BI2025 Experiment Report - Group 008

Milica Aleksic*
TU Wien
Austria

Vidak Grujic†
TU Wien
Austria

Abstract

This report documents the machine learning experiment for Group 008, following the CRISP-DM process model. GitHug repository with code and data is available at: https://github.com/milialeksic/BI_ass3

CCS Concepts

• **Computing methodologies** → **Machine learning**.

Keywords

CRISP-DM, Provenance, Knowledge Graph, Machine Learning

ACM Reference Format:
Milica Aleksic and Vidak Grujic. 2025. BI2025 Experiment Report - Group 008. In *Proceedings of Business Intelligence (BI 2025).* ACM, New York, NY, USA, 5 pages. https://doi.org/https://doi.org/10.5281/zenodo.17955246

1 Business Understanding

1.1 Data Source and Scenario

The dataset is the Supermarket Sales dataset. It contains transaction-level data including branch, customer type, gender, product line, payment method, unit price, quantity, tax, total sales, date, time, and a customer satisfaction rating (4 to 10). Scenario: The supermarket chain wants to monitor customer satisfaction and predict the rating a customer is likely to give. The business particularly wants to detect low predicted ratings early to improve service quality and customer experience.

1.2 Business Objectives

The primary business objectives are: 1. Understand which factors affects customer satisfaction. 2. Predict customer ratings based on available information. 3. Detect scenarios likely to produce low ratings and support decisions to improve service.

1.3 Business Success Criteria

The project will be considered a business success if: 1. The model predicts customer ratings with useful accuracy (e.g., MAE sufficiently low to distinguish low vs. high satisfaction cases). 2. The system can identify transactions likely to result in low ratings, which would enable proactive service improvements. 3. The analysis provides useful insights into factors associated with customer satisfaction.

1.4 Data Mining Goals

The data mining goals are: 1. Train a regression model to predict the continuous customer rating variable. 2. Understand feature importance to identify which attributes most influence satisfaction. 3. Optionally compare multiple ML models. 4. Ensure full reproducibility of all experiments.

1.5 Data Mining Success Criteria

Data mining will be considered successful if: 1. The model achieves meaningful and realistic MAE threshold (*e.g.*, 1.0). 2. The model is interpretable, allowing business users to understand what determines the low satisfaction. 3. The entire workflow is reproducible and computationally efficient.

1.6 AI Risk Aspects

Relevant AI risks include: 1. Customer type (Member/Normal) could result in unfair treatment if misused. 2. Using model predictions to treat customers differently could be unethical or discriminatory. 3. Predictions must be used only as decision support and not as a mechanism for unequal service.

2 Data Understanding

Dataset Description: Transaction-level supermarket sales data including customer attributes, product information, sales amounts, timestamps, and customer ratings.

The following features were identified in the dataset:

2.1 Data Understanding – Outlier Detection

Potential outliers were identified using Z-score based approach for numerical attributes. Treshold of 2.2 is used as cut-off value. The analysis shows that outliers were not detected for Unit price, Quantity, or Rating.

Outliers were identified for Tax 5%, Sales, cogs, and gross income. These outliers occur at the same transaction indices across these attributes, which is expected because these monetary variables are deterministically related (Sales = cogs + Tax 5% and gross income is proportional to cogs).

The detected outliers correspond to unusually big purchase rather than error values. This indicates the presence of high-value transactions rather than data quality issues.

2.2 Data Understanding – Outlier Assessment

The outlier report was inspected to check the authenticity of the detected values. The identified outliers occur only in monetary attributes and correspond to large but realistic purchase amounts.

*Student A, Matr.Nr.: 12424821
†Student B, Matr.Nr.: 12332263

BI 2025, -

ACM ISBN 978-x-xxxx-xxxx-x/YYYY/MM
https://doi.org/https://doi.org/10.5281/zenodo.17955246

Table 1: Raw Data Features

Feature Name	Data Type	Description
Branch	string>	Supermarket branch where the transaction occurred.
City	string>	City where the branch is located.
Customer type	string>	Customer classification: Member or Normal.
Date	date>	Date of transaction.
Gender	string>	Gender of the customer.
Invoice ID	string>	Unique identifier for each transaction.
Payment	string>	Payment method used in the transaction.
Product line	string>	Category of product purchased.
Quantity	integer>	Number of items purchased.
Rating	float>	Customer satisfaction rating on a scale from 4 to 10.
Sales	float>	Total amount paid including tax.
Tax 5%	float>	5 percent tax applied to the purchase.
Time	string>	Time of transaction.
Unit price	float>	Price per unit of product.
cogs	float>	Cost of goods sold (pretax).
gross income	float>	Gross income from the transaction.
gross margin percentage	float>	Gross margin percentage (constant in dataset).

