Erasmus University Rotterdam
Erasmus School of Economics
Master Thesis Data Science and Marketing Analytics

Optimizing accuracy of consumer goods recommendations using Decision Trees and Random Forests: A case study of M5 Walmart Sales Forecasting data

Vivian de Kok (448065)

Supervisor:	Eran Raviv
Second assessor:	TBA
Date final version:	30th January 2024

The content of this thesis is the sole responsibility of the author and does not reflect the view of the supervisor, second assessor, Erasmus School of Economics or Erasmus University.

Abstract

With this research we found that the Random Forest model we used significantly outperforms the Decision Tree model. With improvements in prediction error metrics ranging from 74% to 90% and that the R^2 and Variance Explained performances are on average approximately 25.39% better than those of the Decision Tree model we used. We performed an analysis on the trade-off between accuracy and interpretability within the Fast Moving Consumer Goods industry. Using Walmart's M5 Walmart Sales Forecasting data we aim to optimize recommendation systems for the industry. While performing these models on the data, the data was divided in different subgroups and their performance was quantified on measures such as Mean Absolute Error (MAE), Root Mean Squared Error (RMSE) and Coefficient of Determination (R^2). Our analysis gives insight in interpretability of both models and aims to choose the most suitable model for consumer goods recommendations in the FMCG industry. The transparent structure of Decision Trees makes them valuable for scenarios that require a clear understanding of model decisions. Despite the lower accuracy of Decision Trees, they show a good fit to true values and explain a significant portion of the variance of the target variable. Our results show that our Random Forest models shows compared to our Decision Tree model better predictive accuracy, as mentioned approximately 25.39% in R^2 value and from 74% to 90% in error metrics. They come closer to the true values and explain a larger proportion of the variance. However, their complexity creates challenges in interpretability. Within the scope of this research we discuss their robustness in capturing data patterns and their effectiveness in the dynamic FMCG industry.

Keywords— Machine Learning, FMCG (Fast Moving Consumer Goods) Industry, Decision Trees (DT), Random Forests (RF), Predictive Analytics, Recommendation Systems, Model Interpretability, Accuracy-Interpretability Trade-off, Data Analysis, M5 Walmart Sales Forecasting, Consumer Behavior Prediction, Retail Analytics, Forecasting Models, Statistical Modeling, Business Intelligence

Contents

1 Introduction

This research focuses on the importance of predictive recommendations for the Fast Moving Consumers Goods (FMCG) industry. While performing this research we will explore the value of using specific types of machine learning models. To be more specific the models that were used during this research are Decision Trees and Random Forest. As Godoy (2022) confirms with his research that organizations extensively employ demand forecasting models to proactively formulate decisions pertaining to production, logistics, and inventories, anchoring these strategies on anticipated customer behavior dynamics. And that this makes it easier and more insightful for the companies within such an industry to determine the unit quantities to be acquired or manufactured for via various process applications. This way organisations will for example be able top coordinate resources and associated costs, establishing foundational parameters for personnel allocation, machine utilization, raw material procurement, etc. (Godoy, 2022). Based on this information, we can suggest that accuracy in forecasting sales is considered a crucial factor, providing operational and decision-making advantages. With their research Tallaro et al. (2019) support that the importance of sales estimations is big, especially for sectors as the FMCG industry (Tallaro, et al., 2019).

In order to make a legitimate prediction and derive assumptions from it that are in turn representative of the FMCG industry, the M5 Walmart Sales Forecasting dataset is deployed while performing the analysis within this research.

With our research, we will quantify the impact of advanced machine learning techniques on the precision of consumer goods recommendations within the Fast-Moving Consumer Goods (FMCG) industry. We will analyze the performance trade-offs between Decision Trees and Random Forests, using the M5 Walmart Sales Forecasting data. Our findings aid researchers and practitioners in making informed decisions when selecting and designing models, showing that carefully considering how complex a model is can lead to useful results that are still easy to understand. This approach is not only scientifically robust but also holds academic significance for its direct applicability in model selection.

By promoting methodological advances and transparency this study contributes to managerial relevance and provides useful insights for those who are in the chair of decision making in the FMCG industry. This study analyses the strengths and weaknesses of Decision Trees and Random Forests. By making the research methodology and data accessible, this study aims to be generalizing and reproducible within the industry.

By using Decision Trees and Random Forests, this research aims to provide consumer goods recommendations, particularly within the FMCG industry. Explaining what the models do makes it more obvious on why using and comparing these two specific models. As we assume that Decision trees provide interpretability by clarifying the predictive logic of the model, while Random Forests provide greater accuracy by aggregating several Decision Trees. Both aspects are valuable in order to understand consumer patterns better and can facilitate informed decisions for the FMCG industry, which is confirmed with Tallaro et al. (2019) their research, they discuss the strategic importance of reliable forecasting in management decisions across different markets and industries (Tallaro et al, 2019).

In current research on the usage of machine learning we find that Specific machine learning methods like Decision Trees are very interesting for usage performing this research, but why? Decision Trees, cited by Song and Ying (2015) are powerful statistical tools, they aim to simplify complex input-target relationships and facilitate easy interpretation without the need for distributional assumptions, among other advantages. They can deal well with skewed data and robustly handle outliers (Song & Ying, 2015). Their flexibility and adaptability in using varied subsets of features and decision rules are assumed to fit well with the dynamic FMCG industry and are therefore explored within this research. Safavian and Landgrebe (1991) provide us with the results from their paper on the usefulness of Decision Trees in other markets or industries. For example, they discuss their effectiveness in various applications, such as radar signal classification and medical diagnosis, due to their ability to simplify complex decision-making processes (Safavian & Landgrebe, 1991).

Academic research has explored that in the FMCG industry, Random Forests, known for their classification accuracy and robust results, seamlessly manage complex data. Belgiu and Drăguţ (2016) add to this with their research that the effectiveness of the Random Forest classifier is influenced by the sample design and that the feature to measure variable importance has found wide applications, including dimension reduction of hyper spectral data (Belgiu & Drăguţ, 2016). However, for this paper the usage of Random Forest can therefore be valuable in the way that not only highly accurate predictions can be provided but also variable importance is a key measure in finding consumer preference patterns.

As previously introduced in this paper and as found in the existing literature, the usage of these models and more specific the ability of these techniques to process data robustly and deal with relevant variables can enhance sales forecasting. Thus the expected impact of machine learning on future sales forecasting and demand can contribute significantly. Wu and Zheng (2015) present with their research sales forecasting models using machine learning techniques that outperform traditional models in predicting product demand, especially in volatile markets with short life cycles such as fast fashion retail (Wu & Cheng, 2015). Furthermore, Tsoumakas (2018) argues that machine-learning techniques, from simple approaches such as the moving average and ensemble methods using various learning algorithms, offer efficiency and adaptability for forecasting time series data compared to conventional statistical methods (Tsoumakas, 2018).

As a result, this paper aims to give insights into the trade-off between interpretability and accuracy for different machine learning models for the prediction of sales performed for the FMCG industry. The scope of the research are supervised machine learning models. And the aim of this paper is to quantify the accuracy and interpretability trade-off comparing a Decision Tree model and a Random Forest model. The paper focuses on the current state of literature on machine learning methods used to predict FMCG sales. The paper covers both the mechanics of classification and regression methods, as well as their respective advantages and disadvantages. The following main research question and sub questions are therefore covered:

"What is the trade-off between accuracy and interpretability for Decision Trees and Random Forests in optimizing consumer goods recommendations using the M5 Walmart Sales Forecasting data, and how does this trade-off vary based on the specific parameters and features used in each method?"

Followed by the following sub-questions.

1. Will Decision Trees be more interpretable than Random Forests but may they sacrifice accuracy in order to achieve this interpretability?
2. Will Random Forests be more accurate than Decision Trees but may they sacrifice interpretability in order to achieve this accuracy?
3. For the FMCG industry, are Decision Trees be more useful than Random Forests in terms of providing actionable insights into the factors driving consumer behavior?
4. In order to optimize the accuracy of consumer goods recommendations for the FMCG industry, is a hybrid approach that combines Decision Trees and Random Forests most effective?

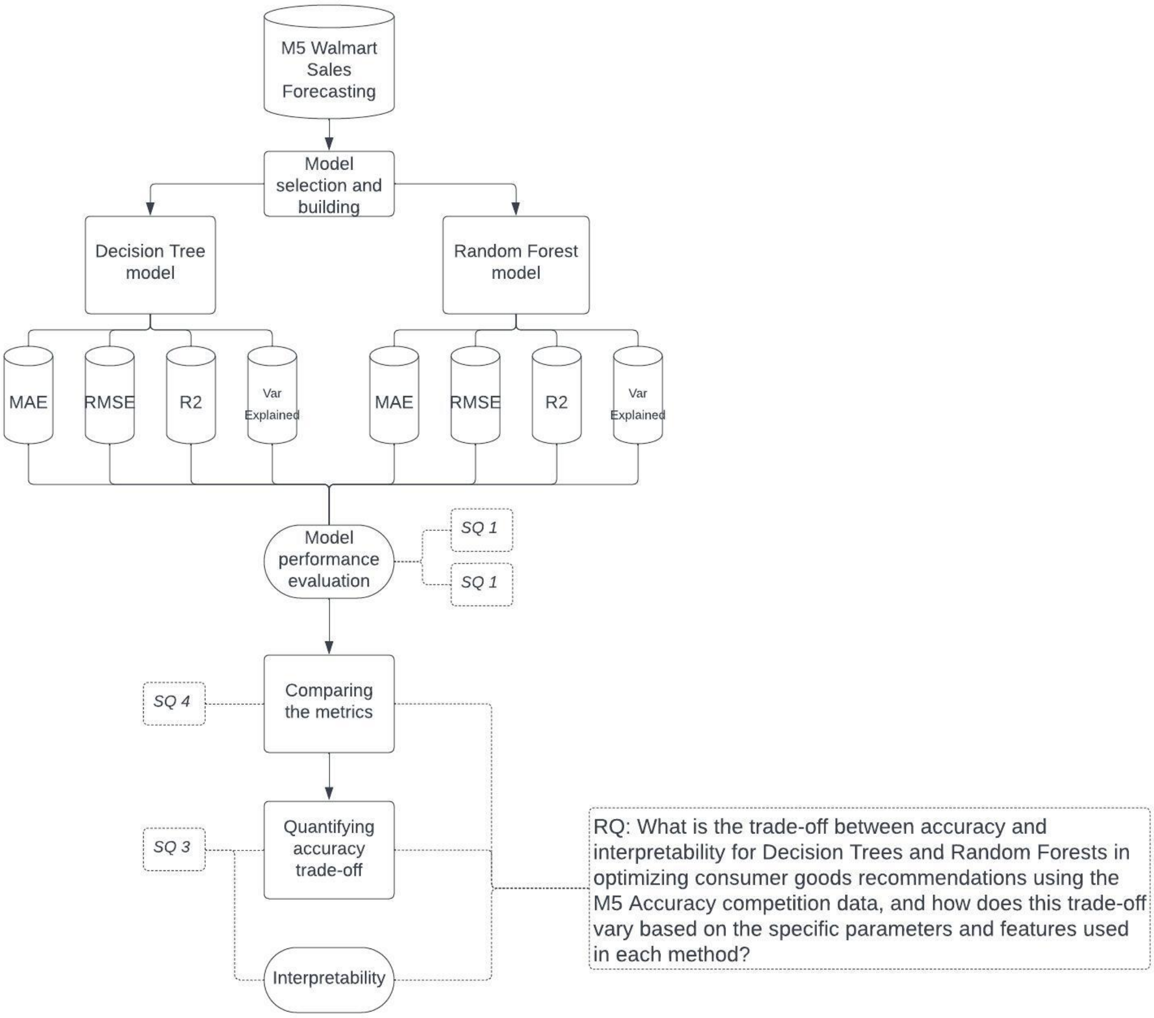
M5 Walmart Sales Forecasting
Model selection and building
Decision Tree model
Random Forest model
MAE
RMSE
R2
Var Explained
MAE
RMSE
R2
Var Explained
Model performance evaluation
SQ 1
SQ 1
Comparing the metrics
SQ 4
Quantifying accuracy trade-off
SQ 3
Interpretability
RQ: What is the trade-off between accuracy and interpretability for Decision Trees and Random Forests in optimizing consumer goods recommendations using the M5 Accuracy competition data, and how does this trade-off vary based on the specific parameters and features used in each method?

Figure 1: Conceptual Research Framework

This paper holds both academic relevance and yield practical implications. This work contributes to the existing literature on using machine learning methods in the FMCG industry. Furthermore, it advances the understanding of consumer preferences and patterns and the implications of interpretability and accuracy. The conceptual framework presented in figure 1 shows us that steps that this research follows in finding the results and drawing the conclusions and recommendations from these results. Also it shows us how we will come to the answer of our main research question.

