Tab 1

**MSBC 5490**

**BUAN Experiential Project**

**Rudi’s Bakery**

**Final Report**

**Submitted by Group 4:**

| **Name** | **Email Address** |
| --- | --- |
| Rafael Cintron | rafael.cintron@colorado.edu |
| Supria Deka | supria.deka@colorado.edu |
| Pranathi Manthri | pranathi.manthri@colorado.edu |
| Murali Prateek Manthri | murali.manthri@colorado.edu |
| Jinal Mehta | jinal.mehta@colorado.edu |

**Index**

[**Executive Summary 3**](#_2umk3xjw27h6)

[Summary of Key Insights from Workable Solutions 3](#_g0ep794539k7)

[**Strategies and Recommendations 4**](#_jmjr5qtkfsgi)

[Introduction 5](#_xsxaidspzf53)

[Problem Statement 5](#_hnrywtpsc8ph)

[Overarching Objectives 5](#_n7s6q367f986)

[**Data 6**](#_pfr1mob0wg4)

[Data Introduction 6](#_lqnk7xh6jvcq)

[Data Cleaning Steps 6](#_g7uckcaszwss)

[**Key Performance Indicators (KPIs) 6**](#_z4jbvinmd5g8)

[**Model/Workable Solution 8**](#_z8tu1njwmzeq)

[Models Used 8](#_dcona1lpgvbh)

[1. Random Forest 8](#_flbbicoeutlr)

[2. Prophet 8](#_8c86nrezlwgw)

[3. Exponential Smoothing 8](#_r8q280fpb152)

[4. Decision Tree 8](#_26454e6f7qj3)

[5. Extra Trees 9](#_usjlwo4colo3)

[6. Clustering Analysis 9](#_ad9cyk1360qa)

[Model Comparison Summary 10](#_ua5fqdyb372v)

[**Workable Solutions Used/Recommended 10**](#_homp7rh78nfq)

[Random Forest 10](#_3nol0ftmjkub)

[Key Observations from the Feature Importance Chart 12](#_ubo0u5utd04t)

[Business Implications for Rudi’s 12](#_59lss6v71avp)

[Extra Trees Model 14](#_ornogtovultz)

[Performance Metrics Breakdown 15](#_ax2lc033hu5l)

[Business Implications for Rudi’s 16](#_jxiz7tgb5e5p)

[**Final Suggestion 17**](#_vmzqiylp6k2l)

[**Future Directions 19**](#_spp15srr0dsm)

[**Adapting this information to an Interactive Dashboard 22**](#_uxr375zfnu8f)

[A. Retailer Summary Statistics 22](#_tv1kf8ywjrus)

[B. Store Summary Statistics 23](#_bz96lnsq07ra)

[C. Products Summary Statistics 24](#_1iq8uu46ie39)

[D. Route Summary Statistics 25](#_89gx9m5jl5mu)

[**Conclusion 26**](#_7s8n5b5fb2th)

[Appendix 27](#_u5wwslw4eejr)

[Generative AI prompts 27](#_2g2y1r6ac48v)

# **Executive Summary**

We performed an in-depth analysis of historical sales data, utilizing forecasting methods and anomaly detection to improve strategic decision-making. We then created an interactive dashboard for visualizing the results on a platform that can be accessible and digestible for all those within Rudi’s.

As we focused on the following to came up with business solution:

* Deploying Moving Average Forecasting & ETS models to project sales trends.
* Applying anomaly detection techniques to identify stores with abnormal product return rates.
* Designing an interactive Tableau dashboard to provide actionable sales insights.

## **Summary of Key Insights from Workable Solutions**

* **Sales Optimization:** Focusing on the top-performing *Costco* stores for inventory expansion while implementing targeted strategies for underperforming locations would be beneficial. Addressing regional demand and store-specific factors which are impacting sales.
* **Product Return Reduction:** Conducting root cause analysis for high-return SKUs like *Multigrain Oat* and *Honey Sweet* to minimize returns. Adjusting packaging, improving shelf life, and providing customer education to enhance satisfaction.
* **Retailer Collaboration:** Engage with retailers showing high return rates (*e.g., Target, Whole Foods Market*) to optimize stock levels and adjust return policies. Implement exchange incentives to reduce overall return rates.
* **Logistics Efficiency:** Analyze route profitability and return rates to streamline distribution, especially for high-return routes. Focus on inventory management and minimize transportation-related product damage.
* **Improved Forecasting:** Refine Random Forest and Prophet models by incorporating external factors like holidays, promotions, and competitor pricing to better predict seasonal demand fluctuations.
* **Mapping Sales and Returns:** Connected zip code to visualize high and low-performing areas, enabling more effective decision-making for store expansions.
* **Continuous Model Refinement:** Regularly update predictive models with new data to maintain accuracy and relevance.
* **Easily Interpreted Metrics:** Put insights into a single place where the previously mentioned summary and model statistics can be easily visualized.

# **Strategies and Recommendations**

Creating a **Tableau** interactive dashboard makes the summary data far more powerful since it is accessible and intelligible by executives and delivery drivers alike. The platform allows for easy adjustment as products, stores, and routes get added. It can even be paired with more complex python based models.

The goal of the predictive modeling and data analysis aims to optimize sales forecasting, inventory management, and improve overall operational efficiency. Among the various models tested, the **Random Forest** and **Extra Trees** models emerged as the most effective tools, offering the highest accuracy and robust predictive capabilities.

The **Random Forest** model demonstrated superior performance in sales forecasting, with its ability to handle complex, non-linear data and provide valuable feature importance analysis. This makes it particularly well-suited for businesses aiming to refine their sales strategies and optimize inventory distribution. By leveraging Random Forest, companies can gain deeper insights into sales drivers and make data-driven decisions that align closely with market demands and customer preferences. Its high accuracy and resilience against overfitting make it ideal for both short-term and long-term sales forecasting.

The **Extra Trees** model further builds on the strengths of Random Forest by introducing enhanced randomness during tree construction, leading to better generalization and reduced variance. This model excelled in handling class imbalances and complex data patterns, making it an excellent choice for anomaly detection and intricate sales forecasting tasks. Its ability to capture subtle relationships within large datasets allows businesses to uncover hidden patterns, optimize marketing strategies, and improve customer targeting.

## **Introduction**

Rudi’s Bakery is committed to enhancing the efficiency and effectiveness of its Direct Store Delivery (DSD) operations through data-driven decision-making. By analyzing sales trends, identifying potential inefficiencies, and optimizing product distribution, this initiative aims to streamline operations, minimize waste, and improve product availability for customers.

Leveraging advanced analytics and agile methodologies, the project seeks to provide actionable insights that empower teams to make informed decisions, enhance operational efficiency, and maintain the highest standards of service. This effort aligns with Rudi’s Bakery’s mission to deliver fresh, health-conscious, and innovative baked goods while continuously improving distribution strategies. Through collaboration and innovation, the initiative is setting a new benchmark in food production and delivery excellence, ensuring that products reach the right stores at the right time meeting customer demand with precision and reliability.

## **Problem Statement**

Rudi’s Bakery operates a Direct Store Delivery (DSD) network across 14 service routes, ensuring fresh product deliveries to retail locations multiple times a week. While the company tracks basic performance metrics, the growing volume of sales data presents challenges in deriving deeper, actionable insights. The lack of granular visibility into sales patterns, product performance, and distribution inefficiencies makes it difficult to refine operational strategies.

