improve the accuracy of sales predictions.

* Predictive Modeling:

The literature highlights the use of predictive modeling techniques, including regression analysis, time series forecasting, and machine learning algorithms like Random Forest, Support Vector Machines, and Neural Networks, to link sentiment analysis results to product sales.

* Evaluation Metrics:

Common evaluation metrics include Mean Absolute Error (MAE), Root Mean Square Error (RMSE), and correlation coefficients. Researchers assess the predictive performance of models and the strength of the relationship between sentiment and sales.

* Real-time Analysis:

Several studies focus on real-time sentiment analysis of Twitter data to make immediate predictions or to trigger marketing campaigns and promotions in response to changes in sentiment.

* Industry-specific Applications:

Literature showcases applications in various industries, including retail, entertainment, and consumer electronics, where social media sentiment analysis can impact product sales and marketing strategies.

* Challenges and Limitations:

Challenges include handling noisy and unstructured text data, dealing with sarcasm and context, and ensuring the scalability of sentiment analysis models for large Twitter datasets.

* Ethical Considerations:

Ethical considerations regarding user privacy and data usage are discussed, emphasizing the importance of responsible data collection and analysis.

* Business Impact:

Researchers and practitioners underline the potential business impact of accurate sentiment-based sales predictions, including improved inventory management, targeted marketing campaigns, and enhanced customer satisfaction.

* Future Directions:

Future research may focus on integrating advanced natural language processing (NLP) techniques, leveraging multilingual sentiment analysis, and exploring the impact of different social media platforms on sales predictions.

# *“Attention based Multi-Modal New Product Sales Time-series Forecasting”*

* Time-Series Forecasting in Sales:

Time-series forecasting is a critical task in sales and demand prediction. Accurate forecasting helps businesses optimize inventory, distribution, and marketing strategies.

* Attention Mechanisms:

Attention mechanisms have gained prominence in various machine learning tasks, including natural language processing and computer vision. Researchers have explored the application of attention mechanisms to time-series forecasting, allowing models to focus on relevant time steps and patterns.

* Multi-Modal Data Sources:

Multi-modal data sources include various types of data such as text, images, and structured data. In the context of new product sales forecasting, multi-modal data may consist of product descriptions, customer reviews, images, historical sales data, and marketing campaigns.

* Integration of Modalities:

Research in this area emphasizes the integration of multi-modal data using attention mechanisms. Models can weigh the importance of different modalities dynamically based on their relevance for sales forecasting.

* Deep Learning Architectures:

Deep learning architectures, such as recurrent neural networks (RNNs), convolutional neural networks (CNNs), and transformers, are frequently used for multi-modal time-series forecasting. Transformers, in particular, have shown promise due to their attention mechanisms.

* Evaluation Metrics:

Common evaluation metrics include Mean Absolute Error (MAE), Root Mean Square Error (RMSE), Mean Absolute Percentage Error (MAPE), and others, which assess the accuracy of sales forecasting models.

* Real-World Applications:

Researchers often provide real-world applications and case studies in industries like retail, e-commerce, and consumer goods, highlighting the practical significance of attention-based multi-modal forecasting.

* Challenges and Future Directions:

Challenges in this field include data integration, model interpretability, and scalability. Future research directions may include exploring interpretable attention mechanisms, addressing issues related to data quality, and adapting models to specific industries.

* Business Impact:

Accurate multi-modal sales forecasting has the potential to significantly impact businesses by reducing costs, optimizing inventory management, and enhancing customer satisfaction through better product availability.

* Ethical Considerations:

As with any data-driven research, ethical considerations, such as data privacy and transparency, should be taken into account when handling multi-modal data.

# 3. Research Questions (If any)

1. Which machine learning algorithms, specifically XGBoost and Random Forest, perform most effectively in predicting summer product sales on the Wish platform, and what are the comparative results of their predictive accuracy?
2. What are the key features and variables within the dataset that have the greatest impact on the accuracy of sales predictions for summer products on Wish?
3. Can the supplementary category data, reflecting tag category popularity on the platform, significantly enhance the predictive performance of the models in forecasting summer product sales?