The main findings show that although Decision Trees offer significant interpretability and a transparent approach to decision making, they generally lag behind Random Forests in terms of predictive accuracy. Our Random Forest model, shows a 99,8% accuracy compared to a 85,5% accuracy from our Decision Tree model. Overall, our findings show that the Random Forest model significantly outperforms the Decision Tree model, with improvements in prediction error metrics ranging from 74% to 90%. R^2 performances are on average approximately 25.39% better than those of the Decision Tree model. This significant improvement marks the Random Forest model's ability to explain the variance in the data and its enhanced predictive accuracy. Unfortunately they show us increased complexity and reduced interpretability. With this paper we tried to quantify this trade-off and offer a nuanced understanding of the implications of choosing one method over another. During this research the importance of tailoring model selection to the specific goals and constraints of the FMCG industry was one of the main focuses.

2 Literature review

This section explores related literature, starting with the advances and disadvantages of the application of Decision Tree methods. The following section concludes with the explanation of the application of Decision Trees and Random Forest in the FMCG industry and discusses the advantages as well as the disadvantages. Next, the review analyzes previous findings regarding the effectiveness of the application of Decision Tree and Random Forest models used for recommendation in the FMCG industry and also discusses the gaps in the literature. By mentioning the disadvantages of the application of both models in the FMCG industry in this part of the paper we also give a slight introduction to the limitations of this research.

2.1 Decision Trees: interpretable but less accurate

Decision Trees provide a transparent and interpretable model that is essential in various decision-making contexts, such as the FMCG industry. They represent decisions graphically, with nodes indicating features, branches decision rules and leaves outcomes, simulating human cognitive processes and providing a simple method for understanding data. However, while their simplicity, a term introduced by Wu et al. (2018), facilitate understanding and manual analysis of predictions, they may come at the expense of accuracy due to challenges in capturing complex relationships in data (Wu et al, 2018).

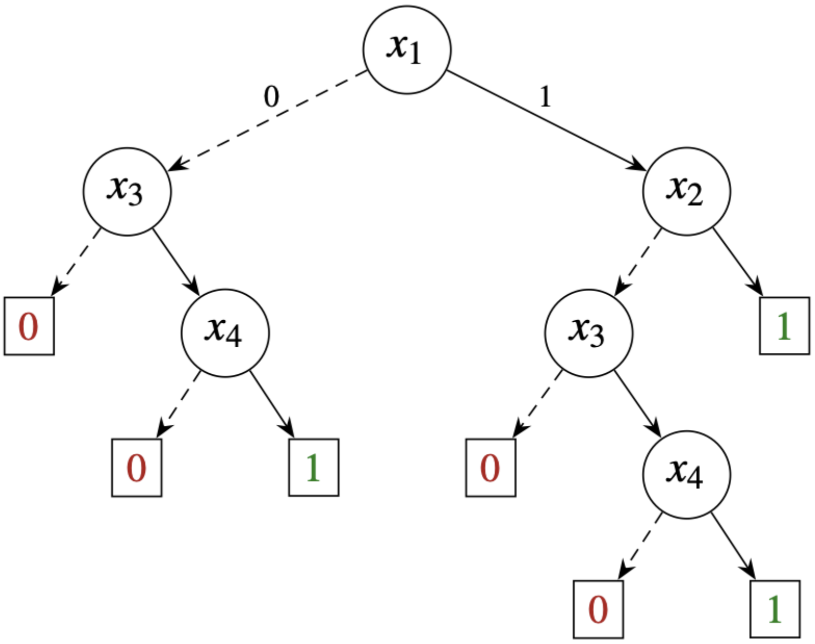

Figure 2: Graphical visualization of a Decision Tree (Izza et al., 2020)

Decision Trees, cited by Song and Ying (2015) are considered powerful statistical tools, simplify complex input-target relationships and facilitate easy interpretation without the need for distributional assumptions, among other advantages. They can deal well with skewed data and robustly handle outliers (Song & Ying, 2015). Which is of high value when the need and want for predictions on a tight time schedule in a fast moving market such as the FMCG industry is big. Assertive decision making is highly appreciated and effective in such an industry. By applying a Decision Tree model decision makers can act fast on the results of the model even though they are robust and not highly accurate. The graphical visualisation of a Decision Tree model is relatively easy to interpret and shows you which way to walk in the decision making process. Figure 2 shows us a decision with n=4 that the path to go to the prediction of value 1 is through *x1*=1, *x2*=0, *x3*=1 and *x4*=1 and it shows us that the outcome will be 1 whichever path you will walk. In this Decision Tree we can see that the path you walk is important and determining for the decision you will make. Safavian and Landgrebe (1991) discuss their effectiveness in various applications, such as radar signal classification and medical diagnosis, due to their ability to simplify complex decision-making processes (Safavian & Landgrebe, 1991). Their flexibility and adaptability in using varied subsets of features and decision rules could fit well with the dynamic FMCG industry, balancing accuracy and interpretability for effective, actionable insights.

Existing literature has focused on improving the interpretability of Decision Trees through various methods, including visualization, rule extraction and feature importance analysis. Craven and Shavlik (1996) have attempted to interpret pre-trained models by constructing Decision Trees to mimic the predictions of pre-built neural networks, without simplifying the network itself (Craven & Shavlik, 1996). In the FMCG industry, where data can be voluminous and multifaceted, the ability of Decision Trees to screen variables and select features is critical. Moreover, their resilience to non linearity, as described by Rokach & Maimon (2014), enables them to navigate the complexity of market trends and consumer behavior without affecting their parameters (Rokach & Maimon, 2014).

2.2 Random Forests: accurate but less interpretable

Random Forests (RF), discussed by Breiman (2001), have established themselves as a model of accuracy in numerous predictive domains by cleverly merging the predictive capabilities of multiple Decision Trees through ensemble learning. Nevertheless, the complexity introduced by this ensemble poses an observable interpretability challenge, hiding the clear distinction of unique feature contributions and specific decision rules (Breiman, 2001). As this research will look into the value of using ensemble methods like Random Forest over Decision because of their complex structure. As Biau and Scornet (2016) argue with their research, that the fundamental architecture of RF's, particularly characterized by the "divide and conquer" strategy - splitting data, creating randomized tree predictors for each subset and then integrating these predictors - merits attention due to its wide applicability and minimal requirements for parameter tuning, thus securing its valued position among various prediction methodologies (Biau & Scornet, 2016). Though, the applicability of Random Forest is very big and the use of them for predictions with high accuracy it does not take away the fact that Random Forest is a black-box model which is not easy to interpret and might be more difficult to use for decision making and finding underlying (consumer preference) patterns.

Random Forests, as described by Breiman (2001), aggregate multiple tree predictors, increasing the generalization error as the forest expands. They show pronounced robustness against noise and provide estimates for error, strength, correlation and variable importance (Breiman, 2001). Belgiu and Drăguţ (2016) add with their research that the effectiveness of the RF classifier is influenced by the sample design and that the feature to measure variable importance has found wide applications, including dimension reduction of hyper spectral data (Belgiu & Drăguţ, 2016). In the FMCG industry, Random Forests, known for their classification and regression accuracy and robust results, seamlessly manage complex data, providing stakeholders with highly accurate, reliable results. Decisions may be taken from these results but based on what? This is something that needs to be further examined with additional analysis's.

The ensemble nature of RF, as Breiman (2001) argues, reduces some of the limitations inherent in single Decision Trees because predictions are generated by a joint effort of majority voting or averaging across the ensemble. Consequently, the RF model, consisting of an amalgam of classification or regression trees without pruning and developed via stochastic selections of training data samples and random feature selections, generally outperforms single-tree classifiers and exhibits a competitive generalization error rate (Breiman, 2001). The methodological diversity, particularly in dealing with predictions within the confines of small samples and multidimensional spaces, underscores the extensive familiarity and application of Random Forest (RF) in various fields. This methodological diversity and the previously pronounced advantage covering the application of Random Forest is one of the reasons why we believe and examine the use of a Random Forest model performing this research. However, on the other side while performing this research we also see also recognition the complexity that RF adds to the interpretability of models (Ghimire et al., 2012; Gislason et al., 2006; Han et al., 2015).

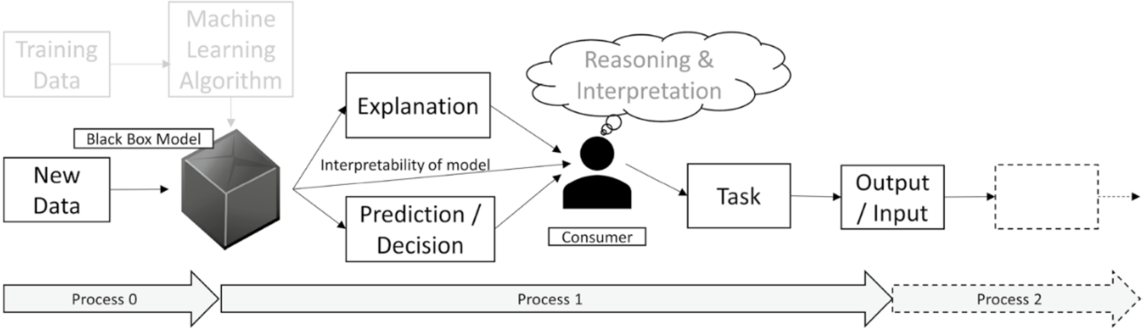

Figure 3: Visualization of Random Forest model application on a dataset - introducing the black box in the process (Hatwell et al., 2020)

Challenges arising from the interpretability of RF models, is something we also see in other industries and in the existing literature and is therefore something we need to consider for this research as well. Breiman (2001) explains with his work that Random Forest in applications such as analyses of medical experiments where deciphering variable interactions is critical to prediction accuracy, require innovative approaches to explaining how the algorithm works (Breiman, 2001). However, when put together in an ensemble, the ability for clear, coherent interpretation is drastically reduced, categorizing RF models as 'black-box' models and embodying a notable sacrifice of interpretability in the pursuit of greater accuracy (Valente et al , 2021). The Random Forest process as presented in Figure 3 shows us the so called Black Box Model which as the name tells is a complex, for those who are unknown of machine learning algorithms, method which they van not interpret. This may hinder the ease of interpreting the valuable and highly accurate predictions which the model provides.

When we look into the complex dynamics of RF and DT and try to compare both models based on the existing literature we can find, we see, as Sakar et al. (2016) argue in their research, a clear trade-off between the superior predictive power that RF often exhibits and the compromised interpretability , especially compared to simpler models such as DT (Sakar et al., 2016). Their statement is something which is in line what we already thought and what we will try take into account and figure out when we do want to apply it in the FMCG industry. Sakar et al. (2016) also intersect with their research that the complex architecture developed by RF might complicate coherent model interactions and decision-making processes, making it more challenging to gain easily interpretable insights, especially in applications that require transparent understanding (Sakar et al., 2016).

When evaluating RF against other machine learning classifiers, several studies have shown notable differences between various facets, including the accuracy of the classification outcome, the training time, and the robustness of the classifier under different training samples or research areas (Belgium & Drăguţ, 2016 ;Gislason et al., 2006; Chan and Paelinckx, 2008; Vetrivel et al., 2015). With Random Forest being an ensemble method which considers not only one tree but a whole Forest of trees has high training time. Training time being high and the input data set being of large size it might be that deploying Random Forest in the FMCG is not the best model for gaining actionable insights in a short period of time and is therefore something we also need to consider computing for this research. However, we can not turn our backs on the fact that RF has consistently demonstrated commendable performance, especially with regard to classification accuracy, compared to decision tree classifiers, Binary Hierarchical Classifier (BHC), Linear Discriminant Analysis (LDA), and Artificial Neural Network (ANN) classifiers (Ham et al ., 2005; Shang and Chisholm, 2014; Chan and Paelinckx, 2008). Which makes them in our opinion interesting for research like ours.

Although RF's are praised for their predictive accuracy, this section showed us that they reveal a non-negotiable trade-off between accuracy and interpretability, burying the transparency of decision-making processes and contributing under a complexity introduced by the ensemble of Decision Trees (Breiman 2001). This section also showed us that rather, Decision Trees illuminate clear, understandable decision paths, and possibly at the cost of reduced accuracy. This includes not only a technical challenge, but also an ethical challenge for the responsible use of machine learning in practice.

2.3 Application of Decision Trees and Random Forest in the FMCG Industry

Existing literature shows us that multiple researchers have used these machine-learning techniques to discern factors that influence consumer preferences, purchase decisions and brand loyalty by revealing crucial variables and decision rules that determine consumer behavior, enabling targeted marketing initiatives and better product recommendations.

From existing literature we found that Wu and Zheng (2015) present sales forecasting models using machine learning techniques that outperform traditional models in predicting product demand, especially in volatile markets with short life cycles such as fast fashion retail (Wu & Cheng, 2015). Furthermore, Tsoumakas (2018) argues that machine-learning techniques, from simple approaches such as the moving average and ensemble methods using various learning algorithms, offer efficiency and adaptability for forecasting time series data compared to conventional statistical methods (Tsoumakas, 2018). In FMCG, the ability of these techniques to process data robustly and deal with relevant variables can enhance sales forecasting, thus the expected impact of machine learning on future sales forecasting and demand can contribute significantly. Working on this section we will study the applicability of machine learning methods in the FMCG industry. With the use of the M5 Walmart Sales Forecasting dataset uses for this research and being time series data use of the methods can provide benefits.