This project aims to conduct an in-depth analysis of DSD data to uncover sales trends across various SKUs, routes, and retail locations. By identifying anomalies such as unexpected product returns or fluctuations in demand, the analysis will help Rudi’s pinpoint inefficiencies and better understand customer purchasing behaviors. Additionally, breaking down sales trends over different time frames spanning the last 4, 12, 26, and 52 weeks will provide a data-driven foundation for improving product placement, optimizing distribution strategies, and enhancing overall sales performance.

## **Overarching Objectives**

The primary objective of this project is to transform raw sales data into actionable intelligence to enhance Rudi’s DSD operations. Key objectives include:

* Analyzing SKU performance across routes, retailers, and stores.
* Identifying inefficiencies through anomaly detection.
* Enhancing demand forecasting using historical data across multiple timeframes.
* Optimizing distribution strategies and improving operational efficiency.
* Providing data-driven recommendations for sales and distribution enhancements.

# **Data**

## **Data Introduction**

The dataset consists of sales data collected from Rudi’s DSD operations, covering multiple routes, retailers, and stores. The data includes product sales, delivery schedules, and return rates. A robust data cleaning methodology was applied to ensure data accuracy and reliability.

## **Data Cleaning Steps**

1. **Removed Duplicate Entries:** Eliminated redundant data to ensure accuracy and consistency in analysis.
2. **Handled Missing Values:** Filled null values with 0 to maintain data integrity and prevent computational errors.
3. **Added Zip Codes to Stores:** Included zip codes for better location-based analysis and geospatial insights.
4. **Standardized Retailer Names:** Ensuring uniformity in retailer names for accurate grouping and analysis.
5. **Cleaned City Names:** Corrected inconsistencies in city names to improve location-based reporting and mapping.
6. **Removed ‘SBT’ from Retailer Names:** Standardized retailer data by removing unnecessary prefixes or suffixes.
7. **Ensured SKU Integrity:** Verified that each SKU has a unique and consistent SKU description to avoid duplication and confusion**.**

In conclusion, the comprehensive data cleaning process significantly enhances the reliability and accuracy of analysis by removing duplicates, handling missing values, standardizing key attributes, and ensuring consistency in SKU descriptions. These improvements facilitate better forecasting, clustering, and return rate evaluations, ultimately leading to more effective decision-making in areas such as sales optimization, inventory management, and retailer performance analysis. By refining the dataset, we ensure that insights derived from it are actionable, precise, and valuable for strategic business planning.

# **Key Performance Indicators (KPIs)**

To effectively measure the impact of data-driven strategies and ensure alignment with Rudi’s Bakery’s business objectives, a set of well-defined Key Performance Indicators (KPIs) has been developed. These KPIs focus on improving sales forecasting accuracy, optimizing inventory management, enhancing operational efficiency, and increasing the overall profitability of the business. Each KPI has been carefully designed to be measurable, actionable, and directly tied to Rudi’s operational goals, ensuring a structured approach to continuous improvement.

**Anomaly Detection Accuracy** serves as a critical metric in monitoring the effectiveness of models designed to detect irregularities in sales trends. This KPI measures the accuracy of anomaly detection models in identifying unusual fluctuations, whether it’s a sudden spike in demand for gluten-free rolls or an unexpected drop in sales of a popular SKU. By reducing false positives and improving detection accuracy, Rudi’s can take proactive steps to adjust deliveries, manage inventory more effectively, and prevent potential losses.

**Trend Forecasting Success** evaluates the capability of predictive models to anticipate changes in demand across different stores, seasons, and product lines. This KPI focuses on how well models can predict peak sales periods and shifts in consumer preferences, enabling Rudi’s to align inventory levels with actual demand, minimize waste, and enhance customer satisfaction.

**Operational Efficiency Gains** measures the time saved in data analysis, reporting, and overall decision-making processes through automation and streamlined workflows. In a business environment where time is as valuable as in baking, this KPI highlights the improvements in efficiency achieved by reducing manual effort in sales reporting and analysis. By automating routine tasks, Rudi’s team can focus on strategic decisions that drive growth.

**Business Impact of Data-Driven Decisions** quantifies how insights derived from data analytics translate into tangible business outcomes. This KPI tracks the percentage of strategic decisions, such as adjusting production schedules, optimizing delivery routes, or launching targeted promotional campaigns, that are directly influenced by data insights. It also evaluates the financial impact of these decisions in terms of revenue growth, cost reduction, and operational improvements.

**Stakeholder Feedback & Report Effectiveness** focuses on assessing how well data reports and insights meet the needs of Rudi’s Bakery’s leadership and key stakeholders. The value of data insights lies not just in their accuracy but in their clarity and actionability. This KPI gauges stakeholder satisfaction with the reports, measuring how effectively they support decision-making processes. Regular feedback loops help refine the reporting approach, ensuring it remains relevant and impactful.

These KPIs collectively form the foundation for Rudi’s Bakery’s data-driven decision-making strategy. By focusing on improving anomaly detection, enhancing forecasting accuracy, increasing operational efficiency, and ensuring data insights directly influence business strategies, Rudi’s can optimize its operations, reduce costs, and drive sustainable growth. Continuous monitoring and refinement of these KPIs will ensure that the company remains agile, customer-focused, and competitive in an evolving market landscape.

This structured approach not only enhances day-to-day operations but also supports long-term strategic planning, positioning Rudi’s Bakery for continued success.

# **Model/Workable Solution**

The comprehensive analysis using various predictive models has provided valuable insights into sales forecasting, inventory optimization, and strategic planning.

**Impact**

Each model contributed unique perspectives, helping to identify strengths, limitations, and opportunities for business improvement. Below are the models used for prediction and sales forecasting:

## **Models Used**

## **1. Random Forest**

The **Random Forest** model demonstrated the highest accuracy in sales forecasting, making it a reliable tool for generating detailed predictions. Its ability to assess feature importance offers deeper insights into key sales drivers, enabling businesses to refine demand planning, optimize inventory management, and craft more data-driven marketing strategies. Utilizing this model will help in predicting sales trends more accurately and adjusting operations proactively.

## **2**. **Prophet**

The **Prophet** model effectively captures seasonal trends, making it ideal for forecasting sales fluctuations driven by recurring events. However, to handle irregular demand patterns more efficiently, integrating external variables such as holidays, promotional campaigns, and competitor activities is recommended. By incorporating these factors, the model's predictive power can be enhanced, leading to better inventory management and resource allocation during peak sales periods.

## **3. Exponential Smoothing**

**Exponential Smoothing** has shown strength in modeling predictable seasonal trends and stable sales cycles. However, it struggles with sudden fluctuations and irregular demand spikes. Combining Exponential Smoothing with machine learning models like Random Forest could create a hybrid approach that captures both stable patterns and unexpected sales shifts, leading to more robust forecasting.

## **4. Decision Tree**

The **Decision Tree** classifier provided a simple and interpretable framework for categorizing sales into high and low-performance groups. While the model effectively identified key relationships, it exhibited a tendency to overfit, which can reduce its generalizability. To improve its accuracy, hyperparameter tuning, decision boundary optimization, and the inclusion of additional features are recommended. These adjustments will reduce false negatives and improve the model's overall reliability.

## **5. Extra Trees**

The **Extra Trees** model outperformed the Decision Tree by handling class imbalances more effectively and improving overall accuracy. Its ensemble learning technique allowed for better generalization and deeper insights into complex sales patterns. This model is particularly useful for high-dimensional data and should be utilized for in-depth sales forecasting and anomaly detection tasks.

## **6. Clustering Analysis**

The **Clustering Analysis** on net sales by ZIP codes revealed clear geographical trends, helping to identify high-performing locations and underperforming areas that require strategic adjustments. This analysis supports targeted marketing, optimized resource allocation, and inventory distribution based on regional demand. Businesses can use these insights to expand in thriving markets and implement corrective measures in struggling regions, ultimately improving overall sales performance.