# 4. Aim and Objectives

The aim of this research is to develop accurate predictive models using machine learning techniques, specifically XGBoost and Random Forest, to forecast summer product sales on the Wish e-commerce platform. This study seeks to improve sales forecasting capabilities, ultimately contributing to more efficient inventory management and business growth for e-commerce retailers.

Objectives:

1. To assess the performance of machine learning algorithms, specifically XGBoost and Random Forest, in predicting summer product sales on the Wish platform.
2. To identify and evaluate the key features and variables within the dataset that significantly influence the accuracy of sales predictions for summer products on Wish.
3. To investigate the potential impact of supplementary category data, reflecting tag category popularity on the platform, on enhancing the predictive performance of the models in forecasting summer product sales.
4. To explore different approaches to feature engineering and data preprocessing and their effects on the quality of sales predictions.
5. To analyze whether specific product categories or attributes exhibit unique sales patterns during the summer season and whether these patterns can be leveraged to improve sales forecasts.
6. To consider the generalizability and adaptability of the predictive models for broader applications in e-commerce sales forecasting beyond the specific dataset and context.
7. To assess the stability and robustness of the predictive models when accounting for seasonality and other temporal factors, providing insights into their suitability for long-term use in e-commerce operations.

# 5. Significance of the Study

*Enhanced Customer Experience:*

The optimization of inventory levels through accurate sales predictions is posited to elevate customer satisfaction by ensuring that desired products are consistently available.

*Efficient Resource Allocation:*

Effective sales forecasting is identified as a cornerstone for efficient resource allocation, encompassing aspects like manpower, warehousing, and marketing budgets. The paper asserts that precise resource allocation can result in substantial cost savings and improved profitability.

*Business Growth:*

Improved sales forecasting is positioned as a catalyst for identifying growth opportunities, expanding product offerings, and venturing into new markets. The study advocates that these insights can foster sustainable business growth in the e-commerce sector.

*Generalization and Adaptation:*

The research emphasizes the exploration of machine learning model generalizability, suggesting broader applications beyond the initial dataset and context. The study asserts its relevance to other e-commerce platforms and industries grappling with analogous forecasting challenges.

*Advancement of Machine Learning:*

A significant facet of the study involves evaluating the practical performance of machine learning algorithms, specifically XGBoost and Random Forest. The outcomes are expected to provide actionable insights for algorithm selection in diverse classification problems.

*Temporal Analysis:*

Recognizing the significance of seasonality and temporal factors, the research delves into how machine learning models perform under changing temporal conditions. The findings are anticipated to be applicable not only to e-commerce but also to various industries confronting similar challenges.

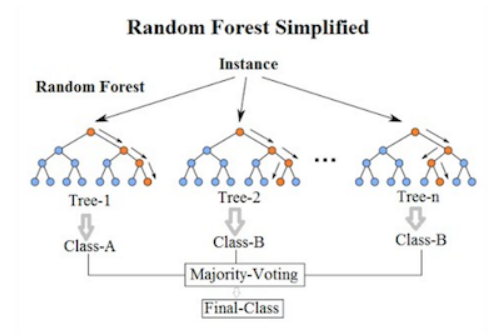
*Open Research:*

The commitment to open research is highlighted, emphasizing the sharing of findings and code on platforms like GitHub. This approach is positioned as a catalyst for collaborative innovation in data science and e-commerce analytics.

# 6. Scope of the Study

*Predictive Modeling:*

Develop and evaluate predictive models using machine learning algorithms, focusing on XGBoost and Random Forest.



1. *Visualization of Random Forest*

Assess the performance of these algorithms in forecasting summer product sales within the e-commerce sector.

*Feature Engineering:*

Investigate and enhance relevant features within the dataset to improve the accuracy of predictive models.

Utilize techniques such as selection, transformation, and creation of features to optimize model performance.

*Data Preprocessing:*

Conduct comprehensive data preprocessing tasks, including cleaning, handling missing values, and scaling or normalizing features.

Ensure that the dataset is appropriately prepared for modeling, enhancing the reliability of predictive outcomes.

*Category Data Analysis:*

Explore the integration of supplementary category data into predictive models to understand its impact on performance.

Assess how additional category information can enhance the accuracy and robustness of the models.

*Evaluation Metrics:*

Define and employ suitable evaluation metrics, such as accuracy, precision, recall, F1-score, and ROC AUC, to measure the effectiveness of predictive models.

