# BI2025 Experiment Report - Group 05

Ana Zrnic[*]
TU Wien
Austria

Daria Alekseienkova[†]
TU Wien
Austria

## Abstract

This report documents the machine learning experiment for Group 05, following the CRISP-DM process model.

## CCS Concepts

• **Computing methodologies → Machine learning**.

## Keywords

CRISP-DM, Provenance, Knowledge Graph, Machine Learning

**ACM Reference Format:**
Ana Zrnic and Daria Alekseienkova. 2025. BI2025 Experiment Report - Group 05. In *Proceedings of Business Intelligence (BI 2025).* ACM, New York, NY, USA, 11 pages. https://doi.org/XXXXXXX.XXXXXXX

## 1 Business Understanding

This section introduces the business context, objectives, and success criteria that motivate the data analytics task.

### 1.1 Data Source and Scenario

This dataset was taken from the website Kaggle. It is a dataset that contains transactional and demographic information about customers of Ifood, a Brazilian online ordering and delivery platform. The dataset originally comes from a public GitHub repository when the Ifood Brain team had a data challenge for the data analyst role hiring process. The company wants to understand the spending behavior of its customers to improve effectiveness of marketing campaigns with historical customer interactions and their past spending patterns. In order to achieve this, we are building a data-driven approach using data mining techniques to analyze customer data and identify patterns and trends in their spending behavior.

### 1.2 Business Objectives

The primary business objective of the customer is to predict how much a customer is likely to spend, in order to improve marketing strategies. This allows for better targeting, budget allocation and campaign planning. Key business questions:

- Which variables (behavioral or demographic) are more significant in predicting customer spending?
- How much is a customer expected to spend based on their profile and past interactions?

### 1.3 Business Success Criteria

Business success is achieved if the implemented prediction model improves the marketing team's effectiveness in targeting customers and allocating budgets.

**Measurable criteria:**

- Identify the top 20% of customers who are the highest spenders before launching a new campaign.
- Increase revenue by 6% in the next quarter compared to the previous quarter by targeting high-spending customers.
- Reduce marketing costs by at least 13% on customers who are inactive.

**Subjective criteria:**

- The model's insights and predictions are positively received and understandable by the marketing team and marketing managers.
- The model is stable and doesn't drastically change within smaller time frames, which the stakeholders would find reassuring.

### 1.4 Data Mining Goals

To support the business objectives, a regression model will be developed to predict customer spending, and to identify which features most strongly influence that spending. The outputs of the data mining process will include:

- Predicted spending value per customer
- A ranked list of significant features influencing customer spending

What is not part of our data mining goals:

- Not predicting campaign success probability

### 1.5 Data Mining Success Criteria

A successful outcome requires that the regression model achieves acceptable predictive performance, as measured by the following criteria:

- The coefficient of determination ($R^2$) is at least 0.75, indicating that a significant amount of the variance is explained by the input features.
- The Mean Absolute Error (MAE) value is below 15% of the average spending value.
- The Root Mean Squared Error (RMSE) value is below 10% of the average spending value.

[*] Student A, Matr.Nr.: 52253331
[†] Student B, Matr.Nr.: 12426894

*BI 2025, -*

ACM ISBN 978-x-xxxx-xxxx-x/YYYY/MM
https://doi.org/XXXXXXX.XXXXXXX

## 2 Data Understanding

In this section, the dataset is explored to assess its structure, attribute semantics, statistical properties, and data quality in order to inform subsequent data preparation and modeling steps.

### 2.1 Attribute Types

Analysis of attribute types in the iFood marketing dataset. Each attribute was classified as: nominal, ordinal, interval or ratio based on the nature of the data:

- **Nominal:** Categorical variables without inherent order (e.g., ID, Marital_Status)
- **Ordinal:** Categorical variables with meaningful order (e.g., Education)
- **Interval:** Numeric variables with meaningful differences but no true zero (e.g., Year_Birth, Dt_Customer)
- **Ratio:** Numeric variables with true zero point (e.g., Income, spending amounts, counts)
- **Nominal-binary:** Binary categorical variables (e.g., campaign acceptance flags, Complain, Response)

#### 2.1.1 *Measurement Level Distribution.*

- Ratio: 15 attributes (numeric with true zero)
- Nominal-binary: 7 attributes (binary flags)
- Nominal: 2 attributes (categorical without order)
- Interval: 2 attributes (numeric without true zero)
- Ordinal: 1 attribute (categorical with order)
- Unknown: 2 attributes

The unknown attributes Z_Revenue and Z_CostContact were found to contain constant values (3 and 11, respectively) across all observations, with no documentation or metadata provided to explain their meaning or purpose. Due to the lack of interpretability these attributes are excluded from subsequent analysis.

### 2.2 Attribute Units

Analysis of attribute units in the iFood marketing dataset. Each attribute was assigned a unit based on its semantic meaning:

- **Identifier:** ID column with no unit
- **Year:** Birth year
- **Calendar date:** Customer enrollment date
- **Days:** Time since last purchase (recency)
- **Monetary amount (BRL):** All spending columns (wines, fruits, meat, fish, sweets, premium products)
- **Count:** Number of purchases, deals, web visits
- **Count of people:** Number of kids/teens at home
- **Yearly income (BRL):** Customer income
- **Binary indicator (0/1):** Campaign acceptance flags, complain, response
- **Categorical label:** Education and Marital_Status

### 2.3 Attribute Semantics

Attribute semantics were documented for all 27 features and are summarized in Table 1.