## **Model Comparison Summary**

| **Model** | **MSE** | **R Squared** | **Strengths** | **Weaknesses** |
| --- | --- | --- | --- | --- |
| Linear Regression | 0.78 | 0.98 | Simple Interpretable | Sensitive to outliers |
| Random Forest | 0.05 | 1.00 | High Accuracy, handles Nonlinearity | Overfitting Risk |
| Prophet | Varies | Varies | Capture Seasonality and Trends | Struggles with irregular demand |
| Exponential Smoothing | Varies | Varies | Effective for seasonal Data | Poor with irregular demand |
| Decision Tree | N/A | N/A | Good for classification task | Prone to overfitting |
| Extra Trees classifier | N/A | N/A | Handles Complex pattern, reduces overfitting | Requires more computational resources |
| Clustering | N/A | N/A | Identify regional Sales pattern | Dependent on cluster quality |

# **Workable Solutions Used/Recommended**

## **Random Forest**

The **Random Forest** model demonstrated superior performance in sales forecasting, with its ability to handle complex, non-linear data and provide valuable feature importance analysis. This makes it particularly well-suited for businesses aiming to refine their sales strategies and optimize inventory distribution.

By leveraging Random Forest, companies can gain deeper insights into sales drivers and make data-driven decisions that align closely with market demands and customer preferences. Its high accuracy and resilience against overfitting make it ideal for both short-term and long-term sales forecasting.

**Impact**

For Rudi’s, the insights derived from the Random Forest model can drive strategic business decisions like:

* **Optimized Demand Forecasting & Return Management** – By understanding the impact of returns on net sales, Rudi’s can adjust production, refine return policies, and improve product quality to minimize returns.
* **Profitability Enhancement Through Promotions** – The model highlights the role of trade discounts in driving sales, allowing Rudi’s to fine-tune promotional strategies to balance sales growth and profit margins.
* **Improved Inventory and Store-Level Strategies** – By focusing on net sales rather than gross sales, Rudi’s can prevent overstocking, reduce waste, and optimize supply chain efficiency. Additionally, stores with higher return rates can be targeted with better inventory management and customer education strategies.
* **Strategic Pricing and Revenue Optimization** – Understanding that **net revenue post-returns and trade discounts** are the strongest sales predictors enables Rudi’s to develop pricing models that maximize profitability while maintaining customer satisfaction.

By leveraging the Random Forest model, Rudi’s can transition from traditional gross-sales-focused forecasting to a more refined, return-aware approach, ensuring smarter decision-making, enhanced profitability, and improved operational efficiency.

**Feature Importance in Random Forest**

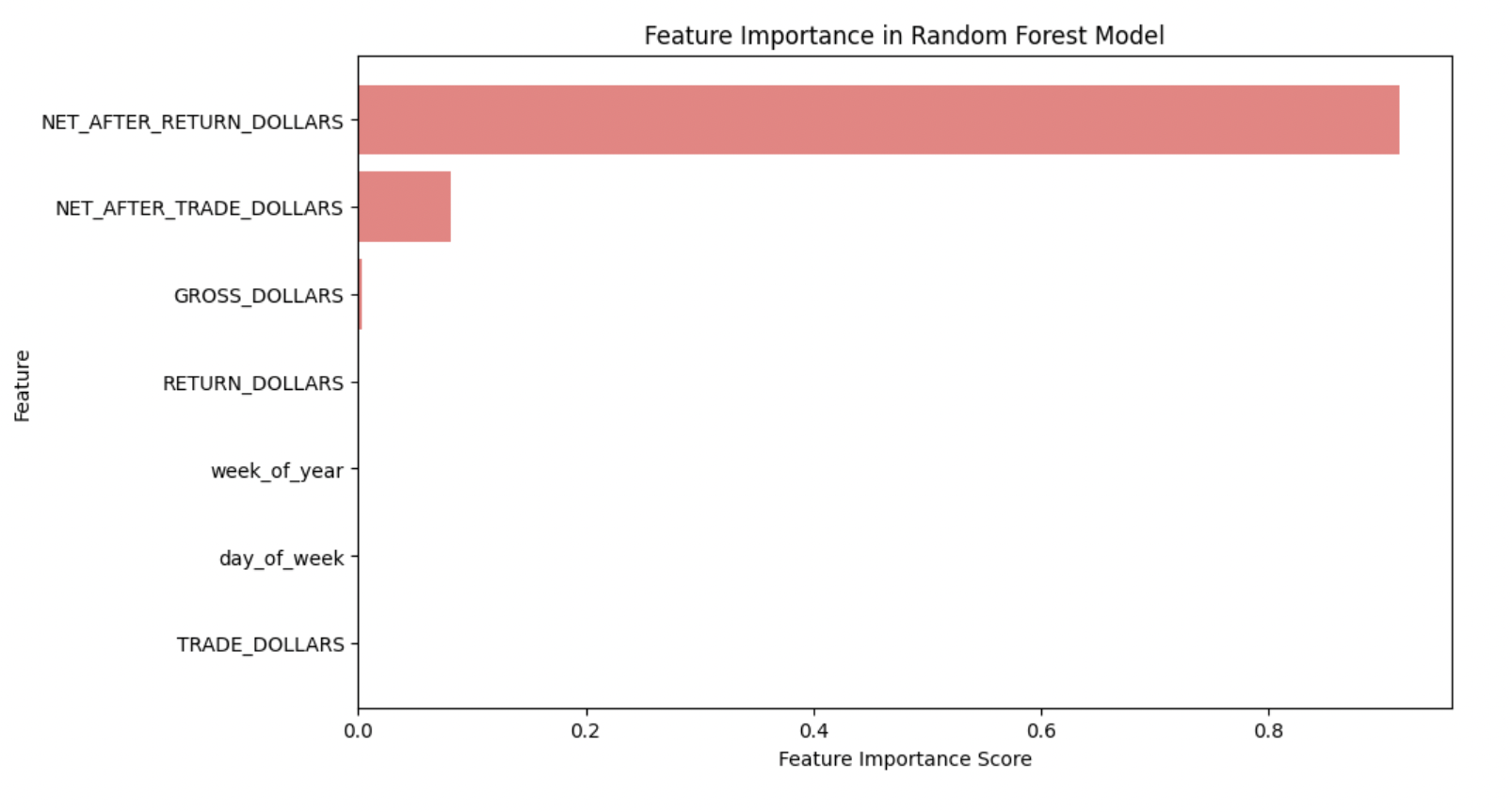


Fig. 1: Feature Importance

The image presents a **Feature Importance Chart** from a **Random Forest Model**, highlighting the most influential factors in predicting sales, demand, or returns for Rudi’s. The feature **NET\_AFTER\_RETURN\_DOLLARS** holds the highest importance, followed by **NET\_AFTER\_TRADE\_DOLLARS**, while other features contribute minimally.

### **Key Observations from the Feature Importance Chart**

* **NET\_AFTER\_RETURN\_DOLLARS** (most important feature)

This indicates that the net revenue after considering product returns has the greatest impact on forecasting. A strong reliance on this metric suggests that **returns significantly affect overall sales performance and profitability**. Understanding this factor can help optimize pricing, promotions, and return policies.

* **NET\_AFTER\_TRADE\_DOLLARS** (second most important feature)

This represents net revenue after accounting for trade discounts and adjustments. It suggests that **trade promotions, discounts, and deals play a critical role in sales forecasting**, influencing purchasing behavior.

