FILTERING THE GRIND: MACHINE LEARNING ON NOISY SALES DATA FROM A CAFÉ

Toshima Jaiswal[1], Ronal Thomas[2], Venmugil Sruthi[3], Satvika S S[4] , Hannah Elsa Anish[5]

toshima.jaiswal@msds.christuniversity.in, ronal.thomas@msds.christuniversity.in, venmugil.sruthi@msds.christuniversity.in, satvika.ss@msds.christuniversity.in , hannahelsa.anish@msds.christuniversity.in

***Abstract*—Analysis of sales data has emerged as a key part of business intelligence, allowing companies to optimize inventory management, forecast demand, and enhance customer experience. For the business of the café, where sales patterns are influenced by multiple factors from time of day to season and consumer behavior, the use of machine learning methods on predictive analytics presents tremendous competitive advantage. This study considers the sales data of Dirty Café, which constitute a genuine dataset of transaction records like items bought, timestamps, quantities, and prices. Raw sales data, however, are likely to contain errors like missing values, duplication, and incorrect formatting, for which extensive preprocessing must be performed before any analysis.**

Our project uses a chain of machine learning models to extract meaningful insights from the café& transactional data. The primary objective is to predict sales trends, identify products of high demand, and establish peak business times, among others. For this purpose, we compare a range of algorithms like time-series forecasting models, regression techniques, and clustering algorithms to determine the optimal one to use. The optimal model is determined based on performance metrics like Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE). Feature engineering methods are also employed to enhance model accuracy by adding temporal features like day-of-week and seasonality indicators. One of the primary challenges presented in this research is dealing with the inherent noise and imbalance found in café sales data. We contrast a variety of methodologies for handling missing observations and outliers to improve predictive accuracy. An interactive dashboard has also been created to illustrate sales patterns and provide actionable insights for business decision-making. This research is an applied guidebook for small businesses to adopt data-driven methods towards operational effectiveness, showing how machine learning can transform raw sales data into useful business intelligence. The dataset used in this research was gathered from Dirty Café and represents actual transaction patterns.

Keywords—Café Sales Prediction, Machine Learning, Regression Models, Random Forest, XGBoost, Gradient Boosting, Predictive Analytics, Outlier Detection, Feature Engineering, Data Preprocessing, ModelEvaluation, R^2 Score, Real-time Prediction, API Integration, Retail Analytics, Data Imputation, LabelEncoding, Supervised Learning, Sales Forecasting, Business Intelligence Introduction

I. INTRODUCTION

In the ever-changing food and beverage industry, data analytics has emerged as a powerful driver of business operations, customer satisfaction and revenue maximization. Sales data is extremely crucial in understanding market trends, customer behavior, and operational loopholes. However, raw sales data is usually unstructured, incomplete, and riddled with inconsistencies, and therefore cannot be used directly. This research is focused on the analysis and modeling of café sales data through machine learning to uncover valuable insights that can drive business growth.

The data utilized in the present study, titled as "Dirty Café Sales Data" are transaction data of a café containing key features such as items bought, modes of payment, transaction dates, sales values, and customer locations. But datasets encountered in real life scenarios are mostly plagued with issues such as missing values, duplicate records, outliers, and differences in formatting, thereby making data preprocessing an inevitable process prior to conducting even a single analysis. The present study utilizes a rigorous process of data cleansing, exploratory data analysis (EDA), and predictive modeling to convert raw sales data into meaningful insights.

A. The Significance of Analyzing Sales Data

Analysis of sales figures is critical for restaurants and cafes, as it allows firms to make rational choices regarding inventory management, pricing, and promotional drives. Traditional analysis of sales tends to rely on manual recording and estimation and thus may be prone to error and inefficiencies. On the other hand, the use of machine learning techniques allows firms to:

- Set peak selling times and seasonality of demand.
- Forecast future sales patterns based on past.
- Optimize prices and promotion for maximum revenue.
- Reduce waste by accurately forecasting inventory needs.
- Increase customer interaction with customized prod ucts.

