ИТМО

# Laboratory 1.

**Authors:**
Jorge Sosa
Elizabetty Guseva
Siraeva Gulnara

**Professors:**
Anna V. Kalyuzhnaya, PhD
Irina Deeva, PhD

**Course:**
Methods and Models for Multivariate Data Analysis

December 11, 2024

# Contents

# 1 Introduction

## 1.1 Objective of the Analysis

The purpose of this analysis is to predict the caloric content of McDonald's menu items based on their nutritional and categorical features. The study involves data cleaning, transformation, and the application of regression models to improve prediction accuracy. Only the graphs showing the distributions of calories will be included in this report, but the transformations for all variables can be reviewed in the accompanying Google Colab notebook.

## 1.2 Dataset Overview

The dataset consists of 260 rows and 24 columns, containing nutritional information about menu items. The key variable of interest is `Calories`, which serves as the target variable. Key issues identified in the dataset include:

- Missing values in several columns, particularly `Calories`.
- Outliers and skewed distributions in nutritional variables.
- Inconsistent scales across numerical features.

# 2 Exploration of the Data

## 2.1 Initial Description

- The dataset contains both numerical and categorical variables.
- A preliminary analysis identified missing values in 13 numerical columns, primarily in `Calories` and `Total Fat`.

## 2.2 Initial Distributions

The next figures show the distributions of key numerical variables before any preprocessing. Skewed distributions and the presence of outliers were evident.

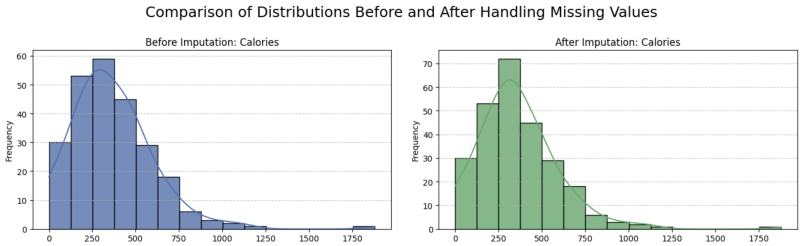


**Figure 1:** *Comparison of Distributions Before and After Handling Missing Values*

# 3 Data Cleaning and Transformation

## 3.1 Handling Missing Values

- Rows with missing `Calories` values were dropped.
- Other missing values were imputed with the median of their respective columns.

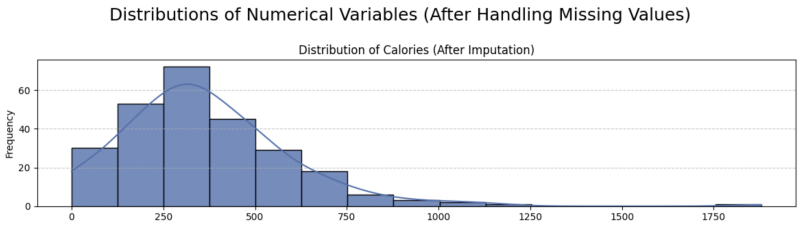


**Figure 2:** *Distributions of Numerical Variables After Handling Missing Values*

## 3.2 Addressing Skewness and Non-Normality

To address skewness and non-normality, we analyzed each variable's distribution and applied transformations as needed. HereâĂŹs the approach:

1. **Assess Skewness:** Calculate skewness values for each numerical variable to identify non-normal distributions.
2. **Select Transformations:**
   - Log Transformation for variables with right skew.
   - Square Root Transformation for moderate skewness.
   - Box-Cox Transformation (only for positive values) to address more complex skew patterns.

The following variables exhibit significant skewness ($|\text{skewness}| > 0.75$), indicating non-normal distributions that could benefit from transformation:

- **Highly Skewed:** Variables such as `Total Fat (% Daily Value)`, `Trans Fat`, `Sodium (% Daily Value)`, and `Vitamin A (% Daily Value)`, which have skewness greater than 1.5, were treated with log transformations where appropriate.
- **Moderately Skewed:** Variables like `Saturated Fat`, `Sugars`, and `Dietary Fiber` underwent square root transformations to address moderate skewness.

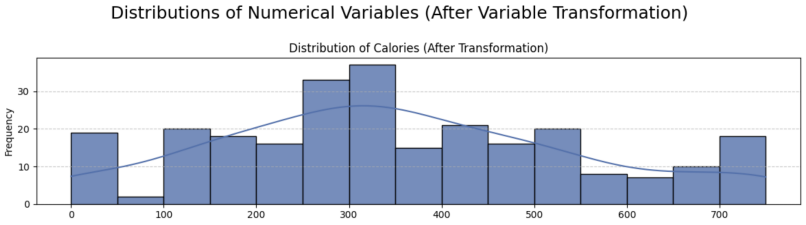


**Figure 3:** *Distributions of Numerical Variables After Transformation*

The transformed distributions appear more normalized, with reduced skewness in previously highly skewed variables:

- Log Transformations for highly skewed variables like `Total Fat (% Daily Value)` and `Sodium (% Daily Value)` helped to compress long tails.
- Square Root Transformations for moderately skewed variables like `Saturated Fat` and `Sugars` smoothed their distributions, balancing the spread of values.