**Table 1: Attribute Semantics for the iFood Marketing Dataset**

| Attribute | Description |
|---|---|
| ID | Unique customer identifier |
| Year_Birth | Customer year of birth |
| Education | Highest education level |
| Marital_Status | Marital status |
| Income | Annual household income (BRL) |
| Kidhome | Number of children in household |
| Teenhome | Number of teenagers in household |
| Dt_Customer | Customer enrollment date |
| Recency | Days since last purchase |
| MntXXX | Amount spent per product category (wine, fruit, meat, fish, sweets, gold) in the last two years (BRL) |
| NumXXXPurchases | Purchases per channel (web, catalog, store, discount) |
| NumWebVisitsMonth | Website visits in the last month |
| AcceptedCmp1–5 | Acceptance of campaigns 1–5 (0 if no, otherwise yes) |
| Complain | Complaint in the last two years (0 if no, otherwise yes) |
| Response | Acceptance of most recent campaign (Cmp6) (0 if no, otherwise yes) |

### 2.4 Basic Statistics

Computed basic descriptive statistics for 16 numeric attributes in the iFood marketing dataset:

- Central tendency: mean, median, mode
- Dispersion: variance, standard deviation, min, max
- Shape: skewness

During the basic statistics analysis a few key observations were made:

- In Year_Birth the presence of the minimal value being 1893 suggests potential outliers or data quality issues that should be verified.
- The median income is around 51,000 BRL with the mean being slightly higher, this along with the very large maximum value and strong positive skewness (6.76) indicates the presence of extreme high-income outliers.
- Household composition variables (Kidhome and Teenhome) are dominated by zeros, as reflected by medians of zero and relatively low means. This indicates that many customers do not have children or teenagers living at home. Customers without children being predominantly a large group of Ifood services is a potential point of interest.
- Among the purchase channel variables, in-store purchases show the highest mean, indicating that physical stores remain the dominant purchasing channel.
- Overall, many numeric attributes exhibit strong positive skewness, highlighting the need for appropriate preprocessing steps such as transformation or outlier handling.

### 2.5 Correlation Analysis

A correlation heatmap (Figure 1) was generated to visualize relationships between numeric variables. Key observations:

- A strong positive correlation is observed between Income and most spending-related attributes (MntWines, MntMeatProducts, MntFishProducts, etc.). This indicates that higher-income customers tend to spend more across multiple product categories and purchasing channels
- Income shows a strong negative correlation (-0.55) with *NumWebVisitsMonth*, suggesting that customers with lower

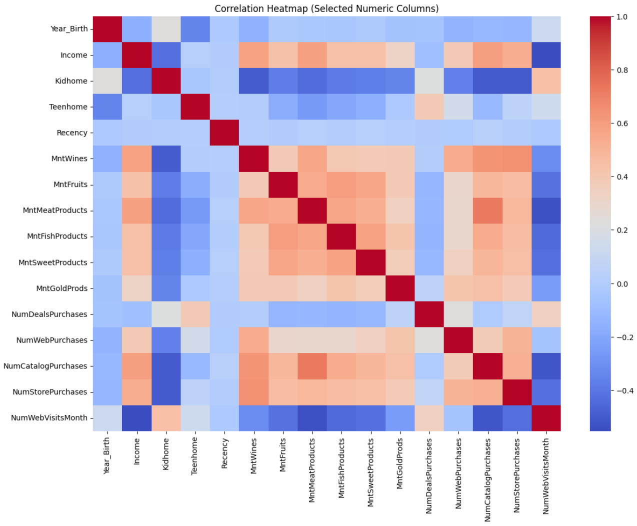

**Figure 1: Correlation heatmap**

income levels tend to browse the website more frequently without necessarily converting these visits into purchases.

- Both *Kidhome* and *Teenhome* are negatively correlated with most monetary attributes and purchase counts, implying that customers with children or teenagers tend to spend less overall.
- *Recency* attribute shows near-zero correlations with most other variables. This indicates that recency cannot be directly explained by the other attributes.

## 2.6 Histogram Analysis

Each generated histogram (Figure 2) includes automated warnings for potential:

- Extreme outliers (|$z$-score| > 4)
- Impossible year values (outside 1925-2025 range)
- Negative income values

Key observations:

- *Year_Birth* has the presence of three extreme values that are inconsistent with realistic customer ages. This confirms the existence of clear data quality issues or encoding errors.
- *Income* displays a highly right-skewed distribution along with eight extreme values far exceeding the majority of observations, reinforcing earlier findings of substantial income outliers.
- *Kidhome* and *Teenhome* are highly imbalanced, which may influence their predictive contribution.
- The detection of extreme outliers across variables indicates that the behavior is a systematic property of the dataset rather than isolated anomalies. While these values may reflect genuine high-value customers, their magnitude suggests that robust scaling or transformation techniques should be considered.

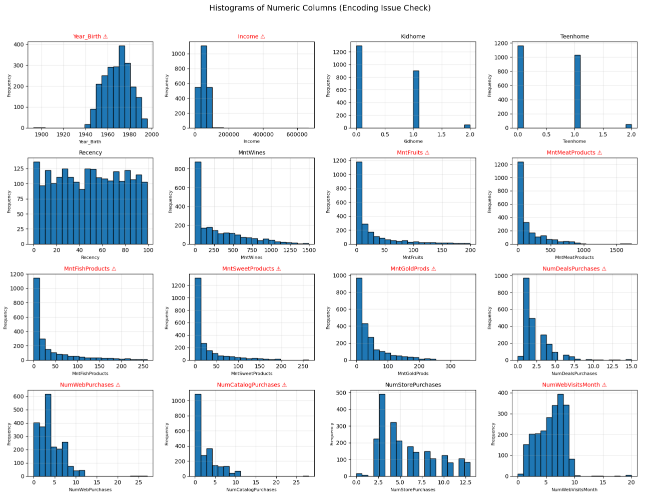

**Figure 2: Histograms of Numeric Attributes**

## 2.7 Categorical Variables

Generated countplots for categorical and binary attributes can be observed in Figure 3. This analysis helps identify class imbalances and category distributions.