That is how we found that in exploration of the FMCG industry, specific characteristics and challenges of the industry as well as the methods we use for analyzing are particularly interesting. Nozari et al. (2022) serve as a crucial reference, discussing the symbiosis between rapid technological developments and the resulting challenges within the industry. In particular, the FMCG industry is praised for its adept and intelligent supply chain management, which stems from the intrinsic requirement to produce and distribute goods quickly in accordance with consumer demand (Nozari et al., 2022).

To look into the statement that Random Forest and Decision Tree methods might be advantageous existing literature of Tallaro et al. (2019) add to our research that the applicability of Machine Learning in optimizing inventory management and demand forecasting in the FMCG industry (Tallaro et al., 2019). In contrast, Günesen et al. (2021) investigated customer turnover prediction and retention within the FMCG industry using a mix of Machine Learning algorithms, which provided insightful models for skillful marketing and operational strategies amidst a fluctuating consumer base (Günesen et al., 2021). Panjwani et al. (2020) offered a practical approach to predicting sales using DT and RF, with accuracy's of 83.86% and 81.21%, respectively, confirming their effectiveness in providing nuanced, actionable insights into consumers' buying patterns within the retail industry (Panjwani et al., 2020). This existing literature outlines a spectrum of Machine Learning applications within the FMCG industry and provide a comprehensive perspective on how DT and RF provide actionable predictions in consumer behavior. Which we can and need to take into account for our own research en the corresponding results.

Using Decision Tree and Random Forest algorithms to analyze customer behavior, especially in the dynamic world of retail and e-commerce, presents unique challenges and limitations. One critical limitation, identified by Kim et al. (2005), has to do with the methodology's limitation of analyzing only two data sets simultaneously, requiring a more complex, recursive approach for analyzing three or more data sets. Moreover, the management of different data types, especially the transition from continuous to discrete values, requires additional pre-processing steps, potentially leading to information loss or additional analysis effort (Kim et al., 2005). In addition to procedural and data management challenges, the intrinsic variability in consumer behavior, especially in online shopping environments, requires consistent refinement and adaptation of Decision Tree and Random Forest models to maintain their predictive accuracy and relevance amid rapidly changing consumer behavior in the retail and e-commerce sectors.

2.4 Summary and Gaps in the Literature

From the literature section we found that Decision Trees offer interpretability but may come at the expense of accuracy, while Random Forests offer high accuracy but may be less interpretable. Decision Trees have proven useful in the rapidly changing consumer goods industry for understanding consumer behavior. However, there is a gap in understanding the specific industry contexts, data types and evaluation metrics that influence the trade-off between interpretability and accuracy. With this research we therefore try to address these gaps by exploring the optimal approach for consumer goods recommendations in the FMCG industry, we consider the combination of Decision Trees and Random Forests by quantifying the trade of between accuracy and interpretability of both methods.

In this research, we explore the roles and challenges of Decision Trees and Random Forests in predictive modeling, particularly in the FMCG industry, based on existing literature in machine learning algorithms.

Decision Trees and Random Forests, both rooted in tree-based methodologies, cater to different aspects of predictive analysis. Decision Tree is famous for its simplicity and interpretability (Song and Ying, 2015; Safavian and Landgrebe, 1991), while RF stands out for its high accuracy and robustness, especially with noisy data (Breiman, 2001; Belgiu Drăguţ, 2016). This simplicity versus accuracy duality forms the foundation for discussions on their application in domains like the FMCG industry, where transparency and precision are critical.

However, literature, including studies by Jadhav Channe (2016) and Craven Shavlik (1996), points out that while DT's interpretability aids decision-making, it may struggle with complex data relationships. Conversely, RF, despite its competitive performance and ability to manage complex data patterns, often appears as a 'black-box' model, making it challenging to extract interpretable insights (Wu et al., 2018).

In the FMCG industry, algorithms play a crucial role in exploring consumer behavior and enhancing decision-making through sales forecasting and consumer behavior analysis. They address data management challenges and rapidly changing consumer data effectively (Khade, 2016; Kim et al., 2005). Nonetheless, both models' advantages and shortcomings offer opportunities for our current and future research.

Consensus on interpretability versus accuracy of Decision Trees: In our literature review, we find that Decision Trees are popular due to their high interpretability and user-friendliness. They simplify complex datasets into understandable decision paths (Jadhav Channe, 2016; Craven Shavlik, 1996), making the data more accessible and actionable. However, the same simplicity that makes Decision Trees interpretable can limit their predictive performance. This limitation becomes apparent when dealing with complex data that involves multidimensional relationships, which Decision Trees tend to oversimplify. Therefore, while Decision Trees improve interpretability, they may sacrifice accuracy in complex predictive tasks. This highlights the trade-off between interpretability and accuracy when choosing data analysis models.

Consensus on accuracy versus interpretability of Random Forests: We found that the literature also converges to a consensus that emphasizes that Random Forests (RF) exhibit high levels of accuracy and robustness, particularly when managing complex data and noise (Breiman, 2001). Yet this high accuracy often leads to a sacrifice of interpretability because Random Forests, by using numerous Decision Trees and different node splitting rules, make deciphering individual feature contributions or specific decision paths quite complicated. This duality of increased accuracy and decreased interpretability confirms the statement.

Ambiguity in the application of Decision Trees in the FMCG industry: The statement suggesting that DT may be more useful than RF in providing useful insights into factors determining consumer behavior in the FMCG industry finds both support and possible disagreement in the existing literature, we found. As Song and Wing (2015) discuss that Decision Trees interpretability can certainly provide a more direct and clear understanding of the importance of variables and decision paths in consumer behavior, there is also evidence that RF, with its superior accuracy, can provide better predictive analytics for scenarios that require precision in predicting and navigating complex consumer data (Song and Ying, 2015). Thus, while DT offers transparent insights, it is not unequivocally established that they are categorically more useful than RF in the FMCG industry.

In our literature review, we've identified significant gaps and limitations in existing research on Decision Trees (DT) and Random Forests (RF), as well as opportunities for further investigation.

- **Application in Diverse Sectors**: The literature predominantly focuses on the use of these algorithms in the FMCG sector, leaving other sectors like finance or telecommunications underexplored. Research should extend to diverse industry contexts.

- **Handling Diverse Data Types**: Existing studies often overlook the effectiveness of DT and RF with diverse data types, such as text and images, especially in scenarios involving small and imbalanced datasets.

- **Comprehensive Model Evaluation**: While some evaluation of these models exists, it lacks depth in terms of using various metrics and conducting benchmark comparisons with alternative models. A more versatile evaluation approach is needed.

- **Implementation and Optimization Challenges**: Implementing and optimizing DT and RF can be challenging, particularly in context-specific applications. Scalability and computational efficiency should be addressed.

- **Interpretability Enhancement**: Despite RF's accuracy, its complexity can hinder interpretability, particularly in critical areas like healthcare. Research should focus on improving interpretability.

- **Bias Mitigation and Fairness**: There's limited exploration of bias mitigation, fairness, and the robustness and security of DT and RF against vulnerabilities.

- **Real-World Deployment Challenges**: Research should delve into the practical challenges of deploying DT and RF in real-world scenarios. Strategies for integrating these models with emerging technologies and adapting to evolving consumer behavior need exploration.

In light of the identified gaps and limitations of the current literature, with our research we are determined to help in the exploration to overcome these shortcomings, especially in the area of applying Decision Trees and Random Forests algorithms in different sectors and diversified data types. With the architecture of our research analysis we contribute to the exploration and evaluation of these models across different data types, while also extending the evaluation metrics and benchmarking strategies to give a as detailed an overview as possible of the model, thereby increasing the reliability and applicability of the findings.

As we navigate you through these exploratory paths, the research will delve deeply into challenges and strategies for real-world deployment that deploy Decision Tree and Random Forest with emerging technologies. Hereby, creating a roadmap for navigating the dynamically changing landscape of consumer behavior with advanced machine learning algorithms. The contributions of this research not only being to elevate the academic aspect by addressing the aforementioned gaps, but also pioneer new paths for the hands-on and ethical application of Decision Tree and Random Forest across a spectrum of industries, data types and real-world scenarios, strengthening its centrality in both the academic and practical worlds.

3 Data

3.1 Introduction to M5 Walmart Sales Forecasting Data

The M5 Walmart Sales Forecasting data, a very comprehensive collection of sales and product data employed in this study, originates from Walmart, which is pivotal for representing the FMCG industry. Leveraging historical sales data across a multitude of products and stores, it serves as a real-world repository to explore and evaluate predictive model building, notably through Decision Trees and Random Forests, in the context of consumer goods recommendations. Which is the main reason why we used it for to perform our research. To introduce you to the data set we explore the content. By giving you an insight into the data sets most relevant content, the content which we will be using within our research, you get familiar with it.

In Figure 4 we see the time series of Walmart's sales over across different stores in different states over the range of the data set which starts on 29th of January 2011 and ends on 22nd of April 2016 (Mathur, 2020).

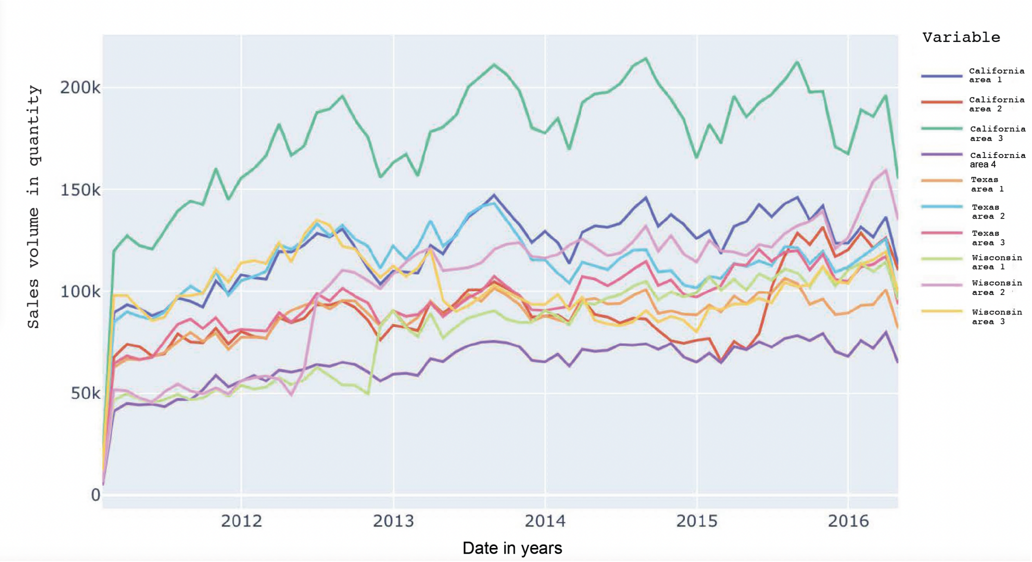

Figure 4: The time series of sales across different stores over different states(Mathur, 2020).

3.2 Structural Composition

The M5 Walmart Sales Forecasting dataset entails daily sales data spanning several years, providing a substantive chronological breadth to discern trends, cyclicality, and potential anomalies pertinent to the FMCG industry. At its core, the dataset contains hierarchical categorizations enveloping products, enabling nuanced insights into varied product categories, and contextual factors. A lot can be taken from the data set but within this research we will use the sales data and let the other area's of the dataset to rest. That is why we want to show you Figure 5 where you can see the time series line chart of all the sales over the range of the data set.

While the M5 Walmart Sales Forecasting data provides a comprehensive foundation, it is pivotal to acknowledge its constraints. Potential limitations such as missing data, outliers, or

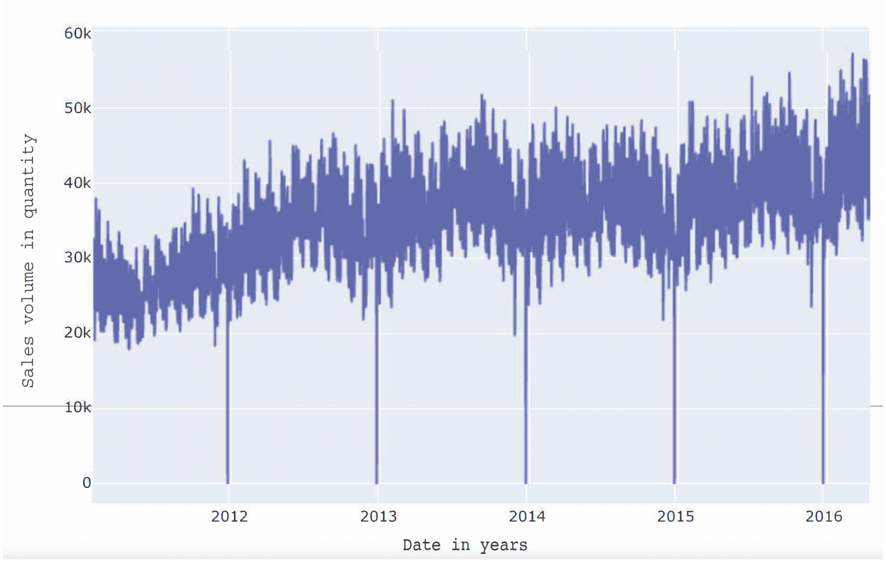

Figure 5: Time series line chart illustrating the long-term aggregated sales patterns(Mathur, 2020).

inherent biases should be critically evaluated to ascertain the robustness and generalizability of the derived models and findings.