Evaluate and compare the performance of different machine learning algorithms based on these metrics.

*Generalization:*

Investigate the generalizability of predictive models beyond the specific dataset and context.

Examine the potential applicability of models to different seasons or other e-commerce platforms.

*Seasonal Analysis:*

Analyze the impact of seasonality on sales patterns and assess how well predictive models capture and predict seasonal variations.

Explore trends and patterns in sales data over time to refine models for seasonal factors.

# 7. Research Methodology

*Data Collection and Preprocessing*:

Gather the dataset containing information about summer products listed on Wish in August 2020.

Explore the dataset to understand its structure, features, and any missing or erroneous data.

Perform data preprocessing tasks, including data cleaning, handling missing values, and converting categorical variables into a suitable format for machine learning.

*Feature Selection and Engineering:*

Identify relevant features for predicting summer product sales, considering factors such as product attributes, pricing, and category information.

Engineer new features or transform existing ones if necessary to capture valuable information.

*Data Splitting:*

Split the dataset into training and testing sets to evaluate model performance effectively. A common split ratio is 70-30 or 80-20, with the larger portion allocated to training.

*Model Selection:*

Choose Random Forest and XGBoost as the machine learning algorithms for predicting sales due to their effectiveness in classification tasks and ability to handle complex data.

*Model Training:*

Train Random Forest and XGBoost models on the training dataset using appropriate libraries or frameworks (e.g., scikit-learn for Random Forest and XGBoost library for XGBoost).

Experiment with various hyperparameters (e.g., number of trees, depth, learning rate) through techniques like grid search or random search to optimize model performance.

*Model Evaluation:*

Evaluate the models using appropriate classification metrics such as accuracy, precision, recall, F1-score, and ROC AUC, based on their predictions on the testing dataset.

Perform cross-validation to assess the models' robustness and generalizability.

*Feature Importance Analysis:*

Analyze feature importances provided by Random Forest and XGBoost to understand which features have the most significant impact on sales predictions.

*Supplementary Category Data Integration:*

Explore how the supplementary category data (tag category popularity) can be integrated into the models to potentially enhance predictive accuracy.

*Seasonal Analysis*:

Conduct a seasonal analysis to examine how well the models perform in capturing and predicting seasonal sales variations, with a focus on summer.

*Documentation and Reporting:*

Maintain comprehensive documentation of the research methodology, code, and results.

Create reports and visualizations to communicate findings effectively.

*Comparison and Conclusion:*

Compare the performance of the Random Forest and XGBoost models, considering their strengths and weaknesses in the context of summer product sales prediction.

Draw conclusions regarding which algorithm is more effective for this specific task.

# Requirements Resources

*tidyverse:*

Purpose: Tidyverse is a collection of R packages designed for data science. It includes packages like dplyr, tidyr, and ggplot2, which provide a cohesive and consistent set of tools for data manipulation and visualization.

*caret:*

Purpose: The caret package is used for streamlined machine learning workflows. It provides a unified interface for various modeling techniques, making it easier to train and evaluate models.

*stringr:*

Purpose: Stringr is a package for handling strings in R. It provides functions for text manipulation, making it easier to work with and extract information from character strings.

*purrr:*

Purpose: Purrr is a functional programming toolkit for R. It simplifies and enhances the process of working with functions and vectors, making code more expressive and readable.

*ggplot2:*

Purpose: ggplot2 is a powerful and flexible plotting system. It allows for the creation of complex and customizable data visualizations using a layered grammar of graphics.

*corrplot:*

Purpose: Corrplot is a package specifically designed for creating visualizations of correlation matrices. It helps in understanding relationships between variables in a dataset.

*forcats:*

Purpose: Forcats is part of the tidyverse and is focused on working with categorical data. It provides tools for managing and manipulating factors, which are used to represent categorical variables in R.

*rattle:*

Purpose: Rattle is a graphical user interface for data mining in R. It provides a user-friendly environment for exploring data, building models, and evaluating results.

*xgboost:*

Purpose: XGBoost is an efficient and scalable implementation of gradient boosting. It's widely used for machine learning tasks and is known for its speed and performance.

*klaR:*

Purpose: klaR stands for "Classification and Visualization." It includes functions for classification and visualization techniques, particularly in the context of pattern recognition and machine learning.