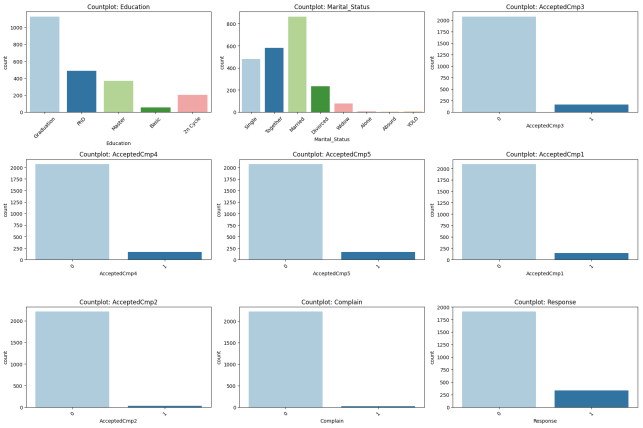

**Figure 3: Countplot of Categorical-like Attributes**

Furthermore, categorical and binary attributes were checked for the following possible encoding issues:

- Placeholder-like values (unknown, null, na, n/a, ?, -): 0 cells
- Inconsistent capitalization across categories: 0 cells
- Rare categories with frequency < 1%:
  - *Complain*: ['1']
  - *Marital_Status*:['Alone', 'Absurd', 'YOLO']
- Mixed numeric and text categories: 0 cells
- Highly imbalanced binary variables (class frequency < 20%): all binary attributes

Additional observations:

- The *Education* attribute exhibits a skewed but overall plausible distribution.

- The *Marital_Status* attribute contains a small number of rare categories with very low frequencies. These categories may require further justification, or removal during the data preparation phase.
- All campaign-related attributes (*AcceptedCmp1–5* and *Response*) exhibit strong class imbalance, with acceptance rates well below 20%. While this reflects typical marketing response behavior, it poses challenges for predictive modeling and must be explicitly addressed in subsequent analysis.
- Similarly, the *Complain* attribute is highly imbalanced, with complaints representing a very rare event (21 cases, approximately 0.009%). Without additional contextual information regarding the nature of the complaints, this attribute is unlikely to provide meaningful predictive value.

## 2.8 Missing Values Analysis

24 missing values were found in the Income column out of 2240 total records (1.07% missing rate). All other columns have complete data with no missing values.

## 2.9 Bias and Risk Assessment

The iFood marketing dataset does not contain explicitly sensitive personal attributes such as race, religion, political affiliation, or health-related information. The *Income* attribute is an example of a socioeconomically sensitive variable. Although not inherently protected, income can act as a proxy for social class and purchasing power. Models trained using this attribute may therefore risk favoring higher-income customers, leading to biased outcomes such as preferential targeting or exclusion of lower-income groups.

Potential risks and sources of bias cannot be fully assessed based on the available data alone. Most notably, the dataset lacks information on how customers were selected, whether any filtering or pre-processing was applied prior to data collection, and over which geographic regions or time periods the data were gathered. Without this context, it is unclear whether the dataset reflects the full customer base or a selectively sampled subset.

## 2.10 Data Preparation Actions

Based on the findings from the Data Understanding phase, several data preparation actions were performed to improve data quality, interpretability, and suitability for modeling. The following attributes are to be removed from the dataset:

- `Z_CostContact` and `Z_Revenue`
- `Complain`

Several transformations and derived attributes to enhance semantic clarity and reduce redundancy:

- *Year_Birth* to *Age*.
- *Dt_Customer* to *CustomerTenure*, representing the duration of the customer relationship in days.
- *Kidhome* and *Teenhome* were combined into a single attribute *TotalChildren* to capture household composition.
- Rare categories in *Marital_Status* need to be removed or justified.
- The campaign-related attributes *AcceptedCmp1–5* and *Response* were aggregated into *TotalCampaignsAccepted*, representing overall customer responsiveness to marketing campaigns.

Target variable *TotalSpending* created through aggregation of monetary attributes into a single spending metric.

Adressing outliers and handling skewed data:

- Handle unrealistic years of birth.
- Scaling and/or logarithmic transformations to highly skewed variables to reduce the influence of extreme values.

Adressing class imbalance: Aggregation of campaign responses instead of binary indicators.

Missing data handled using appropriate imputation strategies to preserve dataset size while minimizing bias.

# 3 Data Preparation

The goal of the Data Preparation phase is to transform the raw marketing data into a clean and consistent form. Following the CRISP-DM methodology, this phase involves selection of relevant attributes, addressing data quality issues, creation of derived features, application of relevant transformations and documentation of preprocessing strategies.

## 3.1 Data Selection

As part of Data Selection, it is necessary to identify which attributes are relevant for prediction of customer spending. The decision is made upon thorough analysis of the results of the Data Understanding phase, where each attribute was examined with respect to its variance, distribution, and correlation with the target variable. Attributes were retained when they demonstrated predictive potential, offered meaningful demographic or behavioral information, or were relevant for constructing derived features. On the other hand, attributes with little or no predictive value were excluded.

It was also decided to remove such variables as *ID*, *Z_CostContact*, *Z_Revenue* and *Complain* due to the absence of analytical relevance. The campaign acceptance variables were combined into *TotalCampaignsAccepted* to limit multicollinearity and retain their overall predictive contribution. Finally, the target variable *TotalSpending* was created by summing all product category spending attributes. Table 2 summarizes the selection decisions and their underlying rationale.

## 3.2 Data Cleaning

The Data Cleaning step involves resolving data quality issues identified during the Data Understanding phase. This includes handling missing values, correcting unrealistic outliers, and standardizing categorical attributes by consolidating rare or inconsistent entries.

*3.2.1 Missing Values.* Only one attribute, *Income*, contained 24 missing values (1.07% of records). Removing these observations would have reduced the sample size without providing any analytical benefit, especially since *Income* is one of the strongest predictors of customer spending. Because the distribution of *Income* is right-skewed, median imputation was applied. After imputation, all missing values were resolved while preserving the overall distribution of the variable.

**Table 2: Data Selection Decisions**

| Attribute | Decision |
|---|---|
| Income | Include |
| Year_Birth | Transform (Age) |
| Education | Include |
| Marital_Status | Include (merge rare values) |
| Kidhome, Teenhome | Combine (TotalChildren) |
| Recency | Include |
| Dt_Customer | Transform (DaysSinceEnrollment) |
| Mnt* (spending variables) | Aggregate (TotalSpending) |
| Num* (purchase frequencies) | Include |
| AcceptedCmp1–5, Response | Combine (TotalCampaignsAccepted) |
| Complain | Exclude (near-constant) |
| ID | Exclude |
| Z_CostContact, Z_Revenue | Exclude (no variance) |

*3.2.2 Income Outliers.* As illustrated in Figure 4, *Income* variable is right-skewed and contains a small number of extremely high values compared to the rest of the customer base. Using the IQR method, 8 observations were identified as outliers. To limit their influence, income values outside the IQR bounds were capped using winsorization. After this adjustment, the income distribution becomes more concentrated around its central range, reducing the impact of extreme values.