Because these values represent valid high-spending transactions and no outliers were detected for the target variable (Rating), no records are removed during the Data Understanding phase. The presence of these values reflects natural variability in customer spending rather than data errors.

Any potential handling of extreme values (e.g. capping or transformation) is moved to the Data Preparation phase if required by the modeling approach.

Outlier Decision: No outliers removed during Data Understanding; decision deferred to Data Preparation.

2.3 Statistical Properties and Correlations

Statistical properties and correlations were computed for all numeric attributes of Supermarket Sales dataset. The dataset contains 1,000 complete records with no missing values.

Descriptive statistics show that unit prices range is from approximately 10 to 100, with a mean = 55.7. Customers purchase on usually 5 to 6 items per transaction, with quantities which have range from 1 to 10. Sales amounts and cost-related attributes The gross margin percentage is constant across all records (4.76%), which results in zero variance.

Customer ratings range from 4 to 10, with mean approx. 6.97 and standard deviation of 1.72, which indicates moderate variability in customer satisfaction.

Pearson correlation analysis with regression target 'Rating' reveals that none of numeric attributes have a strong linear relationship with target variable. All observed correlations are close to zero ($|r| < 0.04$). Money related attributes such as Sales, cogs, Tax 5%, and gross income show similar correlations with target, which confirms their relationships. The constant gross margin percentage shows no defined correlation.

These results indicate that customer satisfaction is not primarily driven by purchase volume or price-related factors only. Instead, categorical attributes, for example, branch, product line, payment method and potential non-linear effects are likely to play a more important role in predicting customer ratings.

Variable	Correlation
Rating	1.000000
Unit price	-0.008778
Quantity	-0.015815
cogs	-0.036442
Tax 5%	-0.036442
gross income	-0.036442
Sales	-0.036442
gross margin percentage	nan

Table 2: Correlation of numerical attributes with Rating

Variable	count	mean	std	median
Unit price	1000.0	55.672130	26.494628	55.230000
Quantity	1000.0	5.510000	2.923431	5.000000
Tax 5%	1000.0	15.379369	11.708825	12.088000
Sales	1000.0	322.966749	245.885335	253.848000
cogs	1000.0	307.587380	234.176510	241.760000
gross margin percentage	1000.0	4.761905	0.000000	4.761905
gross income	1000.0	15.379369	11.708825	12.088000
Rating	1000.0	6.972700	1.718580	7.000000

Table 3: Count, mean, standard deviation, and median of numerical attributes

Variable	min	25%	75%	max
Unit price	10.080000	32.875000	77.935000	99.960000
Quantity	1.000000	3.000000	8.000000	10.000000
Tax 5%	0.508500	5.924875	22.445250	49.650000
Sales	10.678500	124.422375	471.350250	1042.650000
cogs	10.170000	118.497500	448.905000	993.000000
gross margin percentage	4.761905	4.761905	4.761905	4.761905
gross income	0.508500	5.924875	22.445250	49.650000
Rating	4.000000	5.500000	8.500000	10.000000

Table 4: Minimum, quartiles, and maximum of numerical attributes

2.4 Distribution and Skewness Check

Skewness was computed for all numeric attributes. Monetary variables (Sales, cogs, Tax 5%, gross income) show positive skewness (approx. 0.89), which indicates right-skewed distributions which were caused by small number of high-value transactions.

Unit price, Quantity, and the target variable (Rating) shows very low skewness (close to zero), which indicates approximately symmetric distributions.

The observed skewness patterns are typical for retail transaction data and do not indicate data quality issues.