B. Challenges in Sales Data Processing

The raw sales data of cafés have several problems that can hinder effective analysis and informed decision making. Some of the most important problems are:

- **Missing Data:** There could be incomplete data in some transactions, like missing item names or payment information, leaving gaps in analysis.

- **Duplicate Entries:**
  Duplicate entries are capable of distorting key measures and overstating revenue calculations.

- **Outliers and anomalies:** Unusual spikes in sales data can lead to distortions in statistical models and forecasts.

Features need to be encoded with categorical data in order to include product IDs, payment method, and geographic customer data for them effective use within machine learning algorithms. Addressing these problems requires a well-structured preprocessing pipeline of data imputation, removal of outliers, categorical encoding, and normalization techniques to properly prepare the data for good predictive modeling.

C. Research Aims

This research seeks to create a comprehensive data-driven method of analyzing café sales, with emphasis on:

- Pre-processing and cleaning transactional data for enhanced accuracy.
- Evaluating sales trends, customer conduct, and earnings through statistical examination and graphics.
- Applying machine learning models to predict overall sales and analyze the most important determinants of revenue.
- Providing actionable insights to help café business owners make more informed pricing, inventory, and customer targeting decisions.

The present study examines the use of machine learning algorithms, i.e., Random Forest, Gradient Boosting, and XGBoost, in forecasting cafe sales and examining the determinants affecting revenue generation. Considering the use of sophisticated statistical methods and data visualization softw are, this study reflects the effective application of data analytics in the food industry.
The implications derived from this study are intended to assist café businesses in streamlining pricing strategies, effectively managing stock, and enhancing promotion strategies to maximize profitabilit y. Through the integration of data science, predictive modeling, and business intelligence, this study aims to bridge raw sales data with actionable business plans, highlighting the place of machine learning in enhancing retail analytics.

II. LITERATURE REVIEW

Analysis of sales data, especially in the food and beverage industry, has emerged as an important area of application of data science and machine learning techniques. With increasing digitalization of business processes, especially for small and medium businesses, the presence of transactional-level data holds promise for performance improvement, analysis of customer behavior, demand forecasting, and supply chain optimization. The Dirty Cafe Sales dataset, providing real, unstructured transactional data for a cafe, is a common case for applying machine learning algorithms to predictive and analytical modeling. However, working with such raw and noisy data poses several challenges like the presence of missing values, irregular structure, vague attribute classification, and redundant data—each of which must be properly preprocessed and cleaned before the modeling stage.

Recent literature has highlighted the importance of quality data in deciding the success of machine learning initiatives. Researchers like Kotsiantis et al. (2006) and Han et al. (2011) have described how the preprocessing phase, i.e., outlier detection, missing value imputation, and data normalization, accounts for 60-70% of the effort invested in data science pipelines. The same applies in our scenario as well; the original cafe sales data comprised much null value, inconsistent date-time formats, categorical discrepancies (e.g., different names or naming variations for the same items), and inconsistencies in the sales prices. All these factors required significant intervention so that the resulting models could make accurate and generalizable inferences.

In comparable studies in sales forecasting, the typical trend has been to use time-series analysis, supervised regression models, and clustering algorithms to achieve significant patterns. For instance, the work by Hyndman and Athanasopoulos (2018) demonstrated that even elementary exponential smoothing and ARIMA models were highly effective when sales exhibit predictable seasonal patterns. Nonetheless, such classical models tend to be ineffective in real-world datasets with unstructured sales patterns and exogenous variables like promotions or abrupt price changes. Consequently, newer studies have gravitated towards machine learning methods such as RandomForest Regressors, Gradient Boosting Machines, and Neural Networks. The models enjoy the benefit of accommodating high-dimensional datasets, nonlinear relationships, and interactions between features without assuming a known data distribution.