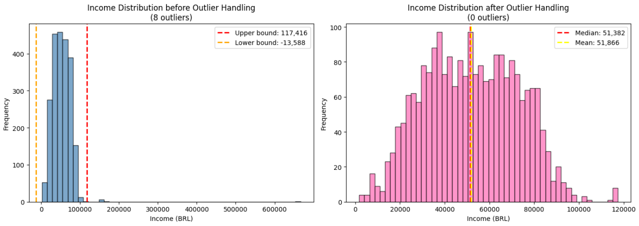

**Figure 4: Income outliers**

*3.2.3 Age Outliers.* Unrealistic values were identified in the *Year_Birth* attribute, with three customers recorded as being born before 1925. These values imply ages exceeding 110 years at the time of data collection and therefore unlikely to be valid. An assumption was made that such values may result from data entry mistakes or deliberate non-disclosure by some customers who choose not to provide their actual date of birth. To correct these values, mode imputation was applied. The mode of the valid birth years (1976) represents the most common and thus most stable value in the distribution, making it a suitable replacement. After imputation, all birth year values fell within a reasonable range. (see Figure 5).

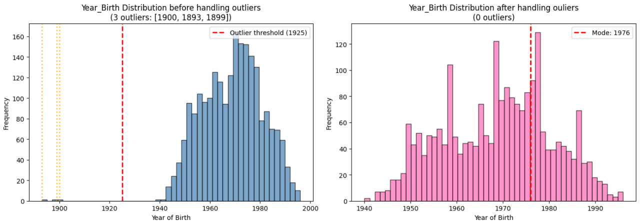

**Figure 5: Age Outliers**

*3.2.4 Marital Status Consolidation.* Inspection of the *Marital_Status* attribute revealed several rare entries (YOLO: 2 records, Alone: 3 records, Absurd: 2 records), each occuring in fewer than 1% of the records. These labels do not represent distinct marital states and instead describe situations that are most consistent with being unpartnered. Rather than dropping these records or creating a separate category with insufficient representation, the values were merged into the *Single* group, which best matches their underlying context. This consolidation improves model stability and interpretability. Following the adjustment that can be observed in Figure 6, the Single category increased from 480 to 487 records.

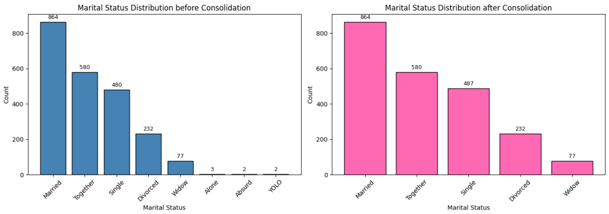

**Figure 6: Marital Status Consolidation**

## 3.3 Data Construction

Data Construction focuses on creating new attributes from existing variables to better represent customer behavior and improve model interpretability and predictive performance.

*3.3.1 Derived Attributes.* Several derived attributes were constructed to enhance interpretability and better capture underlying customer behaviour. First, *Year_Birth* was converted into a customer age by computing the difference to the latest enrollment year in the dataset. Also, *Kidhome* and *Teenhome* were combined into a unified variable,*TotalChildren.* Customer tenure was quantified by transforming *Dt_Customer* into *DaysSinceEnrollment*, representing the number of days since the customer joined the platform.
For modelling purposes, all product-category spending variables were aggregated into a single target variable, *TotalSpending*, capturing overall customer expenditure. Finally, the five campaign responses and the most recent campaign indicator were combined into *TotalCampaignsAccepted*, reducing sparsity across individual binary indicators. There is an important observation of newly combined variable *TotalChildren* which changes the distribution of household size. In the original variables shown in Figure 2, the most frequent value was zero, indicating no children in the household. After aggregation, Figure 7 shows that having one child becomes the most common outcome, providing a more informative representation of household composition.

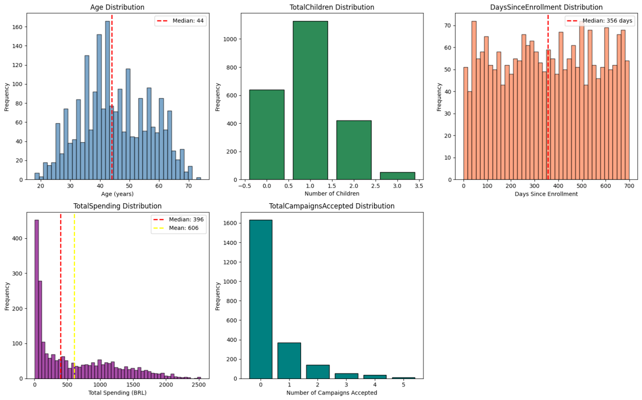

**Figure 7: Derived Attributes**

### 3.3.2 *Feature Scaling.*

Feature scaling was applied selectively based on the distributional properties of the variables and their expected role in regression modeling. As shown in Figure 8, different scaling techniques were used for different attributes. The variables *Income* and *Age* were standardized using the StandardScaler. These attributes are continuous and exhibit bell-shaped distributions without extreme skewness. In contrast, *DaysSinceEnrollment* and *Recency* were scaled using Min-Max normalization to the interval $[0, 1]$. These variables are bounded by construction and represent duration-based measures with fixed and meaningful limits. Other numerical variables related to purchase counts are highly skewed. For these attributes, scaling decisions were deferred and will be finalized once the regression method is selected.