* **GROSS\_DOLLARS & RETURN\_DOLLARS** (less important but still relevant)

**Gross sales revenue** has a lower impact than net metrics, likely because actual profitability depends on returns and discounts. **Return Dollars** contributes, reinforcing that **managing returns is crucial for accurate forecasting**.

* **Time-Based Features (Week of Year, Day of Week)**

These have the least importance in the model, suggesting that **seasonality and weekly trends may not strongly impact sales predictions compared to financial metrics**. However, they may still provide insights when combined with other variables.

### **Business Implications for Rudi’s**

* **Enhanced Demand Forecasting & Return Management:** Since **returns significantly impact net sales**, forecasting models should integrate return patterns to **optimize production and reduce waste**. If certain SKUs consistently show **high return rates**, Rudi’s can refine product quality, packaging, or expiration timelines.
* **Profitability Optimization via Discounts & Promotions:** The impact of **NET\_AFTER\_TRADE\_DOLLARS** highlights the role of promotions in sales trends. This suggests Rudi’s should analyze the effectiveness of trade discounts and optimize promotional strategies to maintain profitability.
* **Store-Level Strategy Adjustments:** Stores with **higher return rates or lower net sales after returns** should be targeted for policy adjustments. This could involve **stricter return policies, better customer education, or retailer-specific inventory management strategies**.
* **Integrating Returns into Sales Forecasting Models:** Traditional demand forecasting often focuses on **gross sales**, but this model suggests **net sales after returns is a better predictor**. Adjusting inventory based on net sales rather than gross sales can prevent overproduction and **improve supply chain efficiency**.

The Random Forest model significantly enhances sales forecasting accuracy by handling complex, non-linear data patterns and identifying key factors driving sales performance. Its ability to analyze feature importance provides valuable insights into the most influential predictors, such as **NET\_AFTER\_RETURN\_DOLLARS** and **NET\_AFTER\_TRADE\_DOLLARS**. This ensures that the model not only predicts sales more precisely but also accounts for critical financial adjustments like returns and trade promotions. Additionally, the model's robustness against overfitting makes it reliable for both short-term and long-term forecasting, ensuring more stable and data-driven predictions.

The **Random Forest Model’s feature importance analysis** reveals that **net sales after returns and trade dollars are the strongest predictors of future sales performance**. This insight allows Rudi’s to **optimize pricing strategies, reduce return rates, and refine promotional campaigns**. By focusing on returns and net revenue, Rudi’s can **develop a more accurate demand forecasting model**, ensuring **better inventory management, improved profitability, and reduced waste**.

## **Extra Trees Model**

This model's gradient boosting mechanism allows it to learn complex patterns more effectively, making it the preferred choice for sales forecasting and anomaly detection.

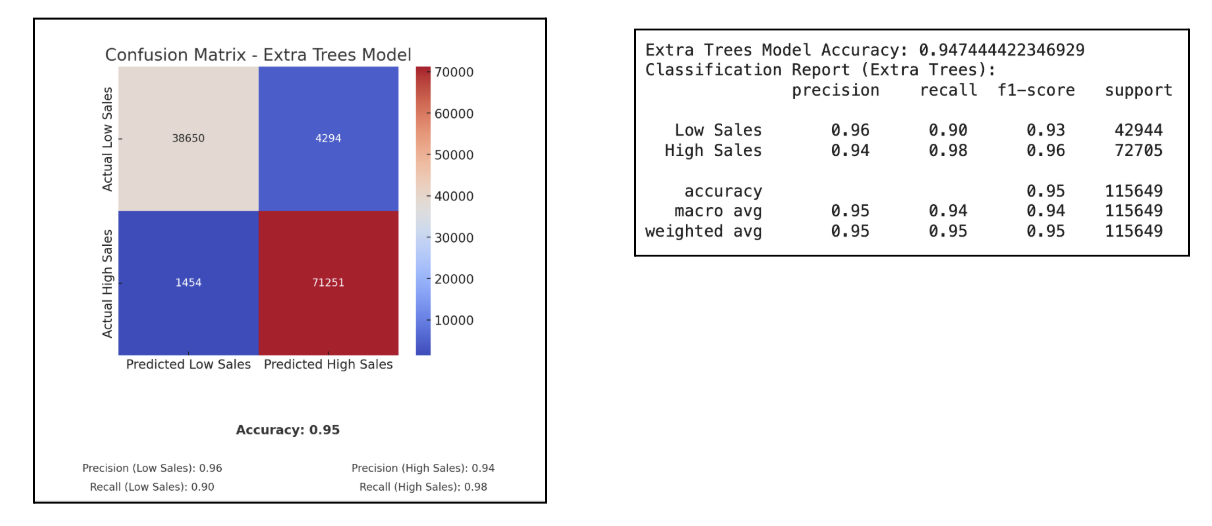


Fig. 2: Extra Trees Model

The image presents a Confusion Matrix and a Classification Report for an Extra Trees Model used to predict sales performance (categorized as Low Sales or High Sales). The model demonstrates high accuracy (95%), making it a strong candidate for forecasting demand and returns for Rudi’s bakery products. Additionally, the **Extra Trees** model further builds on the strengths of Random Forest by introducing enhanced randomness during tree construction, leading to better generalization and reduced variance.

This model excelled in handling class imbalances and complex data patterns, making it an excellent choice for anomaly detection and intricate sales forecasting tasks.

**Impact on Prediction**

1. **Better Handling of Class Imbalances** – The model effectively differentiates between high and low sales periods, ensuring that underrepresented patterns (such as rare spikes in demand) are not overlooked. This makes it more reliable in predicting anomalies and seasonal variations.
2. **Reduced Overfitting** – Unlike traditional decision trees, Extra Trees introduces additional randomness, reducing the risk of overfitting. This ensures that predictions remain accurate across different datasets and time periods.
3. **Faster Training and Scalability** – Since Extra Trees operates by using randomly selected splits rather than optimizing each split, it trains faster than traditional tree-based models. This makes it ideal for businesses dealing with large volumes of sales data.
4. **Robustness to Noisy Data** – Extra Trees can handle noisy, inconsistent, or incomplete sales data better than other models, ensuring that Rudi’s gets reliable insights even when faced with unpredictable market fluctuations.
5. **Early Warning System for Market Shifts** – The model’s anomaly detection capabilities can identify unusual sales patterns early, allowing Rudi’s to respond proactively to changing consumer behavior or external disruptions.

**Impact on Business**

1. **Optimized Pricing Strategies** – By analyzing factors influencing high and low sales, Rudi’s can refine its pricing models to maximize revenue and profit margins while staying competitive in the market.
2. **Personalized Marketing and Promotions** – The model’s ability to detect sales trends enables targeted promotions, ensuring that discounts and campaigns are optimized for the right customer segments and seasons.
3. **Supplier and Logistics Efficiency** – With more precise demand forecasting, Rudi’s can streamline procurement and distribution strategies, reducing excess inventory storage costs and improving order fulfillment.
4. **Store-Specific Performance Optimization** – By identifying locations with inconsistent sales patterns, Rudi’s can implement location-specific strategies, such as tailored marketing efforts, adjusted inventory levels, or improved customer engagement tactics.
5. **Competitive Advantage in the Market** – Leveraging a highly accurate and scalable model gives Rudi’s a **data-driven edge**, allowing it to react faster to market trends, minimize financial risks, and improve decision-making across all business operations.