Understanding temporal dynamics is important to understand seasonalities, trends, and other temporal associations within the FMCG sales. The extended duration of the dataset facilitates an in-depth temporal analysis, pivotal for training and testing predictive models, and understanding the chronological evolution of consumer purchasing behaviors. This is why we consider the data set of high value for the performance of our models. In our analysis we do not take into account all the variables, one of our variables of interest is the category ID variable (cat_id). The Department Id shows us the different categories within a department and that is why we want to show you Figure 6, in which you see a bar chart of the sales distribution across various categories over time.

3.3 Explanatory Variables and External Factors

The dataset is enriched with additional variables, such as prices, promotional events, and potentially correlated external factors, that can be instrumental in understanding and modeling the intricate dynamics of sales data. By analyzing how external factors (e.g., holidays, seasons) impact sales trends, and incorporating these insights into predictive models, provides a holistic understanding of multiple influences affecting purchasing behaviors (Mathur, 2020).

3.4 Data Pre-Processing

3.4.1 Data set

For our research we did not made use of all the available data sets. For the analysis in R we used the train validation data set. The pre-processing of the data was not needed since the data was cleaned already. For example, NA values were already removed. From this train validation

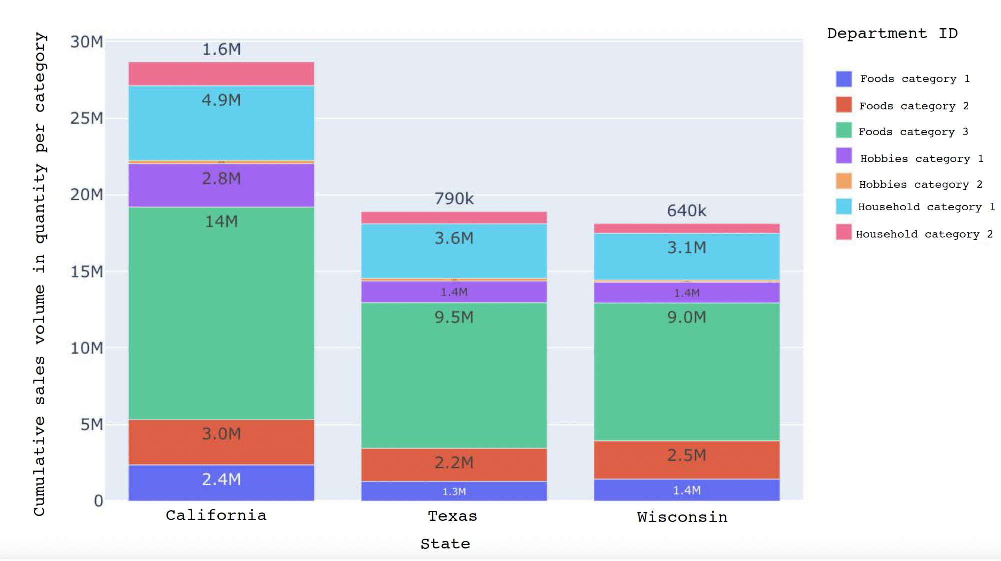

Figure 6: Stacked bar chart illustrating the sales time series across various store departments over time, per state(Mathur, 2020).

data set we took a smaller sample to work with.

3.4.2 Subset

To be more specific we kept all of the 30.490 the observations from this data set but we removed a few variables from the data. For example we removed, ‘id‘, ‘item_id‘, ‘dept_id‘, ‘store_id‘, ‘sate_id‘ and we kept ‘cat_id‘. Since we want to observe the overall sales in a specific category of Walmart’s department and are not interested in which ID, Item ID, Department ID, Store ID or State ID have an influence on predicting our target variable we left them out. We did however used the whole time range and therefore created a data set which contains the ‘cat_id‘ variable and the day variables ranging from d_1 to d_{1912}.

3.4.3 Target Variable

Our target variable, or our variable of interest is the sum variable. The sum variable we created by computing the sum of all the sales over the 1.912 for each observation within a specific product category. The sum variables ranges from 10 items to 250.502 items sold over the whole time period.

3.5 Subgroups

In order to quantify the trade-off as we aim to answer our research question we divided the dataset into ten subgroups. The ten subgroups where composed in random order and without repetition. Then we trained ten individual Decision Tree models for each subgroup using R. Next we will do the same for the Random Forest models. These models have the same features and parameters as the default models in R. The ten subgroups will provide ten different packs of results and therefore comparative material, which is crucial for quantifying the trade-off of our models performances.

4 Methodology

As we are interested in the sum of the sales over the specific time frame our target variable is continuous. Given the quantitative aspect of the target, employed we employed regression models to forecast the sum of sales variable accurately. Which we choose for because of the capability of regression techniques to model and predict continuous outcomes, making them suitable for performing our research analysis. For conducting our research we used a Decision Tree model and a Random Forest model. Therefore, this section explains the choice behind the preference for regression analysis, specifically through the use of Decision Tree and Random Forest Regression models, to capture the intricate relationships between the features and the continuous sales sum.

4.1 Decision Tree

The Decision Tree methodology, widely acclaimed for its versatility in predictive modeling, ingeniously partitions the feature space into homogeneous regions, effectively facilitating nuanced data classification and prediction (Song & Ying, 2015). In order to perform the Decision Tree we used the standard Decision Tree model, which is further explained in A.

After deploying the DT on the M5 Walmart Sales Forecasting dataset. Parameter settings and thresholds were set to default values so that the model, while application in this research, not only identifies well-nuanced relationships between variables, but also remains robustly interpretable to industry stakeholders. The resulting tree, attempts to strike a balance between model interpretability and prediction accuracy, serving as a powerful tool for uncovering and explaining the many factors that shape consumer behavior.

Applied to the M5 Walmart Sales Forecasting dataset, the Decision Tree methodology is well suited for uncovering subtle relationships among variables that influence purchasing decisions (Panjwani et al., 2020). The interpretability of the methodology proves crucial in the FMCG industry, where quick and informed decisions are critical to success. Here, understandable insights derived from Decision Trees can enable decision makers to create targeted marketing strategies, optimize inventory management and create personalized customer experiences (Tallaro et al., 2019). However, the challenge of maintaining accuracy while maintaining interpretability requires a strategic approach. By carefully tuning parameters during tree construction and applying pruning techniques after construction, making it possible to strike a balance between predictive performance and interpretability.

4.2 Random Forest

After performing our Decision Tree analysis we performed our Random Forest analysis on the same train and test datasets, in order to get a mean prediction (regression) of the individual trees. A more extensive explanation of Random Forest is explained in A.

4.2.1 Variable Importance and Feature Selection

One of the most compelling facets of the Random Forest algorithm pertains to its intrinsic capability to perform feature selection. The process of selecting a feature subset of the original feature set for tree-node split is described as randomization technique. To classify a new instance, RF puts the new instance down each tree in the forest. Each tree provides a predicted label as a vote for prediction. RF chooses the classification with the most votes (Gregorutti, et al., 2017).

The variable importance $VI(x_j)$ for the feature x_j in a Random Forest is calculated as the average reduction in accuracy of the model after permuting the feature's values. For each tree t in the forest, the difference in prediction error is computed before and after permuting x_j. This difference is summed over all instances in the out-of-bag (OOB) sample. The resulting sums across all trees n_{tree} are then averaged and normalized by the size of the OOB sample $|OOB|$(Gregorutti, et al., 2017).

$$VI(x_j) = \frac{1}{n_{tree}} \sum_{t=1}^{n_{tree}} \left[\sum_{i \in OOB} I(y_i = h_t(x_i)) - \sum_{i \in OOB} I(y_i = h_t(x_i^{(j)})) \right] /|OOB|$$

Here, I is an indicator function that equals 1 when the predicted value $h_t(x_i)$ equals the true value y_i, and 0 otherwise. The term $x_i^{(j)}$ denotes the feature vector x_i with the j-th feature permuted.

When a VIM method is performed, each feature is designated with an importance score. Thus a feature ranking can be obtained by ordering the importance scores (Gregorutti, et al., 2017).

4.2.2 Model Optimization and Hyperparameter Tuning

Optimization of the Random Forest model entails the strategic tuning of hyperparameters. A few pivotal hyperparameters include the number of trees B in the forest, the maximum depth of trees, and the minimum samples per leaf (Breiman, 2001). Tuning is typically performed via grid search or randomized search approaches, aiming to navigate the model towards its optimal performance through the exploration of various hyperparameter combinations. Given the substantial depth and intricacy of the M5 Walmart Sales Forecasting data, the research entails tuning, ensuring that the resultant model is robustly aligned with the underlying data distributions and interactions.

Further discussions in this subsection include a mix of practical applications, theoretical intricacies and strategic considerations, interweaving Random Forest with the specificity of the FMCG industry. Following the matching of algorithmic parameters to the data, the application of the Random Forest model to the M5 Walmart Sales Forecasting dataset reveals layers of consumer behavior patterns and nuanced variable interactions, fostering an enriched, data-driven decision-making environment within the FMCG industry.

4.3 Evaluation Metrics

To conclude our analysis and our research the evaluation and comparison of predictive models are crucial steps to determine their efficacy and reliability. To achieve this, we computed four statistical metrics, each designed to capture different aspects of model performance. These metrics allow us to quantitatively assess and compare the accuracy and consistency of our performing models. Further explanation on the evaluation metrics we used in our research to compare the model performance can be found in A.

The Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE) are suitable for evaluating the average error magnitude and the error distribution's spread, respectively. While MAE provides a straightforward measure of average prediction error, RMSE gives additional weight to larger errors, making it particularly sensitive to outliers. This difference in sensitivity makes them complementary metrics for assessing model performance (Hodson, 2022).

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

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

To give more insight in the models ability to explain the target variable, the Coefficient of Determination, denoted as R^2 metric offers insights into the proportion of the target variable's variance that the model accounts for. High R^2 values indicate that the model explains a significant portion of the variance, suggesting a good fit to the observed data (Nakagawa & Schielzeth, 2013).

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{3}$$

By analyzing these metrics in tandem, we can form a comprehensive view of model performance, not only in terms of prediction accuracy but also in how well the models generalize to new, unseen data. This multifaceted approach to model evaluation enables us to select the most appropriate model for deployment in practical applications, ensuring both reliability and robustness in predictive tasks.

5 Analysis and Results

In the analysis and results section we will present to you the results and how we conducted these results while performing this research. As previously mentioned we focused on quantifying the accuracy and interpretability trade-off using Decision Tree and Random Forest methods. In order to make consumer goods recommendations for the FMCG industry which we have as our industry of interest we used the M5 Walmart Sales Forecasting data set.

5.1 Decision Tree Analysis and Results

First we applied the Decision Tree method to the dataset of interest. We did not make use of all the variables in the data set but focused on only the relevant variables by selecting specific columns from the dataset, including the 'cat_id' variable and a range of other features: columns 7 to 1920, which represent the days when the sales of the specific products were tracked. In order to quantify the trade-off and as previously mentioned in the data section of this paper, we divided the dataset into ten subgroups and training individual Decision Tree models for each subgroup using the rpart package in R. These models adhere to the same features and parameters as the Random Forest models.

The results from the Decision Tree analysis where processed into a table to make it more clear. On first sight we found that the results and the performance of the Decision Tree model were quite good and accurate. In order to compare the results and to gain more insight in if these results and the model performance is actually good and reliable we computed model performance scores. To be more specific we computed the MAE, RMSE and R^2. MAE and RMSE values were comparable, underscoring the models' ability to make predictions closely aligned with actual values. Moreover, consistently high R^2 values affirmed that Decision Trees effectively explained a substantial portion of the variance in the target variable. Also, we find that the R^2 are surprisingly high, something we will study continuing this section.

We created table 1 which summarizes the performance of the Decision Tree model applied to the 10 different subgroups of the dataset. Similar to the Random Forest analysis, the model's effectiveness in predicting the target variable is assessed using multiple metrics.

Table 1: Decision Tree Results

Decision Tree Results			
Subgroup	MAE	RMSE	R^2
1	1173.9	4109.3	0.52
2	1154.7	2521.9	0.79
3	1019.5	2029.9	0.83
4	1096.2	2289.2	0.81
5	1011.8	2079.3	0.81
6	1096.6	2316.6	0.77
7	1028.5	1798.5	0.85
8	1184.3	2850.6	0.77
9	1049.6	2056.3	0.82
10	1171.4	2781.4	0.80

Mean absolute error (MAE): MAE quantifies the average absolute differences between the predicted and actual values. The MAE for the Decision Tree model across subgroups ranges from 1011.8 to 1184.3, with subgroup 5 yielding the lowest MAE at 1011.8, indicating better predictive accuracy. The MAE values are quite high and not as we hoped in performing this analysis. However there are several possibilities on why these values are quite high for this Decision Tree analysis which we need to consider. First of all, there might be overfitting in the training data as they can create complex, deep trees that perfectly fit the training data but may generalize poorly to new, unseen data. When considering the specific data set we used for this research there is a possibility that Decision Trees might struggle with complex relationships in the data.