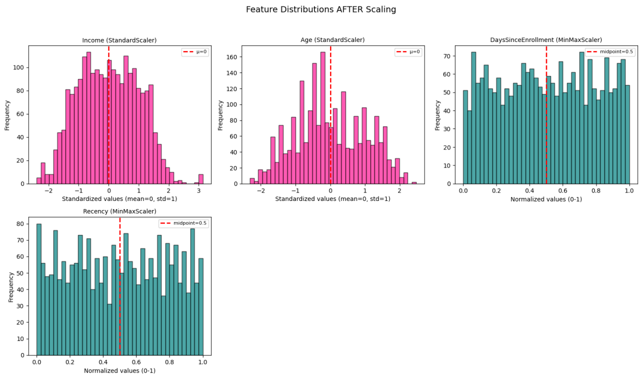

**Figure 8: Distributions of selected features after scaling**

### 3.3.3 *Categorical Encoding.*

The dataset contained two categorical attributes, *Education* and *Marital_Status*, both of which consist of non-numeric labels. Since regression models require numerical inputs, one-hot encoding was applied to convert each category into a separate binary indicator variable, as illustrated in Figure 9. To avoid perfect multicollinearity in subsequent regression analysis, one category per variable was omitted by applying *drop_first=True*. This establishes a reference category against which all remaining categories are interpreted. As a result, although the two categorical attributes contain a total of ten distinct categories, only eight binary variables were generated.

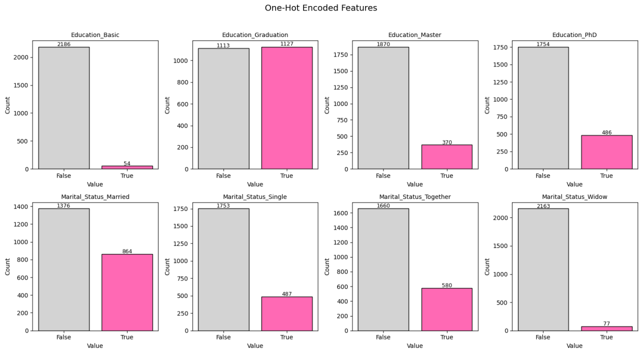

**Figure 9: Categorical Encoding**

## 3.4 Post-Transformation Correlation Analysis

Figure 10 represents the correlation heatmap computed after completing all data preparation steps, including outlier handling, feature construction, scaling, and categorical encoding. Compared to the initial correlation analysis (see Figure 1), this heatmap reflects the final model-ready feature space. Strong positive correlations with the target variable *TotalSpending* are observed for *Income* ($r = 0.804$), *NumCatalogPurchases* ($r = 0.779$), and *NumStorePurchases* ($r = 0.675$). These findings suggest that customer spending is strongly influenced by income level and the frequency of purchases across different sales channels. Negative correlations with *TotalSpending* are visible for *TotalChildren* ($r = -0.499$) and *NumWebVisitsMonth* ($r = -0.500$). This pattern suggests that customers from larger households and those who frequently browse without purchasing tend to spend less overall.

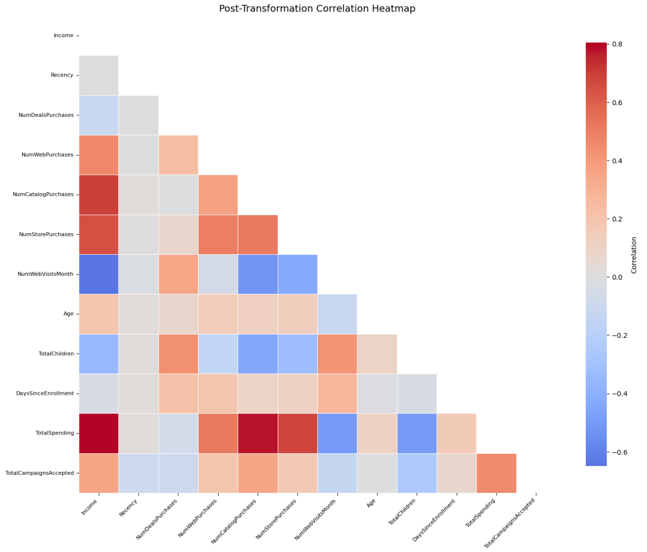

**Figure 10: Correlation heatmap**

## 3.5 Additional Preprocessing Steps Considered but Not Applied

Several further preprocessing techniques were evaluated during the Data Preparation phase but were ultimately not applied, either because their benefit was limited or because they risked distorting relevant patterns in the data:

- Advanced imputation methods (kNN, regression): not used due to very few missing values and risk of introducing artificial relationships.
- Binning of continuous variables: rejected to preserve predictive granularity.
- Non-linear transformations of predictors: evaluated but unnecessary, since only the target variable required transformation.

## 3.6 Potential Derived Attributes Considered

Beyond the derived attributes implemented in Part 3.3.1, several additional variables were evaluated for potential inclusion:

- Average spending per channel: conceptually useful but unstable due to many zero-purchase cases.
- Income–spending ratio: potentially insightful but sensitive to low-income values.
- Purchase frequency rate: Considered as purchases per unit of time, but strongly correlated with existing purchase count variables.
- Household size estimate: dismissed due to strong assumptions not supported by the data.

## 3.7 Potential External Data Sources

Several external data sources could enhance predictive performance or support deeper managerial insight. Although these data were not available for this assignment, their potential value is outlined below:

- Geographic or socioeconomic data:Regional characteristics such as average income levels or population density could capture contextual differences in purchasing power and consumption patterns that are not fully reflected by individual-level income measures. Including this information could enhance model performance and enable more targeted, location-specific marketing decisions.
- Macroeconomic indicators: Macroeconomic conditions, such as inflation and unemployment, can affect customers' purchasing power and contribute to variations in spending over time.

# 4 Modeling

The following section refers to the selection and construction of a suitable regression modeling technique. The modeling phase is iterative, meaning that it requires revisiting earlier steps like data preparation, and even business understanding, in order to refine the results.

## 4.1 Algorithm Selection

Random Forest (RF) was selected to model the regression task of customer spending behavior. The choice of RF model is motivated by the nature of the model to reduce overfitting through averaging across multiple decision trees. It also enables the extraction of feature importance rankings, which directly supports the stated data mining goals. The model was also present in a majority of the SOA researches in similar fields, and was easy to understand for first year students that do not have a lot of machine learning experience, hence this being the primary driving factor.

Model performance was evaluated against the success criteria defined in the Business Understanding phase using metrics computed on the original spending scale. Specifically, the model was required to achieve an $R^2$ of at least 0.75, a mean absolute error (MAE) below 91 BRL (15% of average spending), and a root mean squared error (RMSE) below 61 BRL (10% of average spending).