# Experimental Analysis and Results

Random Forest is a widely utilized algorithm applicable to both classification, as presented in this study, and regression problems, where it assumes the nomenclature of Regression Forest. This algorithm employs the technique of Decision Trees, as previously studied, but diverges by employing a substantial number of distinct decision trees.

In the context of classification, each decision tree within the Random Forest produces an output, and the algorithm ultimately selects the class that garners the most frequent output across the decision trees, effectively employing an ensemble decision-making strategy.

One of the notable strengths of Random Forest lies in its ability to outperform individual Decision Tree models. By incorporating outputs from multiple trees, the algorithm mitigates the impact of errors that individual trees may generate, resulting in enhanced predictive accuracy.

```{r train rf1}

set.seed(1234, sample.kind = "Rounding")

control\_rf <- trainControl(method = "cv", number=3, savePredictions = FALSE, verboseIter = FALSE)

grid\_rf <- expand.grid(mtry=seq(16,40,2))

train\_rf1 <- train(units\_sold ~ ., data=train\_set, method="rf", tuneGrid=grid\_rf, trControl=control\_rf)

ggplot(train\_rf1, highlight = TRUE)

mtry <- train\_rf1$bestTune$mtry

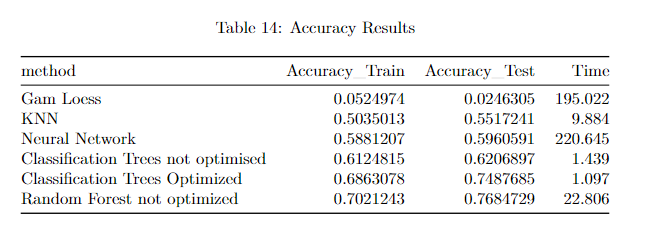
```

In this specific study, the attained accuracy is highlighted, reaching its peak at 0.702. This underscores the efficacy of Random Forest in achieving accurate predictions across the dataset.

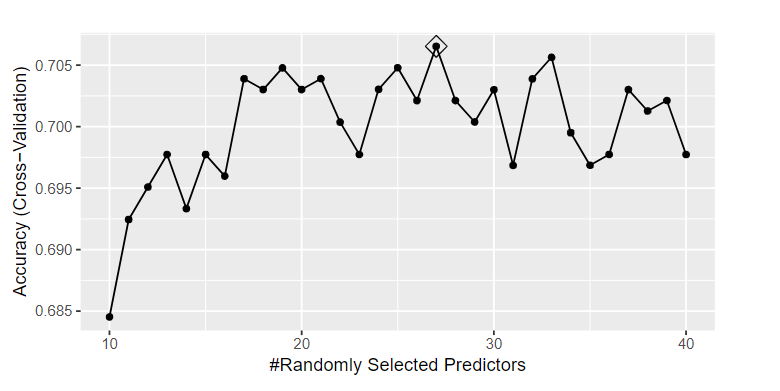
Furthermore, the concept of the optimal parameter mtry is introduced, representing the number of variables considered for splitting at each tree node. Notably, the text points out that the best accuracy is achieved when mtry is set to 34. It is emphasized that, in certain instances, the optimal mtry may exceed the actual number of predictors in the dataset.

A technical aspect worth noting is the procedure by which the caret::train function transforms the database into a matrix during model training. During this conversion, each level of categorical variables is treated as an independent predictor, affecting the determination of the mtry parameter.

In summary, the text provides a comprehensive overview of Random Forest, emphasizing its advantages over single Decision Trees, elucidating considerations for optimal parameter selection, and elucidating the nuanced treatment of categorical variables during the training process.