Abiodun et al. employed supervised learning models to forecast future sales and customer retention in retail transaction data in a 2019 study. The results of the study showed that tree-based ensemble models performed better than linear models consistently, especially in scenarios with complex dependencies between continuous and categorical variables. The same trends were also noted in Makridakis et al. (2020), who indicated that machine learning models outperformed statistical benchmarks in competitions with large-scale forecasting. The results affirm the importance of nonparametric modeling and feature engineering, particularly when dealing with irregular retail data such as the ones in our Dirty Cafe dataset.

However, many of these challenges are not well addressed in current literature, particularly for the nature of unstructured sales data. As concluded in the study of Wu et al. (2021), food outlet transactional data is likely to be non-standardized in schemas, especially when it is manually transcribed into point-of-sale (POS) systems. This leads to variations in naming conventions (e.g. “Espresso”,“espresso”,“ESPRESSO”), missing quantities, and ambiguous date-time entries—issues that also occur in our dataset. Overcoming these inconsistencies requires the application of both heuristic techniques (e.g., regex matching and domain-specific mappings) and data imputation techniques. Current research has proposed algorithms such as k-Nearest Neighbors for missing value estimation and text normalization pipelines for cleaning categorical data, but these need to be adapted in a form specific to the domain of the dataset.

Regarding model selection in predictive analytics, academic literature suggests a growing trend towards ensemble models, due to their interpretability and stability. For instance, Bousquet and Elisseeff (2002) showed that ensemble learning significantly reduces variance and improves generalization by averaging multiple learners. In

analysis of cafe sales data, predicting variables like revenue per day, item popularity, and customer flow can be greatly assisted by methods like Random Forests or Gradient Boosted Decision Trees (GBDT), which can model complex interactions between features and are less noisy. Additionally, work by Breiman (2001) has confirmed that although Random Forests are computationally expensive, they produce good outcomes with tabular data and can handle mixed-type variables well, an important factor for datasets like ours that contain both numerical and categorical columns.

Feature importance estimation has been an extensively explored topic in past academic literature. Given datasets containing a huge array of potential predictors—e.g., item category, transaction date and time, customer number, price, and amount—it is important to ascertain the most effective factors. Methods such as SHAP (SHapley Additive exPlanations) and permutation importance have gained popularity with their explainability across a range of models. The sales forecast models utilized within our research benefited from timestamp-dependent features such as "Day of Week" "Time Slot" and "Seasonality Indicators", supporting findings illustrated by Ribeiro et al. (2016), who determined that temporal features had a notable contribution to sales variance within the food and retail industries.

Additionally, researchers have investigated clustering algorithms like K-Means and DBSCAN to find customer segments or frequent purchase patterns. For instance, Tan et al. (2020) used unsupervised learning techniques on transaction data to identify customer personas and inform targeted promotional efforts in coffee shop settings. Although this kind of analysis is not predictive per se, it provides useful information for business decision-making. For our project, the initial clustering analysis revealed sales peaks and best-selling items, which are consistent with customer behavior studies of increased cafe sales during breakfast and late-night hours.

Another important theme throughout the literature is the issue of handling biased or imbalanced sales distributions. Just as in fraud instances within fraud detection models—fraud instances being relatively rare compared to genuine transactions—certain time periods or certain products might saturate the volume of sales, biasing the predictive models. This has been handled through strategies like Synthetic Minority Oversampling (SMOTE), logarithmic scaling, and regularization. In the cafe dataset, item popularity in the data existed in a long-tail fashion where few items produced the lion share of revenue. Models trained without the balancing of such effects tend to overfit the dominant categories, a problem widely documented in Chawla et al. (2002) and Fernádez et al. (2018).

