Predicting Profitability in Retail Transactions: A Machine Learning Approach on the Superstore Dataset

University of Bologna
Master's Degree in Digital Transformation Management
Course: Machine Learning and Data Mining
Academic Year: 2024/2025

Alessandro Astolfi
Student ID: 0001120878
alessandro.astolfi2@studio.unibo.it

Francesco Russo
Student ID: 0001121020
francesco.russo60@studio.unibo.it

***Abstract*—This project investigates the feasibility of predicting retail transaction profitability using machine learning techniques applied to the Superstore Sales dataset. The primary objective is to identify which factors most strongly influence profit, with particular attention to discount levels, sales volumes, and product categories. Following the CRISP-DM methodology, the analysis encompasses data understanding, preparation, modeling, and evaluation phases. After appropriate data cleaning and feature engineering, several algorithms were compared, including both linear and ensemble-based models. The final optimized Random Forest Regressor achieved an R^2 of 0.829, confirming the inherently non-linear nature of the problem. Feature importance analysis highlighted *Sales* and *Discount* as dominant predictors, underscoring their central role in shaping profitability. Overall, the study demonstrates how interpretable machine learning models can provide actionable insights to support pricing and promotional strategies in retail environments.**

1. Introduction

In today's competitive retail landscape, understanding the underlying drivers of profitability is critical for data-driven strategic decision-making. While higher sales volumes are generally associated with greater profits, additional factors—such as discount policies, product categories, shipping times, and promotional timing—can significantly affect overall financial performance.

This project aims to develop a machine learning model capable of predicting the profit associated with each transaction in the **Superstore Dataset**. The objective is twofold: to accurately estimate transaction-level profit and to identify the most influential variables driving profitability.

To ensure a structured and reproducible workflow, the study adopts the **CRISP-DM** (Cross Industry Standard Process for Data Mining) framework. The pipeline includes phases of data understanding, cleaning, feature engineering, modeling, evaluation, and interpretation. Both **linear** and **ensemble models** were evaluated, with systematic use of preprocessing pipelines, cross-validation, and hyperparameter tuning to ensure robustness and generalization.

Beyond predictive accuracy, the study emphasizes model interpretability—a key requirement for decision support in business contexts. By analyzing feature importance and model behavior, the results provide insights into how discount levels, product types, and operational efficiency interact to determine profit outcomes.

The following sections detail the methodology, experimental results, and conclusions drawn from this analysis.

2. Related Work

The task of predicting sales and profitability in retail has attracted significant attention, with a broad body of literature exploring how data-driven models can support strategic and operational decisions. Early studies primarily relied on traditional regression techniques and time-series models, which often proved inadequate for capturing the non-linear patterns, seasonal fluctuations, and regional heterogeneity typical of retail environments.

In response, more recent research has adopted machine learning methods capable of handling higher complexity and larger volumes of structured data. Among these, supervised learning algorithms—such as decision trees, Random Forest, XGBoost, and other ensemble models—have gained particular prominence. These approaches are appreciated for their ability to capture variable interactions, handle mixed data types, and maintain a relatively high degree of interpretability compared to deep learning models.

Ensemble techniques, in particular, have consistently demonstrated superior performance over linear models in real-world applications, especially when dealing with high-dimensional or non-linear datasets. They are widely used to estimate future sales, evaluate the impact of pricing or discount policies, and identify the key determinants of profitability. In some cases, hybrid approaches have been

proposed, combining machine learning with statistical techniques or expert rules to integrate domain knowledge and improve generalization.

Deep learning models—such as Convolutional Neural Networks (CNNs) and Long Short-Term Memory (LSTM) networks—have also been applied, particularly in contexts involving high-frequency, timestamped data (e.g., e-commerce clickstreams or in-store sensor data). However, their use is generally reserved for large-scale operations where their complexity and data requirements are justified.

From a methodological standpoint, the role of feature engineering and interpretability has gained increasing importance. Tools such as SHAP values, permutation importance, and model-agnostic explainers have been used to clarify how variables like sales volume, discount level, or product category contribute to predicted outcomes. This transparency is especially valuable for business stakeholders, as it supports trust in the model and facilitates actionable insights.

Public datasets such as the Superstore Sales dataset have played a key role in advancing experimentation and benchmarking in the field. Its rich transactional structure—including detailed information on sales, profit, product hierarchies, customer segments, and geographical context—makes it a preferred resource for both academic studies and applied machine learning exercises. Many open-source projects use it to experiment with regression and classification models, analyze profitability, or simulate promotional strategies in a controlled setting.