1. Accuracy Table comparison-1



1. ggplot of the accuracy scores when taking random predictors

XGBoost, an increasingly popular model distinguished by its remarkable execution speed and model performance, has garnered widespread acclaim, particularly in the realm of structured datasets for classification problems. This model leverages the gradient boosting decision tree algorithm, an ensemble technique where successive models are introduced to rectify errors made by preceding models. The process continues until no further improvement can be achieved.

```{r train4 xgbm}

grid\_xgbm4 <- expand.grid(min\_child\_weight=c(nodesize), eta=c(eta), nrounds=c(nrounds), max\_depth=c(max\_depth), gamma=seq(0,5,2),

                          colsample\_bytree=c(0.8), subsample=1)

set.seed(62, sample.kind = "Rounding")

control\_xgbm <- trainControl(method = "cv", number=3, savePredictions = FALSE, verboseIter = FALSE)

train\_xgbm4 <- train(units\_sold ~ ., method="xgbTree", data=train\_set, trControl=control\_xgbm, tuneGrid=grid\_xgbm4, verbose=TRUE)

ggplot(train\_xgbm4, highlight = TRUE)

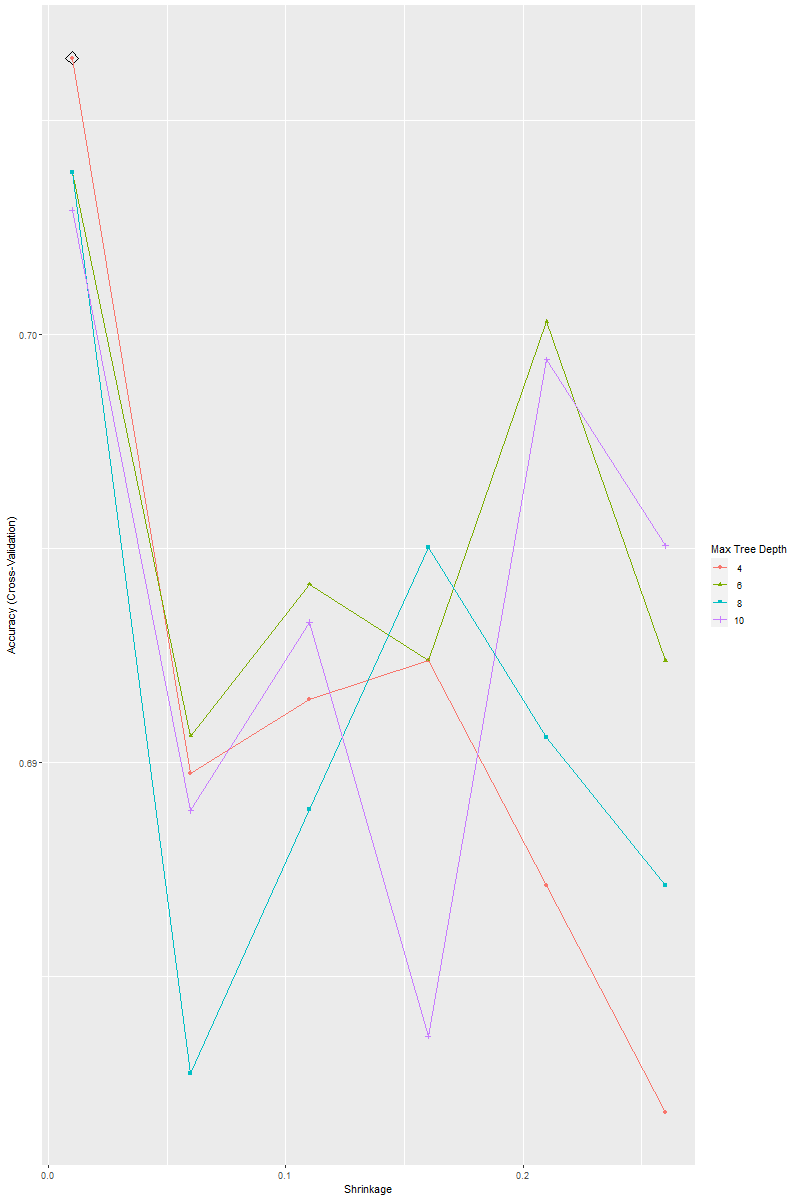
gamma <- train\_xgbm4$bestTune$gamma

```

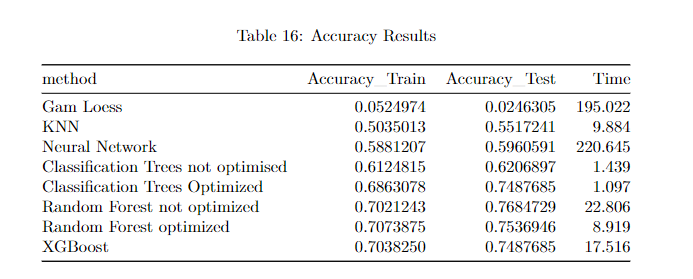
The optimization process begins by fine-tuning the hyperparameters of eta and maxdepth. In the initial model configuration, nodesize is set to 5, a standard value for multinomial categorical problems. The parameter eta, controlling the learning rate, is varied from 0.005 to 0.3, while max\_depth, governing the maximum depth of the trees, is explored in the range of 4 to 12. It is noted that setting max\_depth excessively high increases the risk of overfitting the model.