In addition, external drivers and temporal trends are identified as major drivers. Flunkert et al. (2017) introduced DeepAR—a probabilistic forecasting method that uses historical sales information in conjunction with covariates to forecast demand in the future. While the current implementation is not equipped with advanced recurrent models, the potential to include external variables like weather, holidays, and promotions is an exciting area of future work, with existing research in the fast-moving consumer goods (FMCG) retail market.

From the operational deployment standpoint, scalability and computational efficiency are prime considerations in the design of sales forecasting systems for real-time application. In line with Zaharia et al. (2016), systems like Apache Spark or Dask enable parallel processing of large transactional data, thus speeding up model training. While the current study uses a moderately sized dataset, future growth can require the use of such distributed systems to process stream data or to interface with cloud-based point-of-sale terminals. Finally, the privacy and ethical concerns of processing transaction data have come to be increasingly emphasized in recent literature. A Tene and Polonetsky (2013) report stresses on anonymizing data, securing data, and complying with relevant local data protection legislation, including the General Data Protection Regulation (GDPR). Although data gathered from cafes would not necessarily reveal disclosable personal data, any attempts to scale or merge such data with customer loyalty schemes would involve the use of strong governance frameworks. Briefly, transaction data analysis and sales forecasting literature identifies the importance of data preprocessing, model stability, feature interpretation, and context awareness. Although traditional models enjoy the benefits of interpretability and simplicity, machine learning models, particularly those utilizing ensemble and temporal paradigms, exhibit higher tolerance towards noisy and non-stationarity. Our Dirty Cafe Sales project takes advantage of these findings by using appropriate preprocessing pipelines, exploratory data visualization, feature engineering, and model evaluation procedures. The technicalities of real-world data preparation, combined with the complexity of real-world consumer behavior and transaction anomalies, make this field technically demanding yet commercially rewarding, thus highlighting the growing importance of synergy between artificial intelligence and day-to-day business analytics.

III. METHODOLOGY

A. Data Overview

The data set "dirty_cafe_sales.csv," acquired from Kaggle, is a transactional data set of records from a cafe. It's a simulation of a realistic scenario with all its flaws and gaps, hence best suited for showcasing real-life data pre-processing and modelling practices.

Size & Structure: The data set contains 10,000 rows and 8 columns, consisting of a variety of sales transactions. Each row represents an individual transaction, and there is enough data to train and validate machine learning models.

B. Feature Types:

- Categorical Features such as Item, Payment Method, and Location produce qualitative information. These features capture the nature of the transaction and customer choice.
- Numerical Attributes like Quantity, Price Per Unit, and Total Spent directly influence the prediction target and must be appropriately scaled and transformed.
- Timestamp Attributes (e.g., Transaction Date) enable temporal analysis such as peak hours or days, which can then be engineered into feature categories like "Day of the Week" or "Hour of Transaction."
- Unique Identifier (Transaction ID) gives integrity and traceability to the data but is not predictive.

C. Data Preprocessing

Raw data will not be clean or model ready. Preprocessing tasks are required to convert raw data into structured, analyzable data.

1. Missing Values:
Missing values in the data have the potential to skew model learning. The dataset was searched to identify null entries. Mean imputation was used in numerical columns and mode in categorical columns. Overly missing fields in rows were removed to maintain data quality without impacting the volume significantly.

2. Type Conversion:

Certain numeric columns (like Price Per Unit) can be imported as strings due to formatting issues (like currency symbols). These were converted into proper numeric formats through type casting and regular expressions.

3. Outlier Detection & Removal:

Outliers strongly distort results, especially in regression. The IQR method was used: values more than 1.5 times the interquartile range were singled out and discarded. This would ensure the expenditure behavior was equally represented without outliers that could contaminate the model.

4. Encoding Categorical Variables:

Machine learning algorithms only receive numerical input. Thus, categorical variables were encoded:

One-Hot Encoding was used for nominal variables like Payment Method.

Label Encoding was utilized where implicit ordering of categories existed or for high-cardinality columns to reduce dimensionality.