### **Performance Metrics Breakdown**

* Confusion Matrix Analysis: The model correctly predicts 38,650 instances of Low Sales and 71,251 instances of High Sales. There are 4,294 false positives (predicted High Sales but were actually Low Sales) and 1,454 false negatives (predicted Low Sales but were actually High Sales). The false negative rate (1,454 errors out of 72,705 actual High Sales) is low, meaning the model is reliable in identifying high sales scenarios.
* Precision & Recall: Low Sales: Precision (0.96), Recall (0.90), F1-score (0.93). High Sales: Precision (0.94), Recall (0.98), F1-score (0.96). The model prioritizes recall for High Sales, meaning it is designed to catch as many high sales scenarios as possible, reducing understocking risks. Overall Accuracy: 95%, confirming strong predictive capability.

### **Business Implications for Rudi’s**

* Improved Demand Forecasting: The model accurately differentiates between high and low sales periods, allowing Rudi’s to anticipate demand fluctuations. This can optimize production scheduling, ensuring adequate inventory levels to meet market demand without overproducing.
* Better Inventory Management & Waste Reduction: By predicting low sales accurately, Rudi’s can minimize waste for perishable bakery items. High recall for high sales means the company can avoid stockouts, preventing lost revenue.
* Return Rate Optimization: High false positives (4,294) indicate some instances where demand is overestimated, potentially leading to higher returns due to overstocking. Adjusting the threshold for high sales classification can fine-tune predictions and reduce excess inventory.
* Retailer-Specific Sales Insights: The model’s outputs can help Rudi’s analyze sales trends across different retailers and store locations, identifying patterns of high returns and demand shifts.

To conclude, the Extra Trees Model provides a highly accurate framework for forecasting sales, enabling better inventory control, reduced returns, and optimized production planning. By leveraging this model, Rudi’s can make data-driven decisions to enhance efficiency, reduce costs, and improve profitability across its supply chain.

# 

# **Final Suggestion**

After conducting a comprehensive analysis using various predictive models, it is clear that adopting a data-driven decision-making approach will significantly enhance sales forecasting, inventory optimization, and overall operational efficiency. Based on the evaluation of the models, the *Random Forest* and *Extra Trees* models stand out as the most reliable and effective tools for forecasting and complex data analysis. The Random Forest model, with its high accuracy and ability to capture non-linear patterns, should be adopted as the primary forecasting tool. Its capability to identify key sales drivers through feature importance analysis enables businesses to fine-tune their strategies, optimize inventory distribution, and better align with market demands. This model not only offers strong predictive power but also provides flexibility in adapting to evolving business conditions.

The Extra Trees model further enhances forecasting accuracy by effectively handling complex data structures and class imbalances. Its ensemble learning approach reduces overfitting and increases robustness, making it suitable for large-scale data analysis and anomaly detection. This model is particularly valuable for businesses seeking deeper insights into sales trends, customer behaviors, and operational inefficiencies.

While the *Prophet and Exponential Smoothing models* have their strengths in capturing seasonality and trends, *their limitations in handling irregular demand* suggest that they should be used as complementary tools rather than primary forecasting methods. Combining these traditional time series models with machine learning approaches can create hybrid models that balance stability with flexibility, leading to more accurate and resilient forecasts.

To enhance sales forecasting and operational efficiency, it is recommended to adopt **Random Forest** as the core forecasting tool due to its high accuracy and interpretability, allowing for detailed predictions and the identification of key sales drivers.

The **Extra Trees** model should be utilized for complex data analysis and anomaly detection, providing deeper insights into sales dynamics and improving decision-making. To further refine seasonal forecasting, external factors such as holidays, promotions, and market trends should be integrated into **Prophet** models, enabling a more accurate representation of demand fluctuations.

Additionally, leveraging **clustering insights** will help optimize inventory management, streamline distribution processes, and implement targeted strategies for underperforming regions.

Implementing **hybrid forecasting approaches** that combine traditional time series models with machine learning can further enhance predictive accuracy by balancing stability with flexibility.

Finally, it is crucial to continuously refine predictive models through **hyperparameter tuning** and regular updates with new data to maintain their relevance and ensure long-term forecasting accuracy. These strategies collectively will empower the organization to make data-driven decisions, optimize resources, and drive sustainable business growth.

By adopting these strategies, the organization can improve operational efficiency, reduce costs, and increase profitability. A data-driven approach will empower the company to make proactive decisions, optimize resource allocation, and respond more effectively to market fluctuations, ensuring long-term growth and sustainability.

# 

# **Future Directions**

Building on the insights gained from the current analysis, future efforts should focus on enhancing predictive capabilities and expanding the scope of data-driven decision-making across the organization.

One key direction is to further refine the **Random Forest** and **Extra Trees** models by incorporating real-time data streams and external variables such as economic indicators, competitor pricing, and seasonal trends. This will not only improve forecasting accuracy but also allow for more dynamic and responsive inventory management.

Another important area for development is the integration of **advanced geospatial analytics**. By combining clustering analysis with geographic information systems (GIS), the company can gain deeper insights into regional sales patterns, customer demographics, and market saturation levels. This spatial dimension will enable more precise targeting in marketing campaigns and better-informed decisions regarding store expansions or closures.

**Hybrid modeling approaches** also present a valuable opportunity for future improvement. By combining traditional time series models like **Prophet** and **Exponential Smoothing** with machine learning algorithms, the company can create more resilient forecasting systems that balance short-term fluctuations with long-term trends. Such hybrid models can better handle irregular demand patterns, seasonal peaks, and market anomalies.

In addition, implementing **real-time anomaly detection** using models like **Extra Trees** can help identify sudden shifts in consumer behavior or operational inefficiencies, enabling quicker corrective actions. These models can also be extended to monitor supply chain disruptions, customer return rates, and product performance across different regions.

The organization should also focus on **enhancing data quality and granularity**. Collecting more detailed data on customer preferences, store-level promotions, and local events can enrich the predictive models and improve the precision of recommendations.

Leveraging **customer feedback data** and integrating sentiment analysis from social media and online reviews can offer further insights into consumer behavior and help tailor product offerings more effectively.

Lastly, adopting a more **agile data science framework** that supports continuous model training and real-time analytics will position the company to be more adaptive in a rapidly changing market environment. Regular model retraining, coupled with ongoing hyperparameter tuning, will ensure that the predictive tools remain accurate and relevant as new data becomes available.

These future directions can prove beneficial for the organization to evolve its data analytics capabilities, deepen its understanding of market dynamics, and make more proactive, informed decisions that drive growth and long-term success.