Overall, the literature highlights a convergence toward ensemble learning, hybrid modeling, and model interpretability as key components of modern retail forecasting. This context forms the conceptual and technical foundation for the methodology developed and evaluated in the present work.

3. Proposed Method

The predictive modeling framework followed the **CRISP-DM methodology** and was designed as a supervised regression task aimed at estimating the *Profit* per transaction. The process consisted of five main components: data understanding, data preparation and feature engineering, outlier detection and treatment, modeling, and model optimization. Each phase progressively refined data quality, model robustness, and interpretability.

3.1. Data Understanding

The analysis was conducted using the **Superstore Dataset**, where each record represents a single retail order. The dataset includes commercial variables (*Sales*, *Profit*, *Discount*, *Quantity*), product information (*Category*, *Sub-Category*), geographic attributes (*Region*, *State*), customer segmentation (*Segment*), and temporal information (*Order Date*, *Ship Date*). The target variable, *Profit*, is continuous and expressed in monetary units.

A preliminary **exploratory data analysis (EDA)** was performed to assess the dataset's structure and quality. No missing values were found, and one duplicate entry was identified and removed. Additionally, during the examination of individual transactions, a record with an unusually negative *Profit* despite a very small *Discount* was detected and removed. This early observation suggested the potential presence of additional **outliers**, which were later analyzed and treated more systematically.

The dataset displayed a right-skewed distribution for both *Sales* and *Profit*, indicating that a small number of large transactions contributed disproportionately to total revenue. The *Office Supplies* category represented the majority of sales, suggesting a predominantly **B2B** nature. Correlation analysis revealed a moderate positive relationship between *Sales* and *Profit*, and a strong negative correlation between *Discount* and *Profit*. Visual inspection confirmed that discounts above approximately 20% often resulted in negative margins, highlighting the importance of discount-related variables in profitability modeling.

Temporal features were derived from *Order Date* to create *OrderYear* and *OrderMonth*, capturing potential seasonal effects. A new variable, *ShippingDays*, was computed as the difference between *Ship Date* and *Order Date*, providing a quantitative measure of delivery efficiency.

3.2. Data Preparation and Feature Engineering

The data preparation phase aimed to ensure consistency, interpretability, and readiness for modeling.

Feature engineering expanded the dataset with business-relevant transformations:

- **Discount Range:** categorical segmentation of *Discount* into brackets (e.g., 0–10%, 10–20%), designed to capture non-linear pricing effects.
- **Promo Season:** categorical label assigned to each transaction based on the order date, identifying the associated promotional period (e.g., Black Friday, Christmas); this feature was later one-hot encoded during preprocessing.
- **Shipping Days:** numeric difference between order and shipment dates, representing logistical efficiency.

Variables with high cardinality or limited predictive value (*Order ID*, *Customer ID*, *Product Name*) were excluded. The dataset was divided into numerical and categorical subsets and preprocessed through a unified pipeline built with `ColumnTransformer`. Categorical variables were one-hot encoded, while numerical features were scaled only for linear models, as ensemble methods are inherently scale-insensitive.

3.3. Outlier Detection and Treatment

During the initial exploration and baseline modeling, both the Linear Regression and Random Forest models exhibited moderate predictive performance and instability. Diagnostic plots of *Profit* versus *Sales* revealed a small number of transactions with extreme profit values—either highly

positive or strongly negative. Some of these anomalies were linked to heavy discounts, while others had no clear business justification, indicating that **statistical outliers** were influencing the model's learning process.

A two-step strategy was adopted to identify and mitigate these effects:

1) **Visual inspection**, using scatter plots and log-transformed axes, to detect observations far from the dense cluster of normal transactions.
2) **Statistical filtering**, where records outside the 1st and 99th percentiles of *Profit* were removed.

The scatter plot below provides a visual confirmation of the presence of outliers in the dataset. Most transactions are concentrated in a dense region near low sales values and moderate profits, while a few extreme observations with unusually high or low profit stand out clearly.