Calculated skewness values: { "Unit price": 0.007077447853328846, "Quantity": 0.012941048017172435, "Tax 5%": 0.8925698049581423, "Sales": 0.8925698049581418, "cogs": 0.8925698049581418, "gross income": 0.8925698049581423, "Rating": 0.00900964876573073 }

Variable	Skewness
Tax 5%	0.892570
gross income	0.892570
Sales	0.892570
cogs	0.892570
Quantity	0.012941
Rating	0.009010
Unit price	0.007077

Table 5: Calculated skewness values of numerical attributes

2.5 Plausibility of Values Check

Plausibility check was done based on minimum, maximum, and median values. It shows that all numeric attributes are located within realistic ranges for supermarket transactions. Unit prices range from about 10 to 100, quantities from 1 to 10 items, and total sales from approximately 11 to 1,043. Customer ratings range from 4 to 10.

The gross margin percentage is constant across all records, which explains its zero variance. No implausible or erroneous values (e.g., negative prices or quantities) were detected. Plausibility summary: { "min": { "Unit price": 10.08, "Quantity": 1.0, "Tax 5%": 0.5085, "Sales": 10.6785, "cogs": 10.17, "gross margin percentage": 4.761904762, "gross income": 0.5085, "Rating": 4.0 }, "max": { "Unit price": 99.96, "Quantity": 10.0, "Tax 5%": 49.65, "Sales": 1042.65, "cogs": 993.0, "gross margin percentage": 4.761904762, "gross income": 49.65, "Rating": 10.0 }, "median": { "Unit price": 55.230000000000004, "Quantity": 5.0, "Tax 5%": 12.088000000000001, "Sales": 253.848, "cogs": 241.76, "gross margin percentage": 4.761904762, "gross income": 12.088000000000001, "Rating": 7.0 } }

2.6 Visual Exploration and Hypothesis Generation

Visual exploration shows that customer ratings are evenly distributed between 4 and 10, with no extreme concentration at single value. This indicates balanced target variable suitable for regression modeling.

The boxplot of Rating by Product Line reveals small but noticeable differences in median ratings across product categories, while overall spread of ratings is similar. This suggests that product line may have influence on customer satisfaction, but no single category dominates ratings completely.

Variable	min	max	median
Unit price	10.080000	99.960000	55.230000
Quantity	1.000000	10.000000	5.000000
Tax 5%	0.508500	49.650000	12.088000
Sales	10.678500	1042.650000	253.848000
cogs	10.170000	993.000000	241.760000
gross margin percentage	4.761905	4.761905	4.761905
gross income	0.508500	49.650000	12.088000
Rating	4.000000	10.000000	7.000000

Table 6: Minimum, maximum, and median values of numerical attributes

These observations support inclusion of categorical features, such as product line, in the predictive model. Visual summary: { "figures_generated": 2, "description": "Histogram of Rating and boxplot of Rating by Product Line.", "rating_summary": { "count": 1000.0, "mean": 6.9727, "std": 1.7185802943791215, "min": 4.0, "25%": 5.5, "50%": 7.0, "75%": 8.5, "max": 10.0 }, "mean_rating_by_product_line": { "Electronic accessories": 6.924705882352941, "Fashion accessories": 7.0292134831460675, "Food and beverages": 7.113218390804598, "Health and beauty": 7.003289473684211, "Home and lifestyle": 6.8375, "Sports and travel": 6.916265060240963 } }

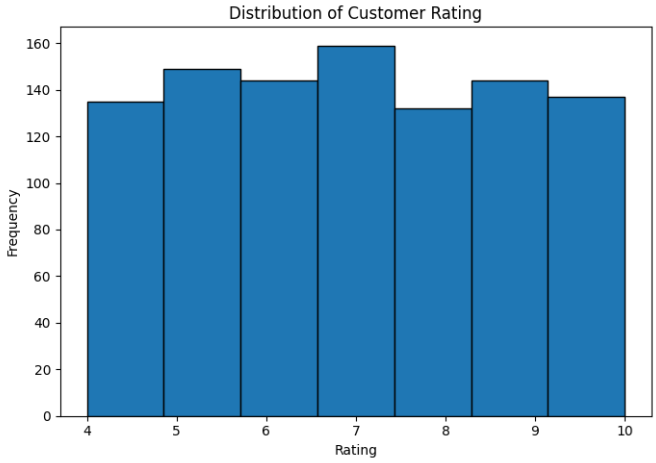

Figure 1: Customer Rating by Product Line

2.7 Bias Evaluation Logging

Ethical risks: Ethically sensitive attributes are present in Supermarket Sales dataset, potentially Gender and Customer type (Member vs. Normal). There is high chance that thhese attributes could potentially introduce bias in the model if model predictions are used to treat customers in different way based on predicted satisfaction.