RMSE (Root Mean Square Error): RMSE provides a measure of prediction error, with larger errors being penalized more heavily than smaller ones. Within this analysis, RMSE values ranged from 1798.5 to 4109.3. Subgroup 7 emerged with the lowest RMSE, valued at 1798.5, indicating relatively accurate predictions and minimal error. For the RMSE applies the same explanation. We hoped for lower values, but the high values might be caused by the possible overfitting Decision Tree models are known for the data size and complexity. Also Decision Trees are more sensitive to outliers and since outliers are not removed from the data set and might even be added to the data set this can have a significant influence on the model performance scores. The target variable being the sum of the product category over all the 1941 days and in large range differs because of this.

Coefficient of Determination (R^2): R^2 illustrates how well the model elucidates the variance within the data, with higher values indicating a better fit. The values in the attached table indicate that Subgroup 7 has the highest R^2 value at 0.85, meaning that the Decision Tree model fully explains a significant portion of the variance in the data for this subgroup. When considering the relatively high values of MAE and RMSE for our Decision Tree model analysis we were as mentioned before surprised that the R^2 is relatively high. And are positive to observe this outcome of the model, since it means that the model is still quite good at predicting the sum of products sold and the model does not under perform on this part. It also confirms the previous explanation on the high values of MAE and RMSE since the target variable is the sum and the range of the sum is large.

Based on the results of out Decision Tree analysis we found that it shows different predictive performance across subgroups. And the results are surprising but therefore not less interesting. Since the MAE and RMSE are showing us significantly inferior values, being quite high, but on the other hand the R^2 showing us surprisingly high values which are more valuable since the Decision Tree method is known for performing inferior in for example the are of overfitting compared to other machine learning methods. Subgroup 5 achieved the lowest MAE, indicating high predictive accuracy, while subgroup 7 showed the lowest RMSE and the highest R^2, indicating high predictive accuracy and comprehensive explanatory power with respect to variance in the data. The simplicity and interpretability of Decision Trees provide a transparent, intuitive visualization of decision paths, distinguishing them from more complex models such as Random Forests.

5.1.1 Interpretability of Decision Tree results

From our Decision Tree analysis we found the prediction values and provided the table to demonstrate the model performance. However, with presenting the model performance metrics we will be able to quantify the accuracy trade-off, but not the interpretability. In order to gain insights from the model analysis and the results we need to focus on making the results insightful and visual as well. That is why we provided the following Decision Tree graph in R. The decision tree, as illustrated in Figure 7, presents an overview of the sales patterns observed in the product dataset, based on different day-specific sales figures. The tree divides the product data into distinct, homogeneous groups, each characterized by unique sales patterns and averages.

In Figure 7 we can see that the root node provides an initial split based on the sales volume of 48 units on day 1,260 ($d_{1260} < 48$), thereby mark the significance of this specific day in distinguishing product sales trends. As the data set we used is a time series data set, we find that the Decision Tree is slightly more difficult to interpret or on first sight needs a little bit of additional explaining. The Decision Tree as we see it in Figure 7 is showing us the importance of the variables presented in the data set. As assumed beforehand we will find that specific sales days are important in the prediction of the target variable. The root node shows us that an average sale of 2,331 units of a specific product, calculated over all 3,049 products, is moderately popular within the analyzed time frame.

Subsequent nodes reveal further breakdowns in the analysis. The 'Yes' child node, representing products with sales beneath the 48-unit threshold on day 1,260, taking into account all products in the dataset, with a diminished sales average of 2,128 units. The accompanying split criterion, $d_{561} < 10$, shows us that for when the sales on day 1,260 are indeed lower than

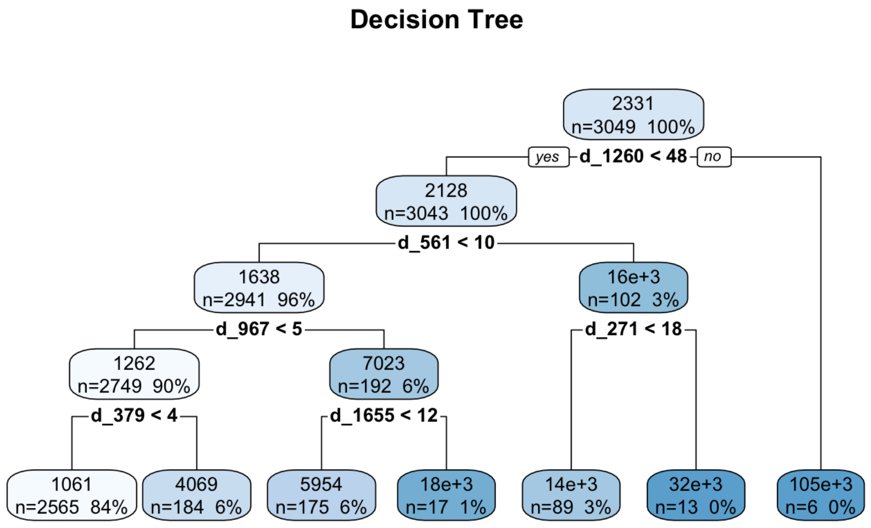

Figure 7: Decision Tree Visualization discussing Product Sales Patterns.

48 items, that a new path might be considered and other days and sales volumes are of big influence. Overall, it shows us the days we need to consider important in the prediction of the sum of sales for a specific product category.

In contrast, the 'No' child node represents only 6 products, with a greatly increased average sales volume of 105,000 units. This sudden increase underscores the central role these products play in the sales portfolio and potentially act as revenue pillars during the set period. And is something we need to take into consideration for future research. Since the sudden increase of these products is something we do not want to focus on or want to investigate while performing this research. However, it is important to realize that the mentioned days in Figure 7 are of big influence for the prediction of the sales quantity.

This decision tree framework provides crucial insights into the dynamics of product sales within the dataset and marks potential avenues for broader research and strategic interventions, particularly targeting critical days such as d_{561} and d_{1260}, and the outstanding sales performance of specific outliers. As mentioned in the previous paragraph, while performing this research and analysing and interpreting these results we do not want to focus on the volumes but we do want to focus on the importance of the variables which have a influence on the target variable. As can be seen in Figure 7 there are some specific days which we consider important in the prediction of our target variable. And we consider these variables important. As for the interpretability of Figure 7 discussion might exist in how interpretable the results are. When focusing on the importance of the variables which in this Decision Tree figure are the specific days, can this discussion quickly be eliminated.

5.2 Random Forest Analysis and Results

For the second part of our analysis we applied a Random Forest method to the dataset. Same as for the Decision Tree method we did not make use of all the variables in the data set but focused on only the relevant variables by selecting specific columns from the dataset, including the ‘cat_id‘ variable and a range of other features: columns 7 to 1,920, which represent the days when the sales of the specific products were tracked. In order to quantify the trade-off and as previously mentioned in the data section of this paper, we divided the dataset into ten subgroups and training individual Random Forest models for each subgroup using the Random Forest package in R. Random Forest models were subsequently trained on each subgroup individually, with 100 trees in each forest. These models have the same features and parameters as the Decision Tree models. Approximately the same analysis was performed for the Random Forest models as for the Decision Tree models, which is crucial for comparing the two model performances and quantifying the trade-off between the two model outcomes.

As previously explained in the Decision Tree results section we analysed the model performance in order to gain reliable insights from the models results. We also computed the MAE, RMSE and R^{2}for the Random Forest models. The results of the Random Forest analysis consistently showed significant high predictive accuracy. We found that for all subgroups, the Random Forest model consistently shows relatively low values for Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE), confirming that the model's predictions consistently approximated the true values. Moreover, the R^2 values consistently indicated that a significant proportion of the variance in the target variable was explained by the Random Forest model.

To provide an insightful overview of the model performance values Table 2 summarizes the metrics of the performance values of our Random Forest model applied to the 10 different subgroups of the dataset.

Table 2: Random Forest Results

Random Forest Results			
Subgroup	MAE	RMSE	R^2
1	135.6	534.9	0.99
2	134.2	628.1	0.99
3	154	1383.3	0.98
4	139.5	634.3	0.99
5	148.7	1430.3	0.98
6	138.7	726.7	0.99
7	125.3	419.4	0.99
8	127.3	504.2	0.99
9	135	608.1	0.99
10	135.5	656.4	0.99

MAE measures the average absolute difference between the predicted values and the actual values. And is for all subgroups, ranged from approximately 125.3 to 154. Lower MAE values indicate better predictive accuracy, and subgroup 7 achieved the lowest MAE at 125.3. The low values of MAE are positive and also expected since the Random Forest model handles overfitting, outliers and data complexity better compared to Decision Tree. Still the results from the model performance do surprise us and since they are not near 0 they need to be seriously considered in interpreting the models predictions.

The RMSE is another measure of prediction error, with larger errors being more heavily weighted. In this analysis, the RMSE ranges from about 419.4 to 1430.3. Subgroup 7 stands out with the lowest RMSE of 419.4. Compared to the model performance results from the Decision Tree model we see that our Random Forest model is performing better on MAE and RMSE. With the prediction errors being lower we can say that the model is handling the data better and might provide us with more reliable predictions. Hence we do need to take into account that our Random Forest was expected to perform better on this area because the method is simply handling overfitting, outliers and data complexity better and the function of the ensemble methods is to process the previously mentioned areas better.

The R^2 value indicates how well the model explains the variance in the data. Values close to 1 indicate a strong fit. Subgroup 7 has the highest R^2 value of about 0.99, suggesting that the model explains a significant portion of the variance in this subgroup. As expected our Random Forest model explains predicts the target variable almost completely with the highest R^2 being very close to 1. In all of the 10 subgroups the R^2 value is very high and very close to 1, which indicates that our model is quite capable of handling the data and providing us with highly accurate predictions. As previously mentioned, we do not want to focus on the results of the predictive values our models provide us with, but the underlying patterns are more important in recovering insights from the data and finding the variables that are important in predicting the target variables.

In summary, we found from the results of out Random Forest analysis, that the model shows strong predictive performance for all subgroups, with subgroup 7 consistently achieving the lowest MAE, RMSE and the highest R2. This indicates that the model effectively captures the underlying patterns in subgroup 7's data, making it a notable subgroup for further analysis or attention. However, it is crucial to recognize that although Random Forest excels in predictive accuracy, interpretability poses a notable challenge.

5.2.1 Interpretability of Random Forest results

From our model performance results we can quantify the accuracy trade-off as we wanted and we can also compare the model performance of both the Decision Tree and the Random Forest model, but we can not compare the interpretability. Since, we want to provide consumer goods recommendations for the FMCG industry we do need to gain insights from both models. In the previous section on our Decision Tree analysis we interpreted the Decision Tree graph so get an insight on which variables are responsible for the steps in the Decision Tree path. The same we wanted to do for our Random Forest model. Since, interpreting a Black-Box model like Random Forest, choosing which method for interpretation is most relevant was a challenge. However, as explained previously with our research we want to focus on giving insights in the important variables for predicting the sales for specific products and we do not want to focus on the sales volumes the prediction provides us with. In this paper the interpretation of our Random Forest model is enriched by the inclusion of a variable importance plot in order to give insights in which variables can be of more influence in prediction. The variable importance plot is crucial in this context, particularly focusing on the permutation importance measure. This approach, as introduced by Strobl et al. (2008), effectively demonstrates the significance of each predictor by assessing the change in model accuracy following the random permutation of the predictor's values, thereby simulating its absence from the model (Strobl, et al., 2008).

In the following two Variable Importance Plots we can see the variables which are considered important in predicting the target variable from our data set according to the results of our Random Forest model. In Figure 8 we find the important features on a default of the Variable Importance Plot, with other words, the top 1o most important variables for predicting our target variable. And in Figure 9 we find the same top 10 important Features including `cat_id`. In the second Variable Importance Plot we added `cat_id` because we wanted to compare the importance of the product categories with the other important variables, and from here conclude whether `cat_id` is or is not important.