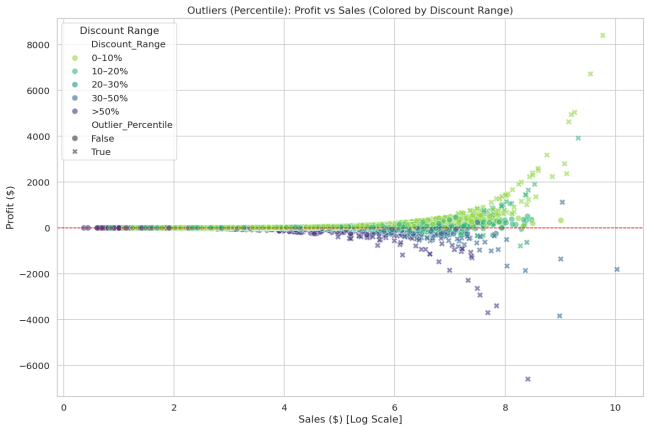

Figure 1. Profit vs Sales (log scale).

This procedure eliminated roughly 2% of the total observations and produced a refined dataset referred to as **"Normal Orders"**. The resulting data preserved the main structure and variability of the original dataset while minimizing the impact of extreme values, leading to a more stable foundation for subsequent modeling.

3.4. Modeling Approach

The predictive task was structured as a **supervised regression problem**, with *Profit* as the continuous target variable. The modeling process followed an incremental approach, progressing from simple linear models to more flexible non-linear ensembles.

In the first phase, two baseline models were trained using a conventional 80–20 train–test split:

- **Linear Regression**, serving as a benchmark for linear relationships.
- **Random Forest Regressor**, introduced to capture non-linear interactions and complex dependencies among features.

Both models were trained within the same preprocessing pipeline, ensuring consistent transformations and avoiding data leakage. While the linear model offered interpretability, the Random Forest demonstrated greater flexibility but signs of overfitting. To reduce variance and improve generalization, regularization parameters such as `max_depth` and `min_samples_leaf` were adjusted in later iterations.

After stabilizing the dataset through outlier removal, a **5-Fold Cross-Validation** strategy was introduced to obtain more reliable estimates of model generalization. At this stage, two additional ensemble algorithms—**XGBoost** and **LightGBM**—were incorporated to benchmark against the Random Forest and Linear Regression baselines. This cross-validated comparison enabled the identification of the most suitable modeling family, balancing accuracy, interpretability, and computational efficiency.

3.5. Model Optimization and Feature Selection

Following the comparative modeling phase, **hyperparameter optimization was performed on the Random Forest model** using `RandomizedSearchCV`. This approach efficiently explored a predefined range of parameters—such as the number of estimators, maximum tree depth, and minimum samples per leaf—evaluating each configuration through cross-validation. The objective was to identify parameter combinations that enhanced the Random Forest's predictive stability and reduced overfitting, while maintaining a reasonable computational cost.

Subsequently, **feature importance analysis** was performed on the tuned Random Forest model to identify the most influential predictors. Features contributing less than 1% to the model's total importance were removed, leading to a reduced yet highly informative subset of variables. This feature selection step simplified the model while maintaining comparable performance, enhancing both interpretability and efficiency.

Finally, to assess the soundness of the manual modeling workflow, an **AutoML benchmark** was executed using the **FLAML** framework. FLAML automatically selected and tuned algorithms within a limited computational budget, serving as an external reference for model quality. The comparison between the AutoML outcome and the manually optimized models validated the robustness and competitiveness of the proposed pipeline, confirming that a structured manual approach can achieve performance comparable to automated systems while maintaining higher transparency and control.

4. Results

The results of the predictive modeling experiments are presented in this section. The evaluation focuses on model performance, the impact of outlier treatment, and the effectiveness of optimization and feature selection. All results were computed using the same preprocessing pipeline and

consistent evaluation metrics: the coefficient of determination (R^2), the Mean Absolute Error (MAE), and the Root Mean Squared Error (RMSE).

4.1. Baseline Models — Train–Test Evaluation

The initial phase involved testing two baseline models, **Linear Regression** and **Random Forest**, using an 80–20 train–test split. Both models were trained on the cleaned dataset, after removal of duplicates and clearly inconsistent records.

The **Linear Regression** model explained less than one-third of the variance in *Profit*, confirming that linear relationships alone were insufficient to capture the complexity of the target. Conversely, the **Random Forest** achieved much higher accuracy but exhibited a notable gap between training and test performance, indicating moderate overfitting due to its flexibility and deep tree structure. Subsequent regularization through parameters such as `max_depth` and `min_samples_leaf` improved generalization, yielding a more balanced bias–variance trade-off.

These early results established the groundwork for further investigation into data irregularities and the potential influence of outliers on model behavior.