**Sales and Return Analysis by Customer/Product Cluster with Strategic Recommendations**

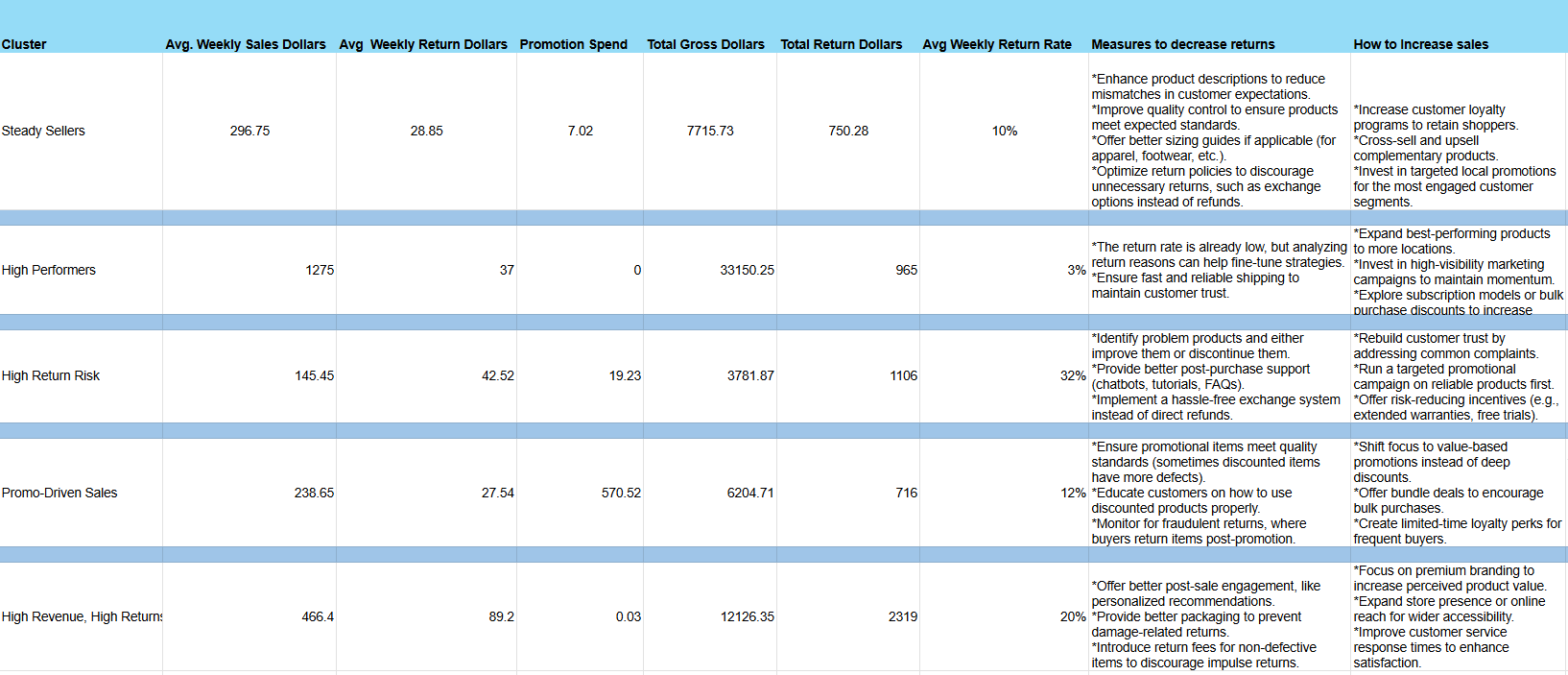


Fig. 3: Cluster Formation on important characteristics

**Top Retailers Associated with Each Sales and Return Cluster**



Fig. 4: Top Retailer based on the cluster formation

The table categorizes retailers based on their association with different sales and return clusters. "Steady Sellers" are linked to King Soopers, Safeway, and Whole Foods Market, indicating consistent sales with relatively low return rates. "High Performers," including Costco, Whole Foods Market, and Vitamin Cottage, likely generate strong sales with controlled returns. "High Return Risk" retailers, such as Safeway, King Soopers, and Vitamin Cottage, face higher return rates, suggesting potential issues with product quality or customer satisfaction. "Promo-Driven Sales," represented by Target and Super Walmart, rely on promotional strategies to drive sales. Lastly, "High Revenue, High Returns" retailers, including Sprouts Farmers Market, Vitamin Cottage, and Safeway, experience high sales volumes but also significant returns, requiring strategies to balance profitability with customer retention.

# **Adapting this information to an Interactive Dashboard**

Our team chose Tableau as the culmination of our workable solutions to get all members of Rudi’s involved in the data. The insights that can be pulled from summary statistics and the predictive models are valuable for drivers and executives alike. We are working on integrating our machine learning models previously mentioned into the Tableau during our next iteration. Although we are using an extract of the data to show its potential as a valuable solution, Tableau can pull from an active dataset and update the figures as they are entered. The value in this is that all people within the network can access real time figures in a quick to interpret manner. We currently have eight data worksheets that are amalgamated into 4 dashboards.

## **Retailer Summary Statistics**

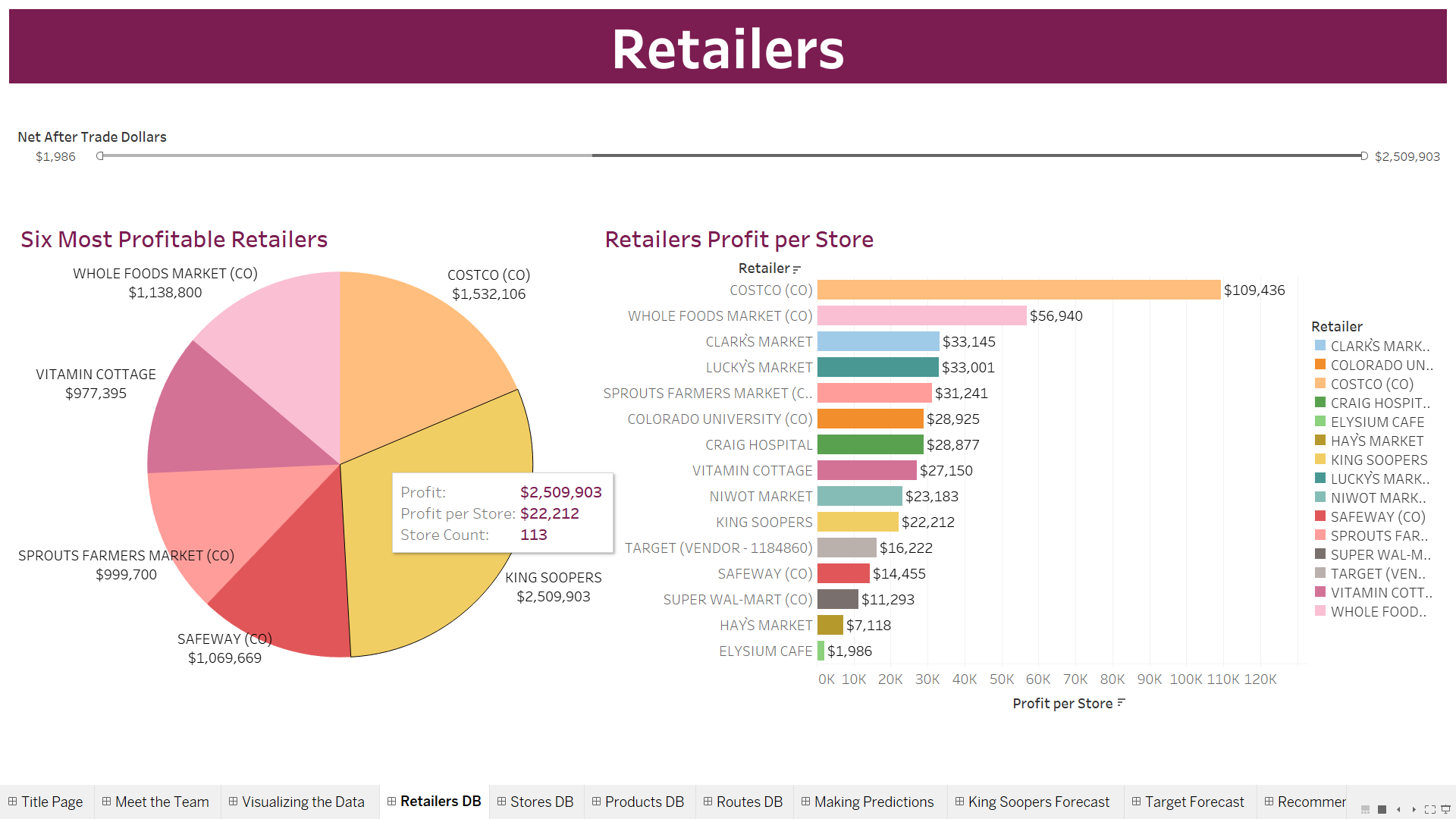


Fig. 3: Retailer Dashboard

This dashboard’s Net After Trade Dollars slider at the top adjusts both the Six Most Profitable Retailers pie chart, the Retailers Profit per Store, and can be used to display how much profit the selected range of retailers make.