## 4.2 Data Split Strategy

The dataset was split into training, validation, and test sets using a 60/20/20 split to support model fitting, hyperparameter tuning, and final evaluation. This resulted in:

- 1,344 samples in the training set,
- 448 samples in the validation set,
- 448 samples in the test set.

A fixed random state was used to ensure reproducibility of the data split and subsequent experiments. Stratification was not applied, as the target variable is continuous. Introducing stratification would have required to categorize the target into bins, which would involve subjective design choices regarding bin size and boundaries and could therefore introduce additional bias.

## 4.3 Hyperparameter Tuning

The model was trained using RandomizedSearchCV with 5-fold cross-validation. The following hyperparameter settings were identified as optimal:

**Table 3: Final Hyperparameter Settings**

| Parameter | Description | Value |
|---|---|---|
| max_depth | Maximum depth of trees | 20 |
| min_samples_leaf | Minimum samples at leaf node | 8 |
| min_samples_split | Minimum samples to split a node | 21 |
| n_estimators | Number of trees in the forest | 53 |

## 4.4 Validation Results

Hyperparameter tuning is performed using RandomizedSearchCV with 5-fold cross-validation. 5-fold was selected as a balance between computational cost and reliable performance estimation given the moderate dataset size. RandomizedSearchCV efficiently samples over given parameter distributions, allowing exploration of a hyperparameter space. Number of random configurations evaluated is 50, that landed on the final hyperparameter configuration from Table 3.

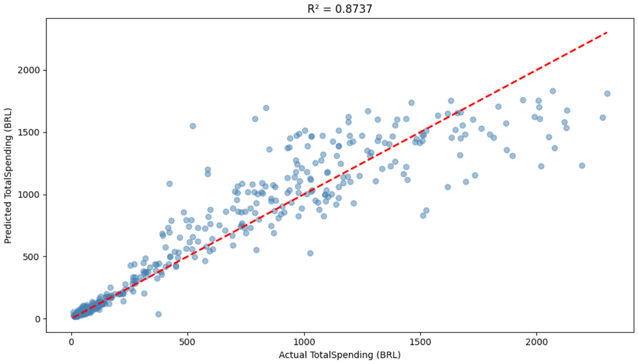

**Figure 11: Validation set: actual vs. predicted values**

Validation set performance (Figure 11):

- $R^2$ Score: 0.8737
- MAE: 121.15 BRL (20.00%)
- RMSE: 209.07 BRL (34.51%)

The model that was fitted during the final step of RandomizedSearchCV with a training $R^2$ of 0.9108 which is slightly higher than the validation $R^2$ of 0.8737, resulting in a performance gap of 0.0371 with no strong indication of overfitting.

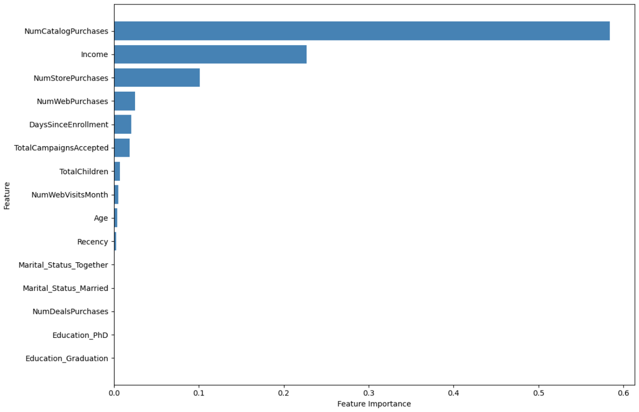

**Figure 12: Feature ranking**

Top 5 important features (Figure 12):

- NumCatalogPurchases 0.584
- Income 0.228
- NumStorePurchases 0.101
- NumWebPurchases 0.025
- DaysSinceEnrollment 0.021

## 4.5 Final Model Retraining

Following hyperparameter tuning, the final model was retrained using the combined training and validation datasets. The model was trained using the optimal hyperparameter configuration identified during the tuning phase. The retrained model achieved an $R^2$ of 0.9192 on the full training set and was subsequently used for final evaluation on the test set.

# 5 Evaluation

The goal of the Evaluation phase is to assess the performance and reliability of the developed data mining model with respect to the objectives defined in the Business Understanding phase. This includes evaluating predictive performance on unseen test data, comparing results against baseline and benchmark approaches, and assessing whether the predefined data mining success criteria have been met.

## 5.1 Test Set Evaluation

The final Random Forest model was evaluated on a held-out test set consisting of 448 samples. Table 4 reports the predictive performance of the model on unseen data.

**Table 4: Final Model Performance on Test Set**

| Metric | Value | Target | Status |
|---|---|---|---|
| R$^2$ Score | 0.8805 | $\geq 0.75$ | Pass |
| MAE | 116.98 (19.31%) | $< 15\%$ | Fail |
| RMSE | 208.96 (34.49%) | $< 10\%$ | Fail |

As we can observe, the model achieves R$^2$ score of approximately 0.88, indicating that approximately 88% of the variance in customer total spending is explained by the model. This value is close to the validation and cross-validation R$^2$ scores observed during model tuning, which suggests good generalization to unseen data.

While the predefined Data Mining Success Critiria for R$^2$ is met, the MAE and RMSE thresholds are not satisfied. It can be explained by the high variance in customer spending, with some customers exhibiting extremely high expenditures that skew the error metrics. Moreover, as students and analysts in this project, our team did not have a proper understanding of acceptable and realistic error margins for this business context, which may have led to setting overly ambitious targets. Overall, while the model demonstrates strong predictive capability as indicated by R$^2$, further refinement may be needed to meet all success criteria.

## 5.2 State-of-the-Art Performance

To evaluate the performance of the developed Random Forest regression model, we conducted a review of state-of-the-art (SOTA) approaches in customer spending and retail analytics. The literature search focused on studies applying Random Forest or ensemble-based machine learning models to customer spending and retail sales prediction problems, which closely relate to the TotalSpending target variable used in this project. Priority was given to peer-reviewed research. However, as no studies were found that apply regression models to the exact iFood dataset, relevant grey literature and publicly available analytical case studies were also considered.