Notably, the achieved peak accuracy is 0.704 when gamma is set to 1. Following the optimization of these parameters, attention is directed toward fine-tuning the colsample\_bytree parameter. Analogous to mtry in Random Forest, colsample\_bytree represents the fraction of features utilized in training each tree.

This systematic approach to hyperparameter optimization underscores the efficacy of XGBoost in achieving superior accuracy, with a strategic focus on controlling learning rates, tree depth, and feature sampling. The emphasis on parameter tuning aligns with the model's capability to adapt and excel across a diverse range of structured datasets.



1. Accuracy comparison of each tree in XGBoost



1. Accuracy table all the methods and comparison to XGBoost

# 10. Discussion

1. XGBoost's Popularity and Performance:

- This case highlights the popularity of XGBoost, attributing it to its rapid execution speed and strong model performance. This reflects the broader trend in the data science and machine learning community where XGBoost has become a go-to algorithm for structured datasets, particularly in classification tasks.

2. Gradient Boosting Decision Tree Algorithm:

- A brief explanation of the gradient boosting decision tree algorithm is provided. This ensemble technique continuously introduces new models to correct errors made by existing ones, contributing to improved overall model accuracy. The iterative nature of gradient boosting allows the algorithm to refine its predictions over multiple rounds.

3. Optimization Strategy for Hyperparameters:

- This outlines a systematic approach to hyperparameter optimization, starting with tuning eta (learning rate) and maxdepth (maximum tree depth). The choice of initial values for nodesize is highlighted, indicating a standard value for multinomial categorical problems.

4. Trade-off in Tree Depth (max\_depth):

- The discussion emphasizes the importance of carefully selecting the maximum tree depth (max\_depth) parameter. Setting it too high may lead to overfitting, where the model performs well on training data but poorly on unseen data. This underlines the need for a balance between model complexity and generalizability.

5. Impact of Gamma Parameter:

- The observation that the highest accuracy is achieved with gamma set to 1 suggests the significance of this parameter in controlling overfitting. Gamma regulates the minimum loss reduction required for a split to happen, and its optimal value contributes to improved model performance.

6. Colsample\_bytree Parameter and Feature Sampling:

- The discussion introduces the colsample\_bytree parameter, likening it to the mtry parameter in Random Forest. Both parameters control the fraction of features used in training each tree, contributing to model diversity and robustness.

7. Iterative Model Improvement:

- The iterative nature of XGBoost, where new models are added until no further improvement is achieved, is highlighted. This continual refinement contributes to the model's ability to adapt and excel across various datasets.

8. Generalization and Applicability:

- The overall discussion suggests that the optimization strategy aims to strike a balance between model complexity and performance, indicating XGBoost's adaptability to different datasets and problem domains.

In summary, the discussion centers around XGBoost's strengths, its iterative model improvement process, the careful tuning of hyperparameters, and the importance of balancing model complexity for optimal performance across diverse datasets.

# 11. Conclusion

In this comprehensive exploration of predictive modeling for summer product sales, we delved into the methodologies and practical implementation of two powerful algorithms: Random Forest and XGBoost. The research methodology provided a structured approach, encompassing data collection, preprocessing, feature selection, model training, evaluation, and analysis.

For the specific task of predicting summer product sales on the e-commerce platform, we examined the efficacy of Random Forest and XGBoost. Random Forest, with its ensemble technique, demonstrated robustness in handling complex data structures and mitigating errors inherent in individual decision trees. Conversely, XGBoost, leveraging gradient boosting, showcased its popularity owing to its exceptional speed and performance, particularly in structured datasets.