When you hover over either chart, the tooltip displays profit, profit per store, and the store count.

## **Store Summary Statistics**

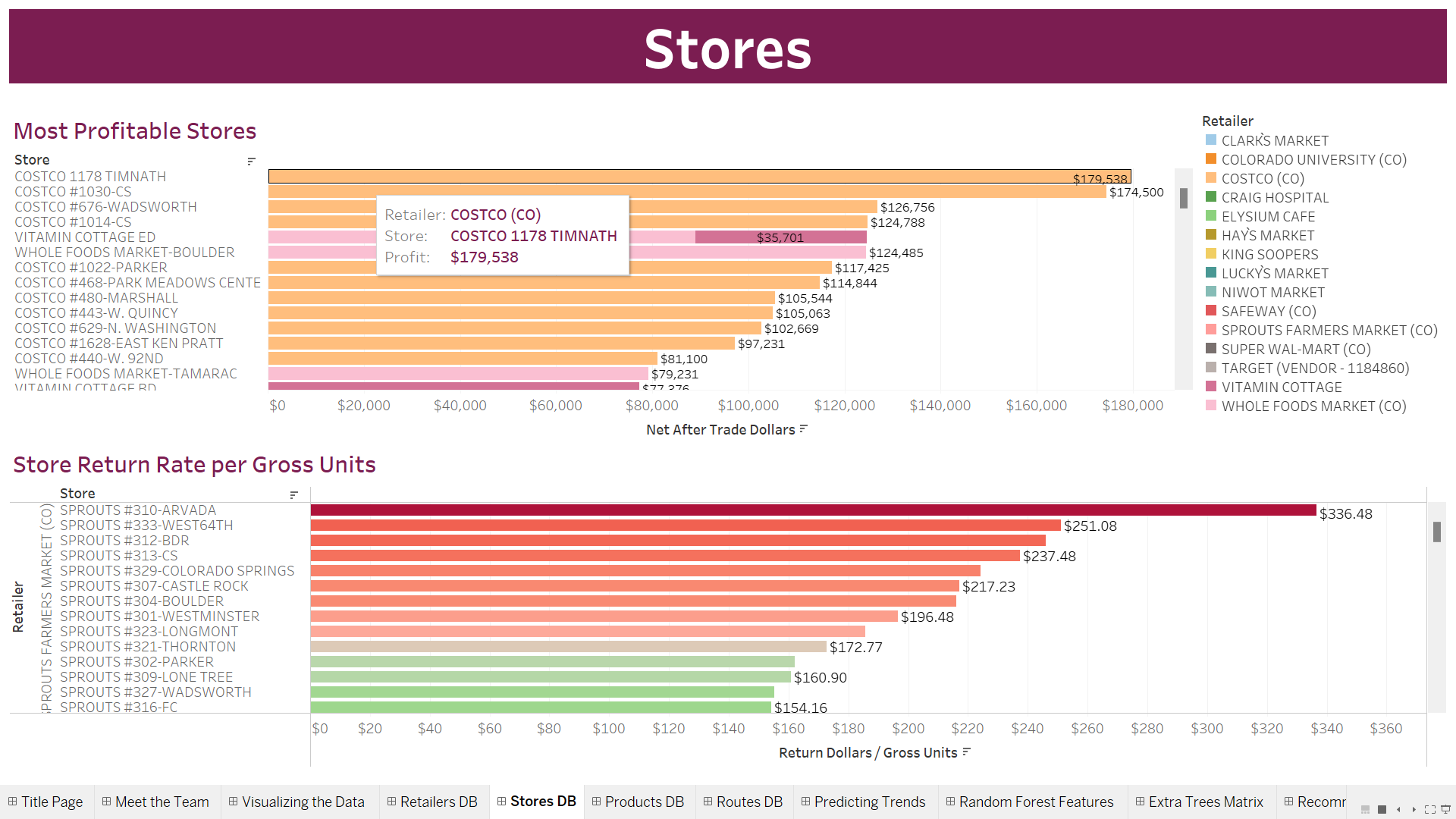


Fig. 4: Store Dashboard

Stores in the Store Return Rate per Gross Units bar graph can be highlighted to be isolated on the Most Profitable Stores chart. The key on the right highlights the retailer on the store profitability chart. The tooltip displays retailer, store, and either profit or return dollars per gross units.

## **Products Summary Statistics**

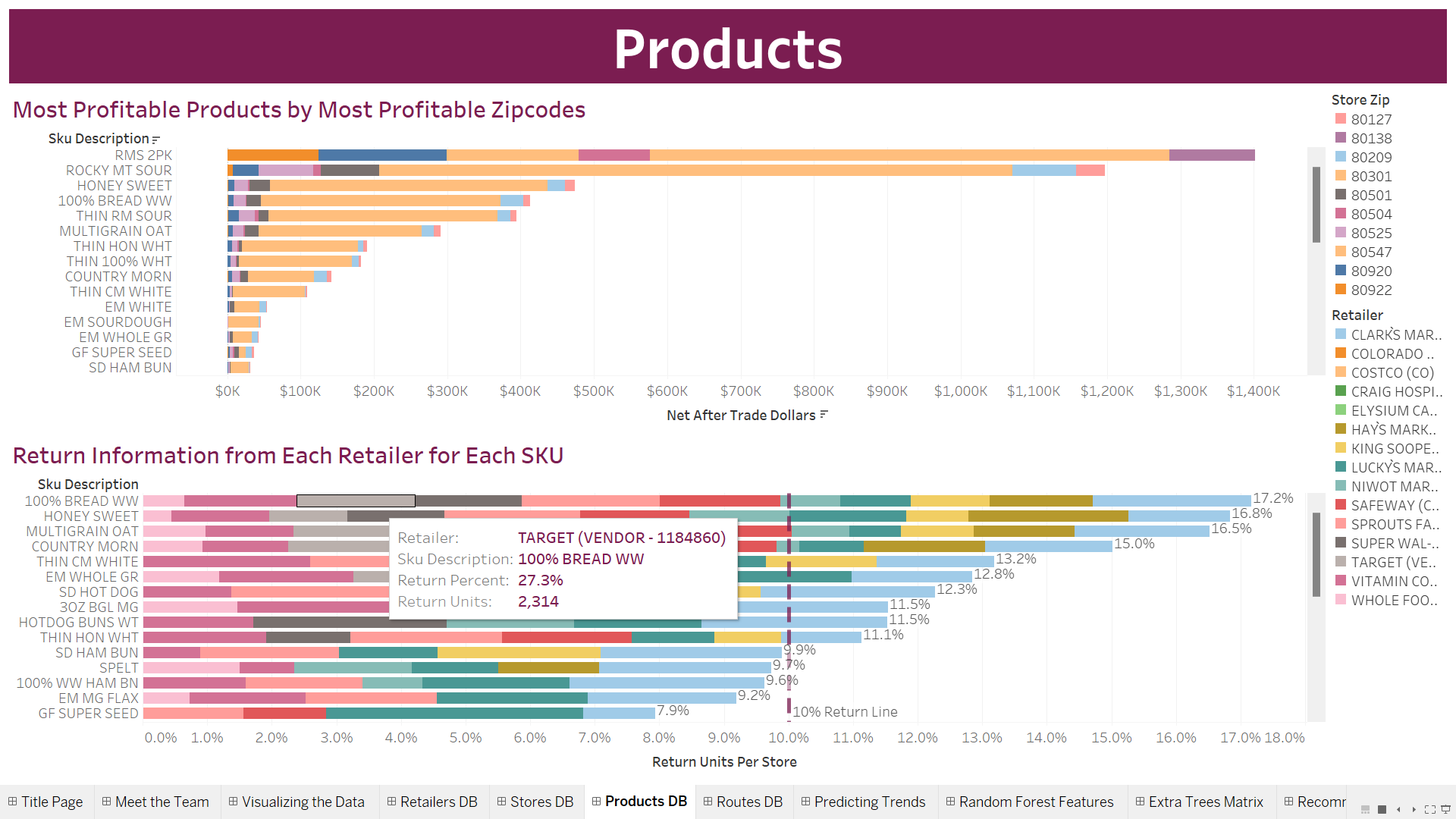


Fig. 5: Product Dashboard

The Most Profitable Products by Most Profitable Zip codes can be isolated by zip code where the tooltip contains SKU description, zip code, city and profit.