Several peer-reviewed studies analyze customer spending behavior using Random Forest regression models. Mukherjee et al. (2022) investigate online buying behavior using multiple machine learning algorithms and report that Random Forest models consistently outperform linear approaches, achieving R-squared values typically between 0.80 and 0.87 depending on feature selection and data preprocessing.

Similarly, Kumar et al. (2020) apply ensemble learning techniques to consumer spending prediction and report R-squared values in the range of approximately 0.82 to 0.88, emphasizing the robustness of Random Forest models in handling nonlinear relationships and heterogeneous customer data.

Further evidence is provided by Sharma et al. (2021), who studied prediction of customer spending scores in retail environments. Their results show that Random Forest models achieve strong predictive performance, with R-squared values around 0.85 to 0.90 on test data. These findings indicate that for individual customer-level spending prediction tasks, Random Forest models typically achieve high but not perfect explanatory power due to the variability in customer behavior.

It is also vital to mention that analysis using the same iFood dataset was also considered. A publicly available GitHub case study explores customer segmentation and campaign response prediction using the iFood dataset. Although this work focuses on classification and therefore reports metrics such as accuracy and recall, it confirms that the dataset contains strong nonlinear relationships between customer characteristics, purchasing behavior, and campaign outcomes. This supports the suitability of ensemble tree-based models, such as Random Forests, for modeling customer behavior in the iFood dataset.

The final Random Forest regression model developed in this project achieves an $R^2$ score of approximately 0.88 on the held-out test set. This result lies at the upper end of the performance range reported in peer-reviewed customer spending prediction studies. Therefore, it can be concluded that the developed model performs competitively with SOTA approaches in the literature.

## 5.3 Baseline Performance

For regression tasks, trivial baseline predictors define the minimum performance level that any useful model must exceed. In this project we consider two standard baselines:

1. Mean baseline: Always predicts the mean of the training target values. By definition, this baseline achieves an R-squared value close to zero. Any model with $R^2 > 0$ therefore improves upon this trivial predictor.

2. Median baseline: Always predicts the median of the training target values. Due to the skewed nature of customer spendig, distributions, the median baseline may achieve slightly lower MAE than the mean baseline, but typically results in a negative $R^2$, indicating worse performance than predicting the mean.

These baselines establish a lower bound for acceptable model performance. A predictive model that does not outperform these trivial predictors has not learned meaningful relationships from the input features. The empirical comparison with these baselines is presented in Section 5.4.

## 5.4 Benchmark Comparison

Table 5 summarizes the predictive performance of the Random Forest model relative to simple baseline predictors.

Compared to the mean baseline, the Random Forest model achieves a substantial performance improvement, with an $R^2$ increase of 0.8805, a 77.8% reduction in MAE, and a 65.4% reduction in RMSE.

**Table 5: Benchmark Comparison on Test Set**

| Model | $R^2$ | MAE (BRL) | RMSE (BRL) |
|---|---|---|---|
| Mean Baseline | -0.0000 | 526.27 | 604.49 |
| Median Baseline | -0.1302 | 502.16 | 642.63 |
| Random Forest | 0.8805 | 116.98 | 208.96 |

To analyze model performance across different spending levels, prediction errors were evaluated (Table 6) within predefined customer spending segments.

**Table 6: Prediction Errors by Spending Segment**

| Segment | N | MAE (BRL) | RMSE (BRL) | MAE (%) |
|---|---|---|---|---|
| Very Low (0–100) | 144 | 16.19 | 34.73 | 32.4 |
| Low (100–500) | 105 | 64.37 | 119.24 | 21.5 |
| Medium (500–1000) | 74 | 201.09 | 288.16 | 26.8 |
| High (1000–2000) | 116 | 206.51 | 283.71 | 13.8 |
| Very High (2000+) | 9 | 497.68 | 517.84 | 19.9 |

As expected, absolute prediction errors increase with higher spending levels. This reflects the greater variance and volatility among high-spending customers.

## 5.5 Business Criteria Comparison

In the Business Understanding phase, the Data Mining Success Criteria were defined as achieving an $R^2$ of at least 0.75, a mean absolute error (MAE) below 15% of average customer spending, and a root mean squared error (RMSE) below 10% of average spending.

Although the $R^2$ criteria were achieved, the error-based criteria were not fully met. The observed MAE and RMSE correspond to approximately 19% and 35% of average spending, respectively, and exceeded the predefined thresholds. The gap in RMSE results is particularly important because RMSE is more sensitive to large prediction errors than MAE, which means that a few extreme errors can substantially increase the RMSE even if most predictions are reasonably accurate.

This outcome is likely driven by the high variability and skewness of customer spending, especially among high-spending customers.

## 5.6 Bias Analysis

In the Evaluation phase, we performed a simple bias analysis to investigate whether the model behaves differently across subgroups defined by education level. Education was chosen as a protected attribute because it is often correlated with socio-economic status and may influence how customers are treated in marketing applications.

Table 7 summarizes the model performance for each education group on the test set. For each group, we report the number of customers, the average actual and predicted spending, the average prediction bias, MAE and $R^2$).

**Table 7: Model performance by education group**

| Group | N | Avg. Actual (BRL) | Avg. Pred. (BRL) | Bias (BRL) | MAE (BRL) | $R^2$ |
|---|---|---|---|---|---|---|
| Basic | 11 | 96.36 | 85.44 | −10.93 | 21.61 | 0.8409 |
| Graduation | 231 | 622.13 | 639.16 | +17.02 | 109.15 | 0.8974 |
| Master | 66 | 647.92 | 619.74 | −28.18 | 107.17 | 0.9239 |
| PhD | 103 | 688.18 | 676.09 | −12.10 | 157.52 | 0.8263 |

Overall, the model achieves reasonably high explanatory power across all education groups, with $R^2$ values ranging from approximately 0.83 to 0.92. This indicates that the model is able to capture a substantial share of the variance in spending for each subgroup, and there is no group with clearly unacceptable predictive performance.

Some differences in error magnitude are visible: MAE ranges from about 22 BRL for the *Basic* group to about 158 BRL for the *PhD* group. However, these differences are consistent with the fact that higher education groups also exhibit higher average spending and higher variability, which naturally leads to larger absolute errors.