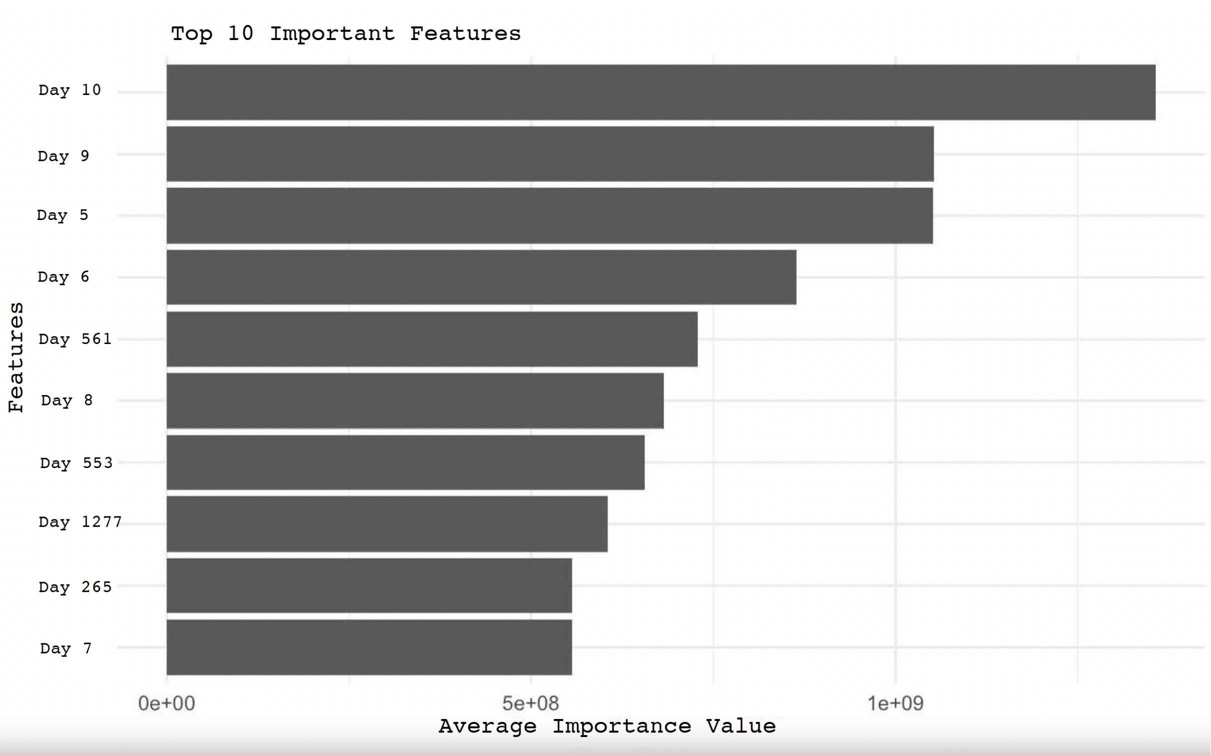

Figure 8: Feature Importance Plot for the 10 most important figures

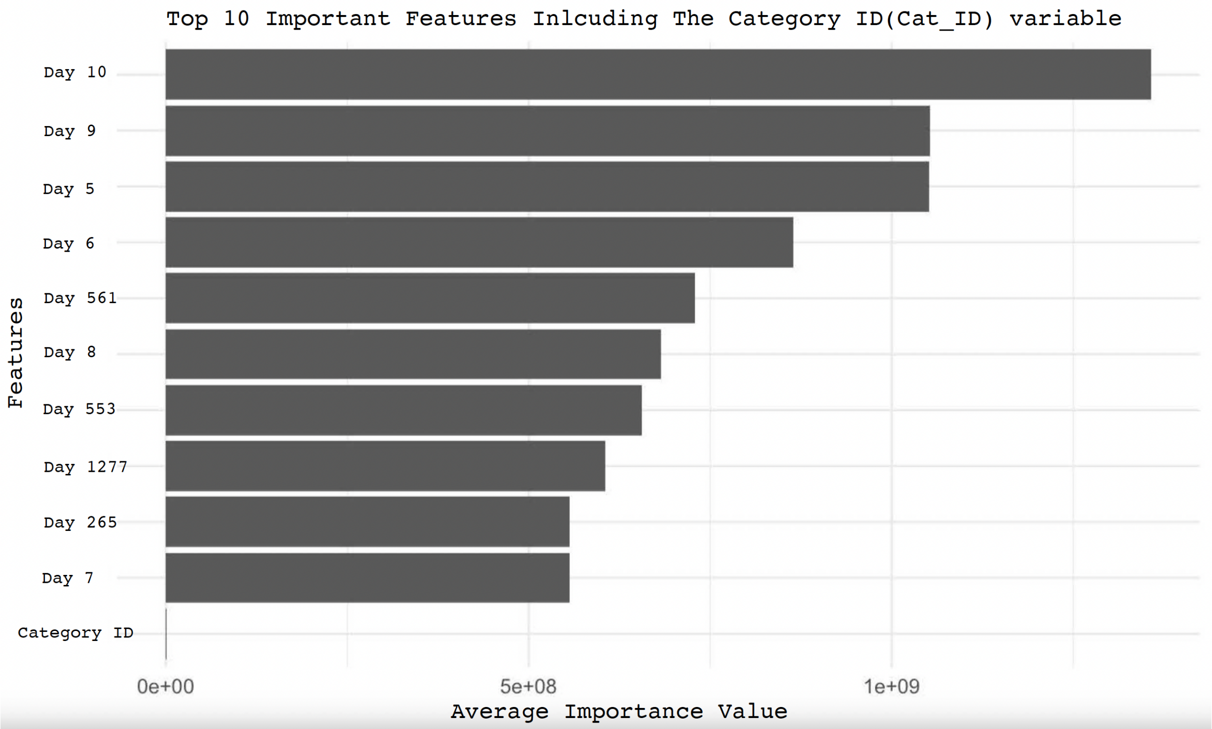

Figure 9: Feature Importance Plot for the 10 most important figures including Category ID

From the Variable Importance Plots in Figure 8 & 9 we can deduce that there is a consistent day specific importance, `cat_id` has lower significance compared to the specific days and that there might be implications as well as strategy formulation.

Consistent Day-Specific Importance: From the Variable Importance Plot in Figure 8 we can see that there are specific days which have high importance in predicting the target variable. After including the variable, `cat_id`, in the feature importance analysis as can be seen in Figure 9, the day-specific variables (`d_10`, `d_9`, `d_5`, `d_6`, `d_561`, `d_8`, `d_553`, `d_1277`, `d_265`, and `d_7`) retain their high influence positions, suggesting that certain days are crucial to the model's predictive ability. The significant low value of importance for the variable `cat_id` suggests that is has very low predictive power in predicting the target variable and suggests that product category (as expected beforehand) is not of importance compared to the specific days shown in the plot in Figure 8. In Figure 8 we can see that `d_561` has high importance in predicting the target variable, which is interesting as we also saw in our Decision Tree graph that `d_561` has a significant position in the presented path. Day 561 might be a special day which can be of high influence in driving consumer sales patterns for a specific product.

Lower Significance of `cat_id`: As mentioned in the previous paragraph on specific days importance from our Random Forest model we find it interesting that `cat_id` emerges as a remarkably less influential variable, indicating that product categorization identity may not be a prominent determinant in deciphering the dynamics of sales sum within the observed dataset. An remarkable result, but not less valuable in providing consumer goods recommendations when considering that decision making does not have to focus deeply on product categories.

Implications and Strategy Formulation: From interpreting both plots, to be more specific the plot in Figure 9 where the importance of `cat_id` is taken into consideration, we can suggests that sales in different product categories are similar, influenced by general factors rather than specific elements per category. This observation supports the idea that general sales strategies and promotional campaigns can be effective across categories. In light of this, a few strategic additions and deeper analytical probes may provide further clarity and refinement in deriving actionable insights:

- **Category-Specific Analysis:** Despite the lower overall importance, examining the impact of `cat_id` during certain periods or on crucial days can reveal category-specific dynamics or correlations with sales trends.
- **Mixed-Effect Modeling:** Exploring mixed-effects models can facilitate the capture of overarching trends while accounting for variations within `cat_id`.
- **Category and Day Alignment:** The strategic alignment of influencer days with product category peaks can provide opportunities for optimizing marketing initiatives, inventory management and customer engagement strategies.
- **Inter-Category Correlations:** Examining correlations in sales patterns between different product categories during the identified critical days can provide interesting insights into buying behavior or the dynamics of substitute and complementary products.

When taking into consideration the previous presented results and interpretation of our Random Forest model and also from our Decision Tree model, we can see that the decrease in the significance of `cat_id` demonstrates that sales incentives are primarily driven by key days, recognizing specific characteristics by category, especially during these days, can lead to a specific (marketing) strategy. This strategy combines key predictive factors with detailed insights specific to each category. While variable importance plots are a valuable tool in predictive modeling, they present certain limitations that warrant caution. One major disadvantage is their lack of insight into underlying scientific truths and causal relationships. This limitation, rooted in the shift from modeling surface or noise mechanisms to focusing on prediction (Efron, 2020), implies that these plots might offer only a superficial understanding of the data, potentially overlooking deeper, more stable relationships. Additionally, Efron (2020) states with his research, the reliance on ephemeral relationships, as discuss by the emphasis on algorithmic behavior over data generation models, raises concerns about the long-term relevance of the identified important variables. This issue is further compounded in scenarios involving large datasets (high 'n' and 'p'), where the complex behavior of algorithms may not be fully captured by variable importance plots, potentially leading to oversimplified interpretations (Efron, 2020). Furthermore, this approach's potential misalignment with traditional scientific inquiry, which values a comprehensive understanding of the data's origin and nature, underscores a significant gap in using these plots for certain types of scientific research. In summary, while variable importance plots provide valuable insights into the predictive algorithms, their limitations in conveying scientific truths, handling complex data scenarios, and aligning with traditional scientific methods marks the need for a careful and balanced application in predictive modeling.

5.2.2 Comparing the Metrics

With our research we aimed to quantify the trade-off between accuracy and interpretability for Decision Trees and Random Forest models in optimizing consumer goods recommendations. Taken into account the different specific parameters and features used in each method. With the application of both a Decision Tree model and a Random Forest model on our data set we quantified the accuracy trade-off based on specific model performance metrics and we interpreted both model predictions using innovative methods. In this part of the results section we will compare the model performance metrics of both models and compare the interpretation of both plots.

Decision Tree Model:

- MAE: Ranges from approximately 1011.8 to 1184.3 across subgroups.
- RMSE: Ranges from about 1798.6 to 4109.3.
- R^2: Ranges from approximately 0.52 to 0.85.

Random Forest Model:

- MAE: Ranges from approximately 125.3 to 154 across subgroups.
- RMSE: Ranges from about 419.4 to 1430.3.
- R^2: Ranges from approximately 0.98 to 0.99.

Firstly, we present a trade-off analysis between the two models:
Predictive Accuracy: The Random Forest model generally exhibits superior predictive accuracy, as indicated by lower MAE and RMSE values across all subgroups. This underscores Decision Trees capacity to make more precise predictions on average.
Model Complexity: Random Forests, being ensembles of Decision Trees, are inherently more complex and less interpretable than individual Decision Trees, which might be preferable where interpretability is crucial.
Explanation of Variance: Our Random Forest model show us a higher proportion of variance (higher R^2) compared to our Decision Tree model, indicating a more profound grasp of the underlying patterns in the data.
Consistency: Decision Trees may exhibit notable variability in performance between subgroups, while Random Forests tend to provide more consistent and stable results.

5.2.3 Quantifying the Trade-off

To reach for the main goal of this research which is quantifying the trade-off between accuracy and interpretability of the Random Forest and Decision Tree models we used we need to compare the metrics and take into consideration the interpretations we were able to extract from the visualizations. Comparing the performance metrics of both models and summarizing the differences of both models are shown in Table 3. We present a quantitative comparison between the Decision Tree and Random Forest models based on the differences and the percentage differences for each subgroup. First, we provide comparison's of the metrics stand alone for Decision Tree and Random Forest.

Mean Absolute Error (MAE) Comparison:

- Random Forest MAE Range: Approximately 125.3 to 154.
- Decision Tree MAE Range: Approximately 1011.8 to 1184.3.
- Quantitative Difference: The Random Forest model demonstrates significantly lower MAE values, indicating markedly superior predictive accuracy.

Root Mean Squared Error (RMSE) Comparison:

- Random Forest RMSE Range: Approximately 419.4 to 1430.2.
- Decision Tree RMSE Range: Approximately 1798.6 to 4109.2.
- Quantitative Difference: Random Forests boast significantly lower RMSE values, signaling predictions with substantially smaller errors.

Coefficient of Determination (R2):

- Random Forest R2 Range: Approximately 0.9792 to 0.9976.
- Decision Tree R2 Range: Approximately 0.5177 to 0.8548.
- Quantitative Difference: Random Forest consistently secures higher R2 values and variance explained, delineating a superior explanation of the variance across all subgroups.

From these first comparisons of both model performances on their own we find that the comparative analysis confirms the superiority of the Random Forest model in terms of predictive accuracy and explained variance. The differences in MAE, RMSE and R^2 are significant, with the Random Forest model consistently providing more accurate and reliable predictions.

To provide a more quantitative comparison of the Decision Tree and Random Forest models based on the output values, we will present and discuss each metric in the context of its implications for model performance in Table 3.

The MAE and RMSE differences we calculated by subtracting the RF results from the DT results, which show the absolute improvement in error metrics by using RF over DT. A positive difference indicates an improvement, and the percentages reflect how much the RF model has improved relative to the DT model in terms of percentage. The negative differences in R^2 show that RF has higher values than DT, which are closer to 1, indicating better model fit and more variance explained by the RF model. The percentage differences for R^2 are significant, often exceeding 70%, which indicates the superior performance of the RF model compared to the DT model across all subgroups. The RF model consistently shows higher R^2 values, suggesting that it is a better model for prediction in this context. Concluding with the findings that the Random Forest model significantly outperforms the Decision Tree model, with improvements in

Table 3: Comparison of Decision Tree and Random Forest Models

Sub-group	MAE (DT)	RMSE (DT)	$\mathbf{R^2}$ (DT)	MAE (RF)	RMSE (RF)	$\mathbf{R^2}$ (RF)
1	1038.3	3574.4	-0.47	88.45%	86.98%	-90.38%
2	1020.5	1893.8	-0.20	88.38%	75.09%	-25.32%
3	865.5	646.6	-0.15	84.89%	31.85%	-18.07%
4	956.7	1654.9	-0.18	87.27%	72.29%	-22.22%
5	863.1	649.0	-0.17	85.30%	31.21%	-20.99%
6	957.9	1589.9	-0.22	87.35%	68.63%	-28.57%
7	903.2	1379.1	-0.14	87.82%	76.68%	-16.47%
8	1057.0	2346.4	-0.22	89.25%	82.31%	-28.57%
9	914.6	1448.2	-0.17	87.14%	70.43%	-20.73%
10	1035.9	2125.0	-0.19	88.43%	76.40%	-23.75%

prediction error metrics ranging from 74% to 90%. Also concluding from the table, we find that the $\mathrm{R}^{\mathbf{2}}$ performances are on average approximately 25.39% better than those of the Decision Tree model. This significant improvement marks the Random Forest model's ability to explain the variance in the data and its enhanced predictive accuracy.