4.2. Outlier Analysis and Impact on Performance

Visual inspection of *Profit* versus *Sales* revealed extreme cases with profits or losses far exceeding the general trend, often linked to heavy discounts. To assess their impact, transactions outside the 1st and 99th percentiles of *Profit* were removed, forming the **"Normal Orders"** dataset.

Retraining the models on this refined dataset produced a substantial improvement in performance for both linear and ensemble models. Linear Regression became considerably more stable, while Random Forest achieved an R^2 exceeding 0.82 on the test data. This confirmed that approximately half of the predictive error in the initial experiments stemmed from outlier distortion, and that the remaining variance was primarily due to non-linear effects beyond the scope of linear models.

In summary, the outlier treatment proved essential for achieving reliable generalization and realistic model evaluations.

4.3. Cross-Validation and Model Comparison

Once the dataset was stabilized, a **5-Fold Cross-Validation** was conducted to compare the performance of several algorithms: **Linear Regression**, **Random Forest Regressor**, **XGBoost**, and **LightGBM**. This approach ensured that results reflected the models' generalization capacity rather than random variation from a single data split.

Models capable of capturing non-linear interactions significantly outperformed the linear baseline. Both XGBoost and Random Forest achieved average R^2 values above 0.82 with low error magnitudes, confirming the suitability of tree-based ensemble methods for this regression problem. LightGBM followed closely, while Linear Regression remained limited by its inability to model complex dependencies.

TABLE 1. AVERAGE CROSS-VALIDATED RESULTS.

Model	R^2	RMSE	MAE
Linear Regression	0.546	49.16	28.30
Random Forest	0.821	30.75	12.34
XGBoost	**0.823**	**30.67**	**12.77**
LightGBM	0.817	31.13	12.99

4.4. Hyperparameter Optimization and Feature Selection

Building upon the cross-validation results, **RandomizedSearchCV** was applied to the **Random Forest model** to optimize its key hyperparameters. The search, performed over 25 random configurations and evaluated with 5-Fold CV, identified a model that balanced expressiveness and generalization, achieving an average cross-validated $R^2 \approx 0.82$.

To visually assess the predictive reliability of the optimized Random Forest model, Figure 2 plots predicted profit values against the corresponding actual profits. The strong alignment along the diagonal indicates high accuracy and minimal systematic bias.

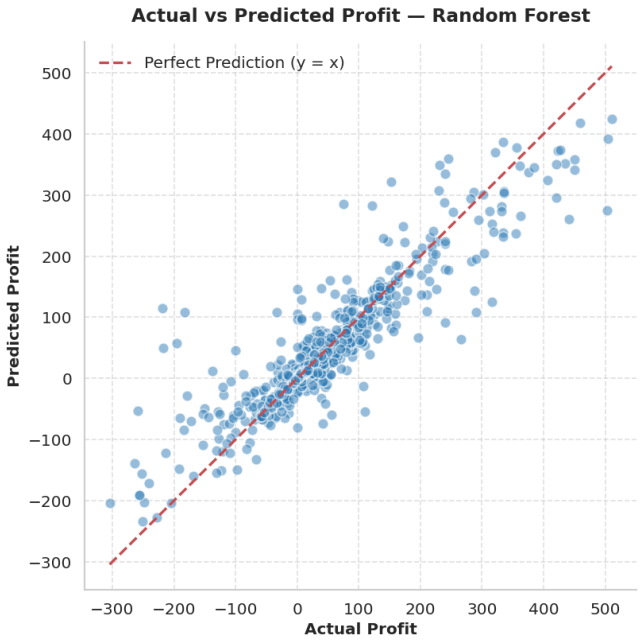

Figure 2. Predicted vs Actual Profit values for the optimized Random Forest model.

Feature importance analysis of the optimized Random Forest revealed that only a limited number of variables contributed meaningfully to the predictive power. By selecting features with relative importance $\geq 1\%$, the dataset was

reduced from 46 encoded features to 11 without notable degradation in accuracy.

TABLE 2. RANDOM FOREST PERFORMANCE WITH ALL VS. SELECTED FEATURES.

Model Variant	R^2	RMSE	MAE
All Features (46)	**0.829**	**30.32**	**11.90**
Selected Features (11)	0.814	31.59	13.25

The minimal drop in performance ($\Delta R^2 = -0.015$) demonstrated that most predictive information was concentrated in a small, interpretable subset of variables—primarily *Sales*, *Discount*, *Discount Range*, *Quantity*, and *Shipping Days*. This confirmed that feature selection improved model simplicity and interpretability without sacrificing predictive quality.