The average prediction bias is relatively small in all groups (between approximately −28 BRL and +17 BRL). The model slightly over-predicts spending for customers with *Graduation* level and slightly under-predicts for the other education groups, but the magnitude of these deviations is modest compared to the overall spending level.

In summary, the analysis does not reveal strong evidence of systematic unfairness across education levels. While moderate differences in accuracy and error magnitude exist, they are plausibly explained by differences in spending level and sample size rather than by clear discriminatory behavior.

## 6 Deployment

The goal of the Deployment phase is to document how the developed model could be used in practice and to identify risks and limitations associated with its application. This includes comparing the achieved performance with the business success criteria, formulating recommendations for model usage, and identifying remaining analytical gaps.

### 6.1 Business Objectives Comparison and Recommendations

The primary goal of predicting customer spending has been achieved. The model provides continuous spending predictions with reasonable accuracy that can support marketing strategy decisions. The secondary objective of identifying significant variables has also been met through feature importance analysis, which revealed that income, recency, and purchase history are among the strongest predictors of customer spending.

The model can support customer segmentation and budget allocation priorities. Additional analyses that would be of benefit for this case include time-series analysis of spending trends to capture seasonality effects and customer lifetime value prediction.

For deployment, the initial phase should deploy the model for customer segmentation as a low-risk high-value use case. Full automation for budget allocation should only proceed after a full fiscal year of monitoring. The model should be fully automated for high-confidence predictions where customers have complete data profiles. Edge cases and anomaly patterns should be flagged for manual review rather than automated processing.

Subsequent analyses should include developing a customer churn prediction model, implementing real-time website personalization, and creating customer micro-segmentation based on spending patterns.

### 6.2 Ethical Aspects and Impact Assessment

As discussed in the Business Understanding phase (Section 1), the model is considered low risk in terms of AI deployment, since it does not use sensitive data.

However, some ethical considerations remain relevant. The model uses demographic attributes such as education level, which may be considered as bias in terms of socioeconomic status. As shown in the bias analysis (Section 5.6), small differences in prediction behavior across education groups were observed. Although no severe or systematic bias was detected, such differences could lead to potentially unequal marketing treatment if not monitored.

### 6.3 Monitoring Plan and Intervention Triggers

Model performance metrics should be monitored monthly including R-squared score with a baseline of 0.8805, MAE as a percentage of average spending with a baseline of 19.31 percent, and RMSE percentage with a baseline of 34.49 percent. Prediction distribution statistics including mean, standard deviation, and skewness should also be tracked. Business outcome metrics should be reviewed quarterly including revenue impact from model-driven targeting, and customer feedback and satisfaction scores.

Automatic alerts requiring immediate action should trigger when R-squared drops below 0.65 representing more than a 0.10 decline from baseline, when MAE exceeds 20 percent of average spending, when the missing value rate in the Income feature exceeds 10 percent. Warning alerts requiring investigation within one week should trigger when R-squared drops below 0.70, when MAE exceeds 17 percent or when prediction variance increases by more than 25 percent. Scheduled quarterly reviews should assess the need for full model retraining, re-evaluate potential bias and fairness metrics, check feature importance stability, and review business objective alignment.

The monitoring infrastructure should include a dashboard for real-time metrics visualization, logging of all predictions with timestamps and input features, automated alerts via email or messaging platforms when triggers are breached, and a complete audit trail of model versions and performance history.

### 6.4 Reproducibility Reflection

Several aspects of this project have been well documented to support reproducibility[1]. The data source has been clearly identified as the iFood dataset from Kaggle and GitHub, with the data loading process documented including timestamps. The original dataset was preserved for reference. The preprocessing pipeline documentation includes all data cleaning steps with before and after counts, feature engineering transformations for derived attributes such as Age, TotalSpending, and TotalChildren, scaling parameters from

[1] Code and data are available at https://github.com/dasha-alekseenkova/Business-Intelligence_Lab3.

the fitted StandardScaler and MinMaxScaler objects, and categorical encoding details. Model configuration is documented including the algorithm selection rationale, hyperparameter search space for RandomizedSearchCV, best hyperparameters found, and random seeds specified with random state set to 123. The evaluation methodology records the train validation test split ratios of 60, 20, and 20 percent, the 5-fold cross-validation strategy, and multiple metrics computed and stored. Provenance tracking uses the PROV-O ontology for activity documentation with timestamps recorded for all major activities and agent associations documented for code writer and executor roles.

The requirements file with pinned versions should be regenerated.

## 7 Conclusion

This project developed a Random Forest Regressor to predict customer spending (*TotalSpending*) for iFood, a Brazilian online food delivery platform, following the CRISP-DM methodology. Overall performance is comparable to state-of-the-art results reported in related marketing analytics studies, indicating that the selected modeling approach is effective for this task.

Lessons Learned:

- Feature Engineering: Derived features (Age, TotalSpending, TotalChildren, TotalCampaignsAccepted) improved model interpretability and reduced multicollinearity
- Provenance Documentation: Using PROV-O ontology for tracking all activities, agents, and entities ensures reproducibility
- Hyperparameter Tuning: By using RandomizedSearchCV it enabled a simpler way to iteratively tune hyperparameters
- Overfitting Control: Monitoring the gap between training and validation $R^2$ to ensure good generalization

## References

[1] *iFood Data Business Analyst Test – Exploratory Analysis and Segmentation*, GitHub Repository, 2019. Available at: https://github.com/nailson/ifood-data-business-analyst-test

[2] S. Mukherjee, A. Banerjee, and R. Ghosh, *An Empirical Study on Understanding Online Buying Behaviour through Machine Learning Algorithms*, Journal of Retail Analytics, 2022.

[3] A. Kumar, R. Singh, and P. Verma, *Machine Learning Analysis of Consumer Spending Behaviour*, International Journal of Data Analytics, vol. 5, no. 2, pp. 45–58, 2020.

[4] R. Sharma, M. Patel, and K. Shah, *Leveraging Ensemble Methods for Accurate Prediction of Customer Spending in Retail*, Proceedings of the IEEE International Conference on Big Data, 2021.