For Return Information from Each Retailer for Each SKU, the key on the right highlights SKU for the selected retailer. The tooltip includes retailer, SKU, percent of product returned, and returned units.

## **Route Summary Statistics**

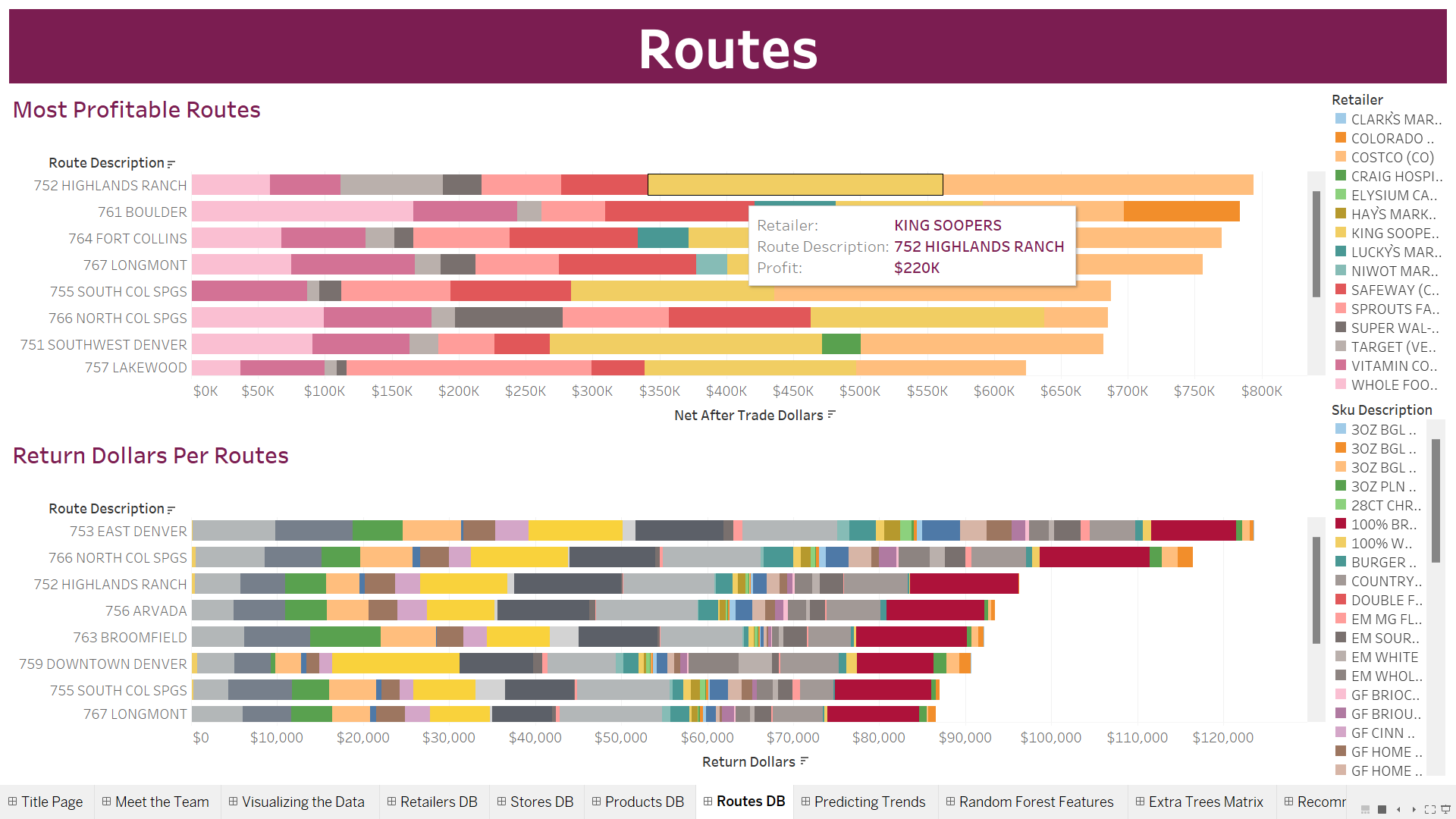


Fig. 6: Route Dashboard

The upper table, Most Profitable Routes, displays Total Net After Trade Dollars for each route, where you can highlight the retailer with the key on the right. The tooltip includes the retailer, route, and profit.

The Return Dollars Per Routes chart graphs route (with description) on the Y-axis and Return Dollars on the X-axis. The tooltip includes the route, SKU, and the Return Dollars.

# 

# **Conclusion**

The **Random Forest** and **Extra Trees** models stand out as the most reliable tools for sales forecasting and complex data analysis. Rudi’s should adopt these models for their high accuracy, scalability, and ability to handle intricate sales dynamics. By refining the Prophet model, leveraging clustering insights, and employing hybrid forecasting methods, companies can improve operational efficiency, reduce costs, and enhance overall profitability. These data-driven strategies will ensure that the organization remains competitive and responsive in an evolving market landscape.

Rudi’s values collaboration and shared knowledge so all aspects of the business can thrive; therefore having a platform where the statistics can be accessed by all stakeholders can be invaluable. Having a **Tableau interactive dashboard**, paired with the statistics derived from the data is another step in creating a successful future for the company.

# 

# 

# 

# 

# 

# 

# 

# 

# 

## 

## **Appendix**

## Generative AI prompts

1. Which predictive models would be most effective for analyzing sales trends and optimizing delivery routes for Rudi’s Bakery?
2. What steps should I take to clean and prepare Rudi’s sales data for analysis? I need to handle missing values, outliers, and ensure the data is ready for modeling.
3. Can you help me draft a clear vision statement for our project to improve Rudi’s Bakery’s DSD operations using data analytics?
4. How can I turn my analysis into practical solutions for improving Rudi’s product placement and delivery strategies?
5. How do I implement and compare various forecasting techniques on machine learning-based approaches?
6. What can we do and why clustering is important in the bakery industry?
7. Why should we focus on the return rate by retailer, store city, and product?
8. About store performances, is it important to showcase the most profitable stores?
9. What are the key business impacts and strategies based on the models I ran?
10. Can you help me understand the main strategies to understand the bakery business?
11. Help me understand the XGBoost ML algorithm to understand the performance.
12. What are the key strategies adopted by various food industries to lower the return rate?
13. What is the difference between return and exchange of products from the retail stores?
14. What are the top reasons customers/retailers return products excluding expiration of products?
15. How do external factors (location, seasonality, competition) influence store performance?
16. What are the key business insights derived from forecasting models in the bakery industry?
17. How can return rate analysis influence supply chain optimization and product quality?
18. How does packaging innovation impact return rates in the food industry?