The interpretability of the optimized Random Forest model can be further examined through its feature importance scores. These quantify the relative contribution of each variable to the overall prediction of profit.

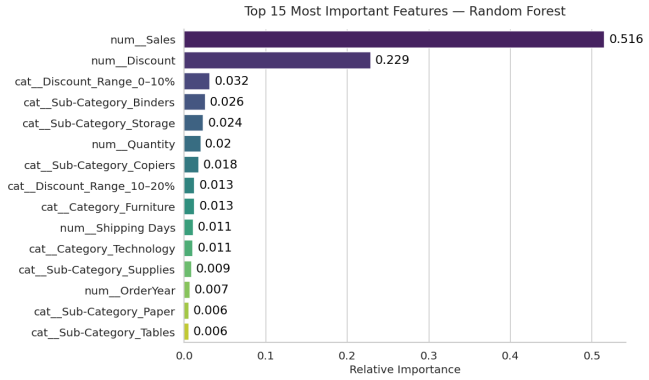

Figure 3. Feature importance for the optimized Random Forest model.

4.5. AutoML Benchmark

To validate the manual modeling pipeline, an **AutoML experiment** was conducted using the **FLAML** framework under a 300-second time budget. FLAML automatically identified **LightGBM** as the best-performing algorithm, achieving results comparable to the optimized Random Forest.

TABLE 3. COMPARISON BETWEEN OPTIMIZED RANDOM FOREST AND AUTOML MODEL.

Model	R^2	RMSE	MAE
Optimized Random Forest	**0.829**	**30.32**	**11.90**
AutoML (LightGBM)	0.827	30.48	12.89

The near-equivalent results confirmed that the manually designed pipeline—featuring systematic preprocessing, outlier control, and hyperparameter tuning—was already performing at the level of automated optimization tools, while maintaining greater transparency and interpretability.

4.6. Key Insights

Beyond predictive accuracy, the modeling process produced several actionable insights:

- **Discount Policy:** Profit remained positive for discounts up to roughly 20%, while higher discounts consistently led to losses, confirming the non-linear "discount cliff" observed during EDA.
- **Operational Efficiency:** Lower *Shipping Days* correlated with higher profit margins, suggesting that delivery performance directly impacts profitability.
- **Seasonal Effects:** Sales during promotional seasons (e.g., Black Friday) displayed greater variance in profit outcomes, underscoring the importance of timing in promotional strategies.

These insights illustrate how interpretable machine learning models can inform both predictive forecasting and strategic business decision-making.

5. Conclusions

This study applied machine learning techniques to predict transaction-level profitability using the **Superstore Sales** dataset, following the **CRISP-DM** methodology. The proposed workflow integrated exploratory data analysis, feature engineering, outlier detection, and model optimization to build an accurate and interpretable predictive system.

Initial experiments showed that **linear models** were insufficient to capture the complex and non-linear relationships affecting profit, achieving limited explanatory power. In contrast, **tree-based ensemble models**—including Random Forest, XGBoost, and LightGBM—achieved significantly higher predictive accuracy, confirming the non-linear structure of the underlying business processes.

The treatment of **outliers** proved to be a decisive step. By removing extreme profit values, model performance improved markedly, particularly in terms of stability and generalization. Subsequent **hyperparameter tuning** and **feature selection** further enhanced efficiency, reducing the number of features from 46 to 11 while maintaining nearly identical predictive accuracy. The optimized Random Forest achieved a test R^2 of approximately 0.83, with low MAE and RMSE values, confirming its robustness.

A final **AutoML benchmark** using the FLAML framework validated the effectiveness of the manual pipeline. Although FLAML identified LightGBM as the best model, its performance was nearly identical to that of the optimized Random Forest, demonstrating that a structured manual approach can match automated optimization while preserving interpretability and control.

From a business perspective, the analysis yielded several actionable insights. Profitability remained positive up to discounts of about 20%, but larger discounts consistently led to losses. Shorter delivery times correlated with higher profit, highlighting the influence of logistical efficiency, while profit variance during promotional periods suggested

that timing and discount strategies are critical for maintaining profitability.

Overall, the project demonstrates how **interpretable ensemble learning**—combined with data cleaning, feature engineering, and systematic validation—can effectively support both accurate profit forecasting and strategic decision-making in retail operations. Future work could explore incorporating **external features** such as market trends or customer sentiment, as well as adopting **time-series forecasting** and **explainable AI (XAI)** techniques to enhance interpretability and decision support.