Unbalanced Distributions and Bias Risk: Target variable (Rating) is well distributed across its range (4–10), and no extremes of low or high ratings was observed. Therefore, we assume that there is no bias which arises from target imbalance.

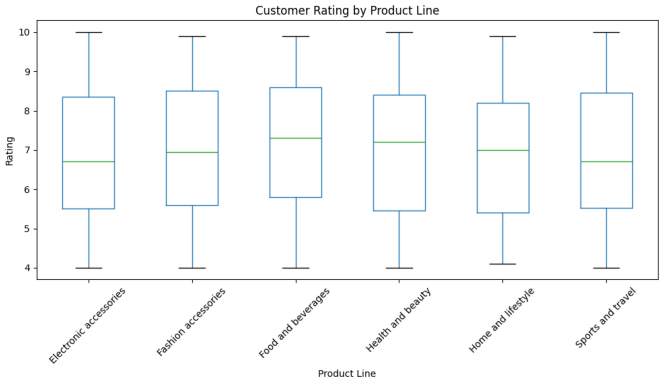

Figure 2: Boxplot of Rating by Product Line

However, the usage of predicted ratings to prioritize service, offers, or attention for specific customer groups could lead to unfair treatment. To overcome this risk, sensitive attributes should be handled carefully and used only for analysis, not for decision making.

Model evaluation should focus on overall performance (e.g., MAE) and include subgroup analysis (e.g., by customer type or gender) to ensure consistent prediction quality across different customer groups.

2.8 Risks and Expert Questions Logging

Potential Risks and Bias: 1. Behavioral Noise: Customer ratings are subjective and can be influenced by not constant factors such as mood, time pressure, or individual expectations. This can introduce noise which cannot be fully explained by transaction data alone.

2. Omitted Variables: Dataset does not detect important contextual factors such as staff behavior, waiting time, promotions, or how much the store is crowded, which may significantly influence customer satisfaction.

3. Usage Bias: If predicted ratings are used to prioritize service or offers, this could lead to unfair treatment for certain customer groups (e.g., Normal vs. Member customers).

Questions for an External Expert: 1. Rating Process: Under what conditions are ratings collected, and how consistent is the rating behavior across customers? 2. Operational Factors: Are there store-level or operational variables (e.g., staffing, queue length, time of day effects) that could be added to better explain customer satisfaction? 3. Ethical Use: What should be applied to ensure predicted ratings are used to improve service quality rather than to discriminate between customer groups?

2.9 Required Data Preparation Actions Logging

Based on the Data Understanding analysis, following actions are required in Data Preparation phase (Section 3):

1. Feature Encoding: Categorical attributes such as Branch, City, Customer type, Gender, Product line, and Payment must be encoded so they can be usable by regression models.

2. Feature Scaling (Recommended): Numerical attributes (e.g., Unit price, Quantity, Sales, cogs, gross income) should be scaled (e.g., using StandardScaler) to ensure they could be comparable across features, especially for models sensitive to feature scale.

3. Feature Selection / Redundancy Handling: Strongly related monetary attributes (Sales, cogs, Tax 5%, gross income) should be reviewed to avoid multicollinearity. One or more redundant variables may be removed or combined.

4. Constant Feature Removal: The attribute 'gross margin percentage' is constant and should be removed because it provides no predictive information.

5. Outlier Handling (Optional): Detected outliers represent valid high-value transactions and do not require removal. Robust transformations may be considered if required by specific models.

6. Extract simple date features: From the 'Date' attribute, extract simple features such as day of week or month. This is done because regression models cannot directly interpret date values.

7. Target Variable Preparation: The target variable (Rating) requires no transformation and can be used directly for regression modeling.

3 Data Preparation

3.1 Data Cleaning

The Data Preparation phase transforms the raw supermarket transaction data into a structured and fully numeric dataset suitable for regression modeling. All preprocessing steps described in this section are directly derived from the provenance graph by querying activities that are part of the Data Preparation phase (`sc:isPartOf :data_preparation_phase`).