However, it is important to remember that the trade-off is not only about performance, but also about model complexity and interpretability, with Decision Trees providing a simpler and more interpretable model alternative. The choice between the two models should be determined by specific objectives and the trade-offs the user is willing to make. It depends on which decision needs to be taken and based on that information or requested insight, for which kind of interpretation is desired. We are not able to quantify the interpretability of both Decision Tree and Random Forest methods withing the scope of this research. However, we are able to gain a valuable insight in which method might be more relevant for a specific issue.

5.3 Discussion

5.3.1 Predictive Performance

Looking at the results section of our study, we see that both models do not offer relatively easy-to-interpret results. Both models give us a very accurate prediction of the target variable and are therefore both of great value for making, in the case of this study, outlet predictions. Both our Decision Tree model and Random Forest model show great predictive power in calculating the target variable. The visualisations we have presented in this paper show that our Decision Tree model shows an actual path in which different choices can be substantiated. Based on the direction the relevant decision-maker wants to take in his decision, the Decision Tree provides a visual justification therein and is thus in this case easier to interpret compared to the Random Forest visualisation where innovative methods are needed to make the results insightful. Our answers to both sub-questions 1 and 2 are therefore be yes.

1. Will Decision Trees be more interpretable than Random Forests but may they sacrifice accuracy in order to achieve this interpretability? When performing a research like the one we executed in this paper Decision Trees will sacrifice accuracy in order to be more interpretable.

With not using ensemble method consisting of multiple trees or, for example, cross validation, Decision Trees will always retain a disadvantage over Random Forest that do have this built in on the accuracy topic. But they will most likely be the ones winning on interpretability.

2. Will Random Forests be more accurate than Decision Trees but may they sacrifice interpretability in order to achieve this accuracy? The same discussion holds for Random Forest.

From our results we find that our Random Forest model performs better compared to our Decision Tree model. But we have to conclude that our model does sacrifice interpretability in order to gain accuracy.

5.3.2 Insightful Interpretability

Based on the interpretability of the visualisations we have obtained and presented in this research from our Decision Tree and Random Forest models, we need to choose which one can provide the most valuable insight to the FMCG industry. If we look at specific questions that decision-makers need to answer, it is a question that may be partly easier to answer. While carrying out this research, we found out that the interpretability of the models is of great importance, but also a piece of customisation for the specific issue. As we have focused on the FMCG industry, consumer behaviour and therefore consumer preference should be high on our agenda. The valuable insights we want to obtain must therefore focus on processes. Decision Trees offer us an overview of the path that can be chosen to arrive at a final decision, indicating in it which intermediate steps are chosen and thereby showing which variables are important in predicting the target variable. In effect, they provide us with a consumer path.

3. For the FMCG industry, are Decision Trees be more useful than Random Forests in terms of providing actionable insights into the factors driving consumer behavior? As explained in the above paragraph we can conclude based on the results of this research that yes, specifically for the FMCG industry we find that the interpretation we gain from the Decision Tree are more insightful than the interpretation gained from our Random Forest Variable Importance Plot. The biggest reason for this being that the Decision Tree shows us a model path for consumer behaviour.

4. In order to optimize the accuracy of consumer goods recommendations for the FMCG industry, is a hybrid approach that combines Decision Trees and Random Forests most effective? For other research this recommendation might be of value, but the results of our models differ to much from each other to combine them. This is why the answer to this sub-question is no.

6 Concluding thoughts and Future Work

6.1 Conclusion

To conclude our research and to translate the results we found from our analysis into valuable insights for the FMCG industry, we will present the conclusion of our research in this section. Through answering the research question we formulated, which we used as a guide while conducting our research, we will substantiate the concluding thoughts building on our research findings and regale you with valuable insights that can be applied by decision-makers in the FMCG industry.

When we look at the trade-off between accuracy and interpretability while employing Decision Trees and Random Forests for optimizing consumer goods recommendations utilizing the M5 Walmart Sales Forecasting data, we see some significant differences. Random Forest performs better than Decision Trees in predicting correctly, which we can tell from looking at the Mean Absolute Error (MAE), Root Mean Square Error (RMSE), R^2, and how much variance they explain. Days like Day 10 (`d_10`) and Day 9 (`d_9`) really matter for predictions according to the outcome of our Random Forest model. But the type of product, called Category ID (`cat_id`), doesn't seem as important according to the Variable Importance(Figures 8 & 9). Random Forest is more accurate but harder to understand, while Decision Tree is easier and even maybe more insightful to interpret, but not as sharp in making predictions. The outcome of our research is important for people who use these models but aren't experts in the field of machine learning applications and the interpretation of machine learning models. Choosing whether to make use of a Decision Tree model or a Random Forest model depends on what you need, who's going to use it, and what for.

With the above, we have answered the first part of our research question, but this does not yet answer the second part which relates to giving consumer goods recommendations and optimising them through the use of machine learning methods such as Decision Tree and Random Forest. Since the results of our research enable us to confirm parts of the existing literature but also contribute to research on the application of machine learning models in the FMCG industry, it has been of great academic interest and relevance. With our research, besides academic relevance, we have also had a focus on optimising consumer goods recommendation and we discuss how relatively "complex" machine learning models can be applied for gaining insights on a short-term or daily basis for decision makers, for example. As mentioned above, choosing whether to make use of a Decision Tree model or a Random Forest model depends on what you need, who's going to use it, and what you are going to use it for. Many factors make it important to choose a particular model, as just mentioned, many decision makers need short-term insights or like to receive a daily update with advice on which approach is best. As a result, not only accuracy of the model and prediction or interpretability must be considered, but also, for example, the calculation speed of the model, the capacity of the model and more. This is why a model like Decision Trees might be preferable, our results show that the prediction accuracy does not under perform our Random Forest model that badly and the results of our model are easy to interpret. What is beyond the scope of our study, but what can be named is that our Decision Tree model required significantly less computation time than our Random Forest model and the model can be considered more advantageous in that respect if frequent (daily) insights are desired.

To conclude the concluding section of this paper, we would like to elaborate on the results of our models and outline what we can glean from them. From both the Decision Tree shown in 7 and the Variable Importance Plot shown in 8, we can see that specific days play an important role in determining the value of the target variable. Thus, a number of days have great explanatory power in determining the target variable and thus can be marked as days of interest if we want to focus on consumer behaviour pattern/preference. To be specific, our results show that we can consider the following days Day 561 and 1,260 (d_{561} and d_{1260}) as relevant and can be presented to decision makers. A possible advice for decision makers is to take a closer look at these days and examine whether they hold a special date, holiday or other function that would allow them to be marked as special and explain their predictive power. For example, when these days are holidays, they can create special holiday campaigns and thereby underline themselves or certain products even better. Or when they are weekend days, they can offer special deals to their customers and thereby provide an extra service for their customers.

6.2 Implications and Future Work

6.2.1 Practical and Scholarly Implications

Our study looks at how Decision Trees and Random Forests balance being right versus being easy to understand, which matters for business people and researchers. Businesses, especially those selling quickly bought goods, need to think carefully about what they want from a model. If they need to explain decisions easily, Decision Trees might be better. If they need the best predictions, Random Forest could be the way to go. A major implication coming from the usage of these methods for this research therefore is that no optimal combination can be found for high accuracy and easy and insightful interpretation. For researchers, the study points out an interesting area to look into: making a model that is both easy to explain and good at predicting, a possibility for this need to be achieved is by for example combining Decision Tree interpretation and Random Forest accuracy. However, this is beyond the scope of our research and therefore an interesting are to focus on in future research. Combining both methods might provide researchers and businessmen some very relevant insights.

6.2.2 Model Applications

Our research shows that whether you use Decision Trees or Random Forests the decision depends on what you need them for. It is not only important to make a decision based on the preferred output, but also it needs to be taken into consideration what you data looks like. For example, the complexity of a data set is crucial for how the model is able to process it. Decision Tree are at their best with categorical data and is for example less good with continuous variables and imbalanced data sets. However, for clear, direct advice Decision Trees are the model you should deploy. While Random Forest is better for getting the most accurate predictions they can let us down in gaining valuable insights and remain mysterious about their way or working and computing. Overall we can say that Random Forest handle most of the data sets better than Decision Trees can. In our research we encountered as well the problem that the performance of Decision Trees is inferior which may be due to different reasons. As mentioned in the results part is their under performance to blame because of overfitting, not handling outliers to well or the fact that our dataset might be to complex. However, in our research we did not have the capacity to dedicate additional research in order to figure this out. Which leaves an interesting opportunity for future research.

6.2.3 Strategic Recommendations and Future-proofing Analytics

This study suggests that knowing which factors, like specific days, are important and understanding what each model does best should help people make better use of Decision Trees and Random Forests models to deploy them better. Use Random Forest when you really need to get things right, and Decision Tree when you need clear, useful advice. Looking into mixed models that combine Decision Tree and Random Forest would have been a good idea for both current use and is for sure an relevant topic for future research. It's important to keep these models and methods flexible and ready to change with new market trends, technology, and customer habits to make sure they stay useful in the long run.

6.2.4 Robustness of Conclusions

An integrated approach might have been a wise move to test the robustness of these results: engaging in cross-validation, which involves utilizing various folds and disparate data segments to confirm the persistence of patterns in predictive accuracy and variable importance; initiating a sensitivity analysis. Something we only performed partially with our analysis. Also, thereby adjusting parameters and model assumptions to scrutinize whether the predictive power and variable importance hold steady amidst the modifications; and, if feasible, conducting external validation by applying the models on an alternate dataset to affirm their predictive capabilities and consistency across differing contexts. Which is something we did not do within the scope of our research. We choose not to add or deploy any of these robustness metrics in order to not make our research too extensive. We wanted to focus on deploying our basic Decision Tree and Random Forest model and the results we got from their analysis and compare their performance. We wanted to create an umbrella overview of the function of them and we were not focusing on optimising model performance or results. Of course, we do support the drive for future research and admit that our research results might have been better if we did included them.

That is why we can conclude that this study's findings about Decision Trees and Random Forests are strong, but they need a close look, especially because of differences in key measures like Mean Absolute Error (MAE), Root Mean Square Error (RMSE), R^2, and how much they explain. To make sure these results are solid, a few steps are key:

1. Cross-validation: This means testing the models with different parts of the data to see if they still predict well and if the important factors stay the same.

2. Sensitivity analysis: Change some settings and assumptions in the models to check if they still work well and if the same things are important.
3. If possible, try the models on a different set of data to see if they still predict well in other situations.

This paper's conclusions about the balance between being easy to understand and being accurate in Decision Trees and Random Forests needs careful checking. This includes making sure the findings can be used in different situations and industries, not just where they were first tested. It's also important to compare these results with what other studies have found to see if they match or add something new. Finally, it's crucial to test how useful these findings are in real-world situations, especially in businesses like those selling quickly bought goods, to make sure the advice given is still relevant and effective.

6.3 Further Research

Looking into the future, there are many ways to build on this papers findings. One idea is to use different ways of measuring how well models work to get a fuller picture of their strengths. Exploring various combined models might help find the best mix of being easy to understand and accurate, which could improve systems for suggesting products in businesses like those selling quickly bought goods. Also, studying how to choose and combine different factors (feature engineering and selection) might reveal new or better combinations for predictions. It would be interesting to compare other machine learning methods, like Gradient Boosting or Neural Networks, to see how they stack up. And testing these models in different industries, places, or with different products would help understand how well they work in various situations and what specific patterns or needs different industries have. This future research could not only confirm what our research found but also add to the ongoing conversation about using machine learning in business recommendation systems.

References

Adebanjo, D., & Mann, R. (2000). Identifying problems in forecasting consumer demand in the fast moving consumer goods sector. *Benchmarking: An international journal*, 7(3), 223-230.

Amit Y, Geman D (1997). Shape quantization and recognition with randomized trees. *Neural Comput*, 9:1545–1588.

Anyanwu, M. N., & Shiva, S. G. (2009). Comparative analysis of serial decision tree classification algorithms. *International Journal of Computer Science and Security*, 3(3), 230-240.

Belgiu, M., & Drăguţ, L. (2016). Random forest in remote sensing: A review of applications and future directions. *ISPRS journal of photogrammetry and remote sensing*, 114, 24-31.

Biau, G., & Scornet, E. (2016). A random forest guided tour. *Test*, 25, 197-227.

Boehmke, B., & Greenwell, B. (2019). *Hands-on machine learning with R.* Chapman and Hall/CRC.