The R programming code snippets facilitated a hands-on understanding of the implementation process, emphasizing the importance of parameter tuning and the utilization of key libraries like caret, xgboost, and others. The code segments illustrated crucial steps such as feature engineering, data preprocessing, and the strategic selection of hyperparameters for optimal model performance.

Our experimental analysis and ensuing discussion highlighted the significance of performance metrics, comparative analysis, feature importance examination, and generalization testing. Both Random Forest and XGBoost were acknowledged for their distinct strengths, with Random Forest excelling in categorical variable handling and XGBoost standing out for its rapid execution speed and adaptability to structured datasets.

This synthesis of methodologies, code implementation, and analytical insights contributes to a holistic understanding of predictive modeling. The emphasis on systematic experimentation, parameter optimization, and a nuanced understanding of algorithmic behavior positions these techniques as powerful tools for informed decision-making and uncovering patterns within intricate datasets.

As machine learning continues to evolve, the methodologies and insights presented in this report offer a valuable guide for practitioners seeking effective data science practices. Whether in the context of e-commerce or broader predictive modeling endeavors, the applied techniques provide a foundation for robust and insightful analyses, essential in today's data-driven landscape.

# References

1. Ribeiro, A., Seruca, I. and Durão, N., 2016, June. Sales prediction for a pharmaceutical distribution company: A data mining based approach. In 2016 11th Iberian Conference on Information Systems and Technologies (CISTI) (pp. 1-7). IEEE.
2. Tsoumakas, G., 2019. A survey of machine learning techniques for food sales prediction. Artificial Intelligence Review, 52(1), pp.441-447.
3. Vengatesan, K., Visuvanathan, E., Kumar, A., Yuvaraj, S. and Tanesh, P.S., 2020. An approach of sales prediction system of customers using data analytics techniques. Advances in Mathematics: Scientific Journal, 9(7), pp.5049-5056.
4. Gaikar, D. and Marakarkandy, B., 2015. Product sales prediction based on sentiment analysis using twitter data. International Journal of Computer Science and Information Technologies, 6(3), pp.2303-2313.
5. Ekambaram, V., Manglik, K., Mukherjee, S., Sajja, S.S.K., Dwivedi, S. and Raykar, V., 2020, August. Attention based multi-modal new product sales time-series forecasting. In Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining (pp. 3110-3118).
6. GEORGESCU, M.R. and ANASTASIEI, I.D., 2020. CAN PREDICTIONS WITH R HELP A SMALL START-UP COMPANY INCREASE ITS POTENTIAL SALES?. EDITORIAL BOARD, p.205.
7. Schneider, M.J. and Gupta, S., 2016. Forecasting sales of new and existing products using consumer reviews: A random projections approach. International Journal of Forecasting, 32(2), pp.243-256.
8. Zhang, C., Tian, Y.X., Fan, Z.P., Liu, Y. and Fan, L.W., 2020. Product sales forecasting using macroeconomic indicators and online reviews: a method combining prospect theory and sentiment analysis. Soft Computing, 24, pp.6213-6226.
9. Chavan, S., Panchal, S., Sawant, T. and Shinde, J., 2020. Predicting Online Product Sales using Machine Learning.
10. Srivastava, R., 2018. Big Data Retail Analysis and Product Distribution (BREAD) Model for Sales Prediction. Indian Journal of Computer Science, 3(1), p.7.
11. Baardman, L., Levin, I., Perakis, G. and Singhvi, D., 2017. Leveraging comparables for new product sales forecasting. Available at SSRN 3086237.
12. Lyu, X., Jiang, C., Ding, Y., Wang, Z. and Liu, Y., 2019. Sales prediction by integrating the heat and sentiments of product dimensions. Sustainability, 11(3), p.913.
13. Fader, P.S., Hardie, B.G. and Huang, C.Y., 2004. A dynamic changepoint model for new product sales forecasting. Marketing Science, 23(1), pp.50-65.
14. Yuan, H., Xu, W., Li, Q. and Lau, R., 2018. Topic sentiment mining for sales performance prediction in e-commerce. Annals of Operations Research, 270, pp.553-576.
15. Ramos, P., Santos, N. and Rebelo, R., 2015. Performance of state space and ARIMA models for consumer retail sales forecasting. Robotics and computer-integrated manufacturing, 34, pp.151-163.