Documented preprocessing steps. The following preprocessing activities were executed and logged in the provenance graph as individual `prov:Activity` instances, together with their associated comments and execution timestamps:

- Outlier handling was executed based on the Data Understanding decision. The detected outliers correspond to valid high-value supermarket transactions and were therefore retained. No rows were removed in this step.
- Feature selection was performed based on Data Understanding results. Non-predictive identifiers such ar Invoice ID, constant attributes which refers to "gross margin percentage", and redundant monetary features like "Tax 5%" and "cogs" were removed to reduce dimensionality and similarly corelated variables.
- Raw Date and Time attributes were transformed into interpretable temporal features (month, day of week, hour and minute) to capture seasonal, weekly, and intra-day effects. The original Date and Time columns were removed to avoid incorrect numerical interpretation.
- Categorical attributes were transformed using One-Hot Encoding. The reason for it is to avoid introducing artificial ordering or distance assumptions, because this kind of categorical variables does not have ordinal relationship.This tries to ensure correct interpretation by regression models and improves model robustness. One-hot encoding was applied with setting parameter to True, which results in implicit reference categories for each categorical variable. For example, Gender_Female and CustomerType_Member are

represented as baseline (value 0) and they are not explicitly stored as separate variables.
- Numerical features were standardized using StandardScaler. This is done in order to to ensure comparable feature scales and stable regression behavior. One-hot encoded categorical features and the target variable were excluded from scaling.
- 3b Preprocessing Steps Considered but Not Applied
- 3c Derived Attributes
- 3d External Data Sources and Attributes

3.1.1 3b Preprocessing Steps Considered but Not Applied. Several preprocessing options were evaluated but deliberately not applied. 1. Outlier removal was considered based on statistical calculation. However, identified outliers correspond to valid high-value transactions and were retained. 2. Label (ordinal) encoding of categorical attributes was rejected to avoid introducing artificial ordering into nominal variables. 3. Binning of the target variable (Rating) was not applied, because analysis is formulated as regression task and binning would reduce information granularity. 4. The raw Date and Time attributes were not used directly, because regression models cannot meaningfully interpret timestamps without transformation.

3.1.2 3c Derived Attributes. Derived attributes were introduced to improve interpretability and predictive performance. 1. Nominal categorical variables were transformed using one-hot encoding to avoid artificial ordering assumptions. 2. Temporal information was extracted from Date and Time attributes by in a way that month, day of week, hour and minute features were derived to capture seasonal, weekly, and intra-day effects. More complex derived attributes were considered but not applied to avoid not necessary feature expansion.

3.1.3 3d External Data Sources and Attributes. Additional external data sources could potentially improve the prediction of customer ratings. Examples include promotional calendars, public holidays, local events, or regional economic indicators. This type of data could help to explain contextual and temporal variations in customer satisfaction. These data sources were not integrated because of limited availability.

Each activity explicitly records its input and output datasets using provenance relations such as `prov:used`, `prov:wasGeneratedBy`, and `prov:wasDerivedFrom`, ensuring full traceability of all data transformations.

Final prepared dataset. The outcome of the Data Preparation phase is the dataset `:prepared_data` (label: *Prepared Dataset for Modeling*), which is documented in the provenance graph as a `prov:Entity`. The dataset description retrieved from the graph is as follows:

> This dataset represents final output of the Data Preparation phase. It includes selected numerical features, one-hot encoded categorical variables, derived temporal attributes (month, day of week, hour, minute), and standardized numerical values. This dataset is fully numeric, reproducible, and ready to be used as input for regression modeling.

4 Modeling

4.1 Hyperparameter Configuration

The model was trained using the following hyperparameter settings:

Table 7: Hyperparameter Settings

Parameter	Description	Value
Learning Rate	...	1.23

4.2 Training Run

A training run was executed with the following characteristics:

- **Algorithm:** Random Forest Algorithm
- **Start Time:** 2025-12-16 17:45:16
- **End Time:** 2025-12-16 17:45:16
- **Result:** R-squared Score = 1.2300

5 Evaluation

6 Deployment

7 Conclusion