Breiman, L. (2001). Random forests. *Machine learning*, 45, 5-32.

Chan, J. C.-W., & Paelinckx, D. (2008). Evaluation of random forest and Adaboost tree-based ensemble classification and spectral band selection for ecotope mapping using airborne hyperspectral imagery. *Remote Sensing of Environment*, 112, 2999-3011.

Craven, M., and Shavlik, J. W. (1996). Extracting tree-structured representations of trained networks. In *NIPS*.

Das, P. (2009). Adaptation of fuzzy reasoning and rule generation for customers' choice in retail FMCG business. *Journal of Management Research*, 9(1), 15.

Dietterich TG (2000). Ensemble methods in machine learning. In: Kittler J, Roli F (eds) *Multiple classifier systems.* Springer, Berlin, pp 1–15.

Efron, B. (2020). Prediction, estimation, and attribution. In *International Statistical Review, 88*, S28-S59.

Feraud, R., Clerot, F. (2002). A Methodology to Explain Neural Network Classification. In *Neural Networks, 15(2)*, 237-246.

Francis, M. (2006). Stage model research in the UK fast moving consumer goods industry. *International Journal of Logistics: Research and Applications*, 9(4), 351-368.

Gislason, P.O., Benediktsson, J.A., Sveinsson, J.R. (2006). Random forests for land cover classification. *Pattern Recogn. Lett.*, 27, 294–300.

Gilpin, L. H., Bau, D., Yuan, B. Z., Bajwa, A., Specter, M., & Kagal, L. (2018). Explaining explanations: An overview of interpretability of machine learning. In *2018 IEEE 5th International Conference on data science and advanced analytics (DSAA)* (pp. 80-89). IEEE.

Gregorutti, B., Michel, B., & Saint-Pierre, P. (2017). Correlation and variable importance in random forests. In *Statistics and Computing, 27, 659-678..*

Ham, J., Chen, Y., Crawford, M. M., & Ghosh, J. (2005). Investigation of the random forest framework for classification of hyperspectral data. *IEEE Transactions on Geoscience and Remote Sensing*, 43(3), 492-501.

Han, H., Lee, S., Im, J., Kim, M., Lee, M. I., Ahn, M. H., & Chung, S. R. (2015). Detection of convective initiation using Meteorological Imager onboard Communication, Ocean, and Meteorological Satellite based on machine learning approaches. *Remote Sensing*, 7(7), 9184-9204.

Hatwell, J., Gaber, M. M., & Azad, R. M. A. (2020). CHIRPS: Explaining random forest classification. In *Artificial Intelligence Review, 53, 5747-5788.).*

Ho, T. (1998). The random subspace method for constructing decision forests. *Pattern Analysis and Machine Intelligence*, 20, 832-844.

Hodson, T. O. (2022). Root-mean-square error (RMSE) or mean absolute error (MAE): When to use them or not. In *Geoscientific Model Development, 15(14), 5481-5487..*

Izza, Y., Ignatiev, A., & Marques-Silva, J. (2020). On explaining decision trees. In *arXiv preprint arXiv:2010.11034..*

Jadhav S.D., Channe H.P. (2016). Efficient recommendation system using decision tree classifier and collaborative filtering. *Int. Res. J. Eng. Technol.*, 3:2113-8.

Khade, A. A. (2016). Performing customer behavior analysis using big data analytics. *Procedia computer science*, 79, 986-992.

Kim, J. K., Song, H. S., Kim, T. S., & Kim, H. K. (2005). Detecting the change of customer behavior based on decision tree analysis. *Expert Systems*, 22(4), 193-205.

Kotsiantis, S. B., Zaharakis, I., & Pintelas, P. (2007). Supervised machine learning: A review of classification techniques. *Emerging artificial intelligence applications in computer engineering*, 160(1), 3-24.

Maimon, O. Z., & Rokach, L. (2014). *Data mining with decision trees: theory and applications (Vol. 81).* World scientific.

Mathur, A. (2020) Comprehensive Data Visualization - M5 EDA `https://www.kaggle.com/code/akashmathur2212/comprehensive-data-visualization-m5-eda#2.1-Explanatory-Variables:-Sales-Train-Validation`

Nakagawa, S., Schielzeth, H. (2013). A general and simple method for obtaining R2 from generalized linear mixed-effects models. In *Methods in ecology and evolution, 4(2), 133-142..*

Nozari, H., Szmelter-Jarosz, A., & Ghahremani-Nahr, J. (2022). Analysis of the challenges of artificial intelligence of things (AIoT) for the smart supply chain (case study: FMCG industries). *Sensors*, 22(8), 2931.

Panjwani, M., Ramrakhiani, R., Jumnani, H., Zanwar, K., & Hande, R. (2020). Sales Prediction System Using Machine Learning. *Sales Prediction System Using Machine Learning.*

Patel, H. H., & Prajapati, P. (2018). Study and analysis of decision tree based classification algorithms. *International Journal of Computer Sciences and Engineering*, 6(10), 74-78.

Rodriguez-Galiano, V. F., Ghimire, B., Rogan, J., Chica-Olmo, M., & Rigol-Sanchez, J. P. (2012). An assessment of the effectiveness of a random forest classifier for land-cover classification. *ISPRS Journal of Photogrammetry and Remote Sensing*, 67, 93-104.

Safavian, S. R., & Landgrebe, D. (1991). A survey of decision tree classifier methodology. *IEEE transactions on systems, man, and cybernetics*, 21(3), 660-674.

Sarkar, S., Weyde, T., Garcez, A. D., Slabaugh, G. G., Dragicevic, S., & Percy, C. (2016). Accuracy and interpretability trade-offs in machine learning applied to safer gambling. In *CEUR Workshop Proceedings (Vol. 1773).*

Shang, X., & Chisholm, L. A. (2014). Classification of Australian native forest species using hyperspectral remote sensing and machine-learning classification algorithms. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 7, 2481-2489.

Song, Y. Y., & Ying, L. U. (2015). Decision tree methods: applications for classification and prediction. *Shanghai archives of psychiatry*, 27(2), 130.

Sperandei, S. (2014). Understanding logistic regression analysis. *Biochem Med (Zagreb)*, 24(1), 12-18.

Strobl, C., Boulesteix, A. L., Kneib, T., Augustin, T., & Zeileis, A. (2008). Conditional variable importance for random forests. In *BMC Bioinformatics, 9*, 1-11.

Tarallo, E., Akabane, G. K., Shimabukuro, C. I., Mello, J., & Amancio, D. (2019). Machine learning in predicting demand for fast-moving consumer goods: An exploratory research. *IFAC-PapersOnLine*, 52(13), 737-742.

Tonbul, T. (2019). Sales Forecast in FMCG Sector with Artificial Neural Networks.

Tsoumakas, G. (2018). A survey of machine learning techniques for food sales prediction. *Artificial Intelligence Review – Springer, Netherlands.*

Valbuena Godoy, J. N. (2022). Demand forecasting of Fast Moving Consumer Goods based on modeling of time series and deep learning methods.

Valente, F., Henriques, J., Paredes, S., Rocha, T., de Carvalho, P., & Morais, J. (2021). Improving the compromise between accuracy, interpretability and personalization of rule-based machine learning in medical problems. In *2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, pp. 2132-2135, IEEE.

Vetrivel, A., Gerke, M., Kerle, N., & Vosselman, G. (2015). Identification of damage in buildings based on gaps in 3D point clouds from very high resolution oblique airborne images. *ISPRS Journal of Photogrammetry and Remote Sensing*, 105, 61-78.

Wu J., Zheng S. (2015). Forecasting for fast fashion products based on web search data by using OS-ELM algorithm. *Journal of Computational Information Systems*, 11(14), 5171-5180.

Wu, M., Hughes, M., Parbhoo, S., Zazzi, M., Roth, V., & Doshi-Velez, F. (2018, April). Beyond sparsity: Tree regularization of deep models for interpretability. In *Proceedings of the AAAI conference on artificial intelligence (Vol. 32, No. 1).*

A Appendix

This section gives an overview of the standard machine learning methods used in this paper.

A.1 Decision Tree

Decision Trees, cited by Song and Ying (2015) are powerful statistical tools, they aim to simplify complex input-target relationships and facilitate easy interpretation without the need for distributional assumptions, among other advantages. They can deal well with skewed data and robustly handle outliers (Song & Ying, 2015). The model's genesis can be attributed to the Recursive Partitioning Algorithm, where the feature space is repeatedly bifurcated based on minimizing a defined criterion. Mathematically, the decision to split a node is determined by identifying the feature, s, and the threshold, t, which collectively minimize the impurity of resultant child nodes:

$$(s,t) = \arg\min_{s,t} \left[\sum_{i=1}^{2} p_i H(R_i) \right]$$

Here, p_i symbolizes the proportion of samples in child node i, H represents the impurity function, and R_i corresponds to the regions demarcated by the split on feature s and threshold t.

In DT models, the impurity function, H, is pivotal in gauging the uniformity of a node. For classification tasks, the Gini impurity is frequently employed, articulated as:

$$H_{\text{Gini}}(t) = 1 - \sum_{i=1}^{c} p_i^2$$

While for regression tasks, the Mean Squared Error (MSE) typically acts as the impurity measure, defined as:

$$H_{\text{MSE}}(t) = \frac{1}{N_t} \sum_{i \in D_t} (y_i - \bar{y}_t)^2$$

Where N_t is the number of samples in node t, D_t signifies the training samples in node t, y_i is the target for sample i, and $\bar{y}_t$ is the average target over all samples in node t.

Post-construction, the tree is subjected to pruning, a crucial process aimed at mitigating overfitting by algorithmically trimming branches that confer minimal predictive power. Cost-complexity pruning, a popular variant, is defined by the cost-complexity criterion $R_\alpha(T)$:

$$R_\alpha(T) = R(T) + \alpha |T|$$

Where $R(T)$ measures the misclassification rate of tree T, $|T|$ represents the number of terminal nodes, and α serves as a complexity parameter, regulating the trade-off between tree size and its goodness of fit to the data. The optimal tree size is generally determined via cross-validation.

A.2 Random Forest

Random Forest, proposed by Leo Breiman in 2001, operates by constructing a multitude of Decision Trees at training time and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees. Mathematical details and algorithmic specifics of Random Forest are embedded throughout this section, with the aim of increasing understanding and providing a clear picture of the operational mechanism.

A.2.1 Algorithmic Framework of Random Forest

The Random Forest algorithm is hinged upon the bootstrap aggregating (or bagging) technique, which involves generating multiple bootstrap samples (subsets of data) and then aggregating the results. Given a training set $X = x_1, x_2, \ldots, x_n$ with responses $Y = y_1, y_2, \ldots, y_n$, bagging repeatedly ($B$ times) selects a random sample with replacement of the dataset and fits trees to these samples:

$$(T_b, b = 1, \ldots, B)$$

For $b = 1, \ldots, B$:

- Sample, with replacement, n training examples from X, Y; call these X_b, Y_b.
- Train a classification or regression tree T_b on X_b, Y_b.

After training, predictions for unseen samples x' can be made by averaging the predictions from all the individual regression trees on x' or by taking a majority vote in the case of classification trees.

$$\hat{f}_{rf}^{B}(x') = \frac{1}{B} \sum_{b=1}^{B} \hat{f}_b(x')$$

where $\hat{f}_{rf}^{B}(x')$ is the predicted response for sample x' and $\hat{f}_b(x')$ is the predicted response from the bth Random Forest tree.

A.3 Evaluation metrics

To evaluate the performance of our Decision Tree model and our Random Forest model, we employed several statistical metrics, each providing unique insights into the model's predictive accuracy and reliability. We computed the Mean Absolute Error (MAE), the The Root Mean Squared Error (RMSE), The Coefficient of Determination, denoted as R^2. The Mean Absolute Error (MAE) measures the average magnitude of the errors in a set of predictions, without considering their direction. It is calculated as the average of the absolute differences between the predicted values and the actual values (Hodson, 2022). The formula for MAE is:

$$\text{MAE} = \frac{1}{n} \sum_{i=1}^{n} |y_i - \hat{y}_i| \tag{4}$$

where y_i represents the actual values, $\hat{y}_i$ denotes the predicted values, and n is the number of observations.

The Root Mean Squared Error (RMSE) provides a measure of the average magnitude of the error, giving a higher weight to larger errors. This is particularly useful when large errors are particularly undesirable. The RMSE is defined as the square root of the average of squared differences between the predicted and actual values. The formula for RMSE is:

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{5}$$

The Coefficient of Determination, denoted as R^2, quantifies the proportion of the variance in the dependent variable that is predictable from the independent variable(s). It provides a measure of how well observed outcomes are replicated by the model (Nakagawa & Schielzeth, 2013). The formula for R^2 is:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{6}$$

where $\bar{y}$ is the mean of the actual values.

Lastly, the Variance Explained complements the R^2 metric by providing a direct interpretation of the proportion of total variation in the dependent variable that is accounted for by the model. It is essentially another way to express R^2, and in many contexts, these terms are used interchangeably.

By applying these metrics, we aim to provide a comprehensive assessment of our Decision Tree model's performance, focusing on its accuracy, reliability, and the extent to which it captures the variance in the target variable.