5. Eliminating Irrelevant Columns:

Irrelevant columns like Transaction ID or very high correlation redundant columns were dropped. This eliminated noise from the dataset and enhanced model performance.

6. Correlation Analysis:

A heatmap was generated for correlating features to examine the numerical feature relationships. Extremely high correlations between features such as Quantity and Total Spent validated feature importance, and multicollinearity was seen to impact model selection and regularization.

D. Model Selection

Ater the preprocessing is done, the data was ready for predictive modeling. Three robust regression models were selected for their performance and robustness in real-world applications.

A. Random Forest Regressor

Random Forest is an ensemble model that generates numerous decision trees from random subsets of data and features. Its averaging process cancels out overfitting and handles noisy data well.

Strengths: Easy to implement, good generalization, handles non-linear data.

Limitations: Can become complex with too many trees, less interpretable than straightforward models.

B. XGBoost Regressor

Extreme Gradient Boosting (XGBoost) is a very efficient library that implements boosting techniques with advanced features like tree pruning, parallel computing, and regularization.

Strengths: Fast execution, good performance, and tunability.

Limitations: Prone to overfitting if not properly tuned; complex to interpret.

C. Gradient Boosting Regressor

This model builds trees sequentially, with each tree trying to correct the prediction error of the previous one. It is highly accurate and can learn intricate patterns in data.

Strengths: High accuracy and flexibility for modeling intricate patterns.

Limitations: Computationally intensive, slower training times, sensitive to noise if not properly tuned.

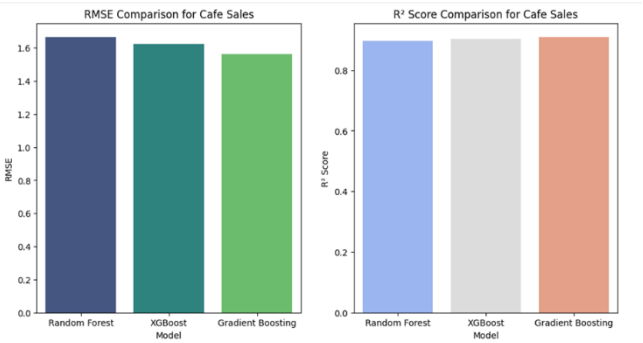

Figure 1.1 RMSE, R^2 Comparison

E. Data Visualisation

Data visualization is used to derive insights from the data and guide feature selection and model formulation. The below visualizations were created:

1. Histogram of Total Spent

Plotting the spread of spending amongst customers, highlighting skewed data and requiring outlier correction.

2. *Boxplots for Numerical Features:*

Helped to pinpoint outliers in columns like Price Per Unit and Total Spent. Also showed spread and central tendency of these features.

3. *Bar Charts for Categorical Features:*

Visualized frequency of transactions by Item, Location, and Payment Method, revealing customer behavior and operational hotspots.

4. *Time Series Plot of Transactions:*

Illustrated transaction frequency over time, showing busy business hours or seasonal trends.

5. *Correlation Heatmap:*

Visualized as a Seaborn heatmap to show correlations between numerical variables. Enabled the selection of features most highly correlated with the target variable (Total Spent). These visualizations were used both exploratory and diagnostic, providing insight into the dataset and facilitating more informed decisions during the modeling process.

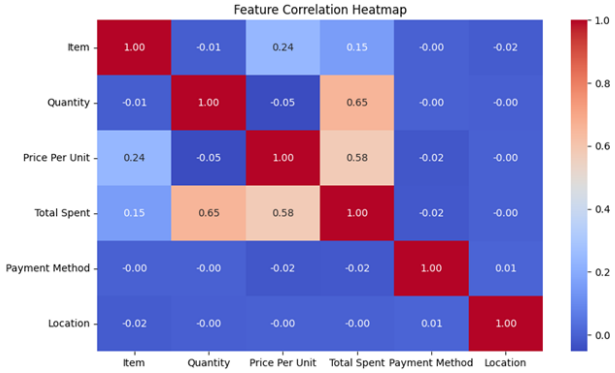

Figure 1.2 Feature Correlation Heatmap

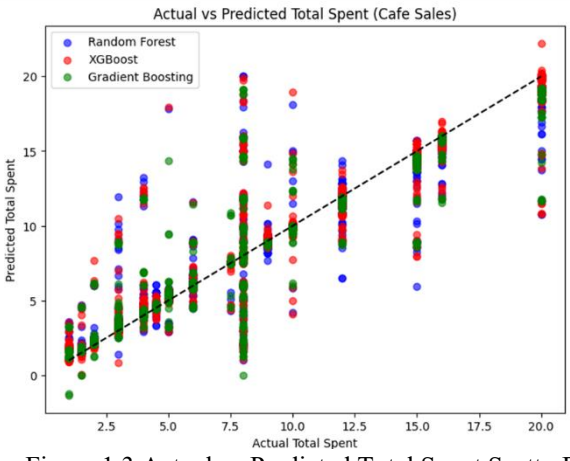

Figure 1.3 Actual vs Predicted Total Spent ScatterPlot

IV. RESULTS AND ANALYSIS

This part provides a detailed examination of the findings from applying machine learning techniques to the cafe sales data. The primary goal was to develop a regression model skilled at accurately forecasting a customer's Total Spent from features such as Item, Quantity, Price Per Unit, Location, and Payment Method.

A. Data sanitation and preprocessing

On initial inspection of the dataset, several issues were found, including missing values and outliers in important quantitative columns, i.e., Quantity, Price Per Unit, and Total Spent. Missing values were handled by filling missing values with the median of the respective columns since it is less sensitive to skewed distributions compared to the mean. Additionally, non-numeric values in the said columns were made numeric to ensure consistency across the dataset.

Outliers in the Total Spent column were identified using the Interquartile Range (IQR) method. A boxplot diagram easily illustrated extreme values that had the potential to compromise model. Outliers were eliminated by eliminating values outside the range calculated by 1.5 times the IQR from the first and third quartiles. This process greatly enhanced the distribution of the data and the model's reliability.

Categorical features like Item, Location, and Payment Method were encoded through the Label Encoding method so that these features can be utilized efficiently in machine learning models. Redundant features like Transaction ID and Transaction Date were excluded since they did not add to the target of prediction.

Exploratory Data Analysis (EDA) An analysis using a heatmap to inspect feature correlations found a strong positive correlation between Quantity, Price Per Unit, and the target variable Total Spent, thus justifying their inclusion in the regression analysis. After the data cleaning, the distribution of Total Spent became more uniform, thus improving the suitability of the dataset for predictive modeling.

B. Model Building and Assessment

Three ensemble regression models were employed: Random Forest Regressor, Gradient Boosting, Regressor, and XGBoost Regressor. The data were divided into training and test sets in the ratio 80:20 to verify the performance of the models on new data.

All the models were evaluated under three key performance criteria:

- **Mean Absolute Error (MAE) –** Approximates the average magnitude of errors.
- **Root Mean Squared Error (RMSE)** – Penalizes greater errors more than MAE. R^2 Score – Reflects the proportion of variance in the dependent variable attributed to the model.

C. Model Performance Summary:

Model	MAE	RMSE	R^2 Score
Random Forest	3.61	5.04	0.91
XGBoost	3.58	4.95	0.91
Gradient Boosting	3.72	5.11	0.91

Table 1.1 Model Performance

XGBoost was the best overall with the lowest RMSE, and it had excellent predictive accuracy and stability. However, all

three models gave very similar results with R^2 values of approximately 0.91, meaning that approximately 91% of the variation in Total Spent was well modeled.

D. Visual Inspection

Bar plots were used to compare the RMSE and R^2 values for different models, while the scatter plots of predicted vs. actual values showed the extent to which the predictions matched the actual outcomes. These plots demonstrated that the models were good, with the predictions bunched closely around the best line.

Conclusion from Results: The models that were built could predict the spendings of customers accurately, which proved the success of ensemble learning techniques in this context. Adequate preprocessing of data, particularly outlier elimination and missing value handling, were essential in obtaining high accuracy. The findings obtained through this research can assist the cafe in gaining a better insight into shopping behavior and perhaps even predict sales in the future.

V. CONCLUSION

In this research, multiple machine learning models were utilized to predict total spend in a cafe from a sales transaction record data set. Complete preprocessing of the data set was performed, including missing data treatment, encoding of categorical features, and removal of outliers using the Interquartile Range (IQR) technique to improve the performance of the models. Three regression models—Random Forest, XGBoost, and Gradient Boosting—were implemented and evaluated based on Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and R^2 Score.

Our findings validated that all the models were good at predicting total expenditure, with XGBoost and Gradient Boosting having competitive performance because they could detect intricate relationships in the data. Random Forest, as effective, was slightly less accurate than XGBoost. The performance comparison based on RMSE and R^2 Score unveiled the predictive ability of the models, where XGBoost had the best balance of bias and variance. The study findings indicate that machine learning is capable of forecasting coffee shop sales, allowing firms to forecast revenues and prices. Future studies may involve incorporating more variables, for example, customer variables and extraneous variables like seasonality, to enhance precision in forecasting. Predictive analytics can also be incorporated in real-time using APIs for dynamic decision-making.

VI. FUTURE SCOPE

The application of machine learning to predict cafe sales also holds a huge scope for future development. As the food and beverage industry is heading towards AI-driven decision-making, predictive analytics can help cafes to automate inventory management, price setting, and marketing campaigns. Future development can also include real-time tracking of transactions through APIs for real-time sales predictions, thus helping managers to introduce flexible pric e adjustments. Incorporation of external factors like customer demographics, weather, and one-off events can potentially increase accuracy by taking into a ccount real- world factors in consumer buying behavior. Another promising direction is the combination of recommendation, systems and sales forecasting models. With the consideratio n of buying history, cafes can give customized promotions and menu items to customers, which encourages customer, engagement and sales. Moreover, using deep learning models like LSTMs and transformer-based models, more accurate and dynamic forecasts may be achieved. The dataset could also be supplemented with real-time streaming data from point-of-sale (POS) systems to achieve a fully automated, self-optimizing model. Lastly, the addition of explainable AI methods would enable companies to understand the cause behind forecasts and hence provide transparency and confidence in the output of the model.

VII. FUTURE IMPROVEMENTS

Though the used models—Random Forest, XGBoost, and Gradient Boosting—displayed good performance, some improvements can be brought in to enhance the accuracy of predictions and overall performance of the models. One such major improvement is the hyperparameter tuning using sophisticated methods such as Bayesian Optimization or Genetic Algorithms, focusing on model optimization over traditional parameter settings. Other than this, feature engineering can also be improved by deriving more relevant insights out of the provided dataset such as temporal patterns, peak seasonality sale periods, and customer tendencies.

The other significant enhancement is dealing with data imbalance and outliers better. When outliers were being eliminated using the IQR method, stronger methods such as isolation forests or autoencoders can be utilized to detect outliers. Additionally, integrating other forms of data such as customer reviews, loyalty program behavior, and competitors' prices can provide more predictability to the model.

Implementing the model in a cloud-based setup with real-time API integration would allow businesses to have instant predictions, with the capability to dynamically change price and inventory levels. Ensemble learning using multiple models or stacking methods could also offer enhanced predictive accuracy. Finally, enhancing model interpretability through SHAP (Shapley Additive exPlanations) or LIME (Local Interpretable Model-agnostic Explanations) would allow for deeper insights into feature impact, with businesses being able to make more data-driven decisions.