This transformation is expected to enhance model performance by aligning the data closer to normality.

## 3.3 Outlier Treatment and Normalization

- Winsorization was applied to limit extreme values in key nutritional columns.
- Logarithmic and square-root transformations were used to normalize highly skewed variables.
- Standardization was performed to scale numerical variables to have zero mean and unit variance.

After applying outlier treatment, the skewness of each numerical column was recalculated to assess the remaining asymmetry. Variables with significant skewness ($|\text{skewness}| > 0.75$) include:

- **Highly Skewed:**
  - `Total Fat (% Daily Value)`: 2.315082
  - `Trans Fat`: 1.888848
  - `Sodium (% Daily Value)`: 1.664776
  - `Vitamin A (% Daily Value)`: 4.704827
  - `Vitamin C (% Daily Value)`: 1.409683
  - `Iron (% Daily Value)`: 1.273594
- **Moderately Skewed:**
  - `Saturated Fat`: 0.753011
  - `Saturated Fat (% Daily Value)`: 0.778384
  - `Sodium`: 0.999428
  - `Carbohydrates (% Daily Value)`: 0.958184
  - `Dietary Fiber`: 1.275916
  - `Dietary Fiber (% Daily Value)`: 1.238667
  - `Sugars`: 1.131317

This analysis highlights variables that may still benefit from further transformations to reduce skewness and improve normality. Likewise, the plots confirm that the standardization has centered each variable around zero and aligned them within a similar range:

These transformations should help improve model stability and ensure that each feature contributes comparably.

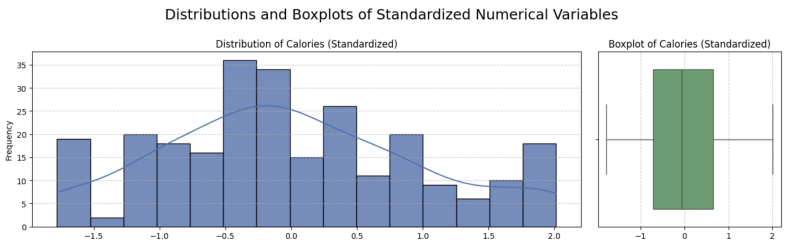


**Figure 4:** *Distributions and Boxplots of Standardized Numerical Variables*

- **Histograms:** Each variable now centers around zero, showing the effect of standardization.
- **Boxplots:** The range of each variable has been adjusted, reducing the influence of outliers while keeping most values within similar limits.

# 4 Results

## 4.1 Model Comparison

The next table summarizes the performance metrics for both models.

| Metric | Initial Model | Improved Model |
|---|---|---|
| $R^2$ | 0.9996 | 0.9781 |
| RMSE | 4.72 | 0.122 |

**Table 1:** *Comparison of Model Performance*

# 5 Conclusion

## 5.1 Model Evaluation

**Baseline Model:**

- **R-squared (0.9996):** The baseline model explained nearly all the variance in `Calories`, suggesting a close fit to the training data. However, a very high R-squared can sometimes indicate overfitting, especially if the model relies on noise or uncleaned data points.
- **RMSE (4.72):** This Root Mean Square Error means that, on average, the modelâĂŹs predictions for calories were off by about 4.72 calories. While relatively low, itâĂŹs still noticeable and might reflect the model's response to noise and uncleaned outliers.

**Cleaned and Transformed Model:**

- **R-squared (0.9781):** After cleaning and transformation, the R-squared decreased slightly. While it still explains a high percentage of the variance, the slight reduction suggests that the model is no longer overfitting to noise or outliers. This is often a sign of improved model generalizability and stability.

- **RMSE (0.12):** The RMSE dropped significantly to 0.12, indicating much higher precision. This improvement means that the cleaned modelâĂŹs calorie predictions are, on average, closer to the actual values by several calories. This reduction shows the positive impact of removing noise and transforming skewed distributions.

The cleaned model, while showing a slight decrease in R-squared, has a much lower RMSE, reflecting a substantial improvement in accuracy. The model now likely generalizes better, avoiding the influence of noise and outliers that could lead to misleadingly high R-squared values. The cleaned and transformed dataset is thus more reliable for prediction, with the model making tighter, more accurate predictions.

# 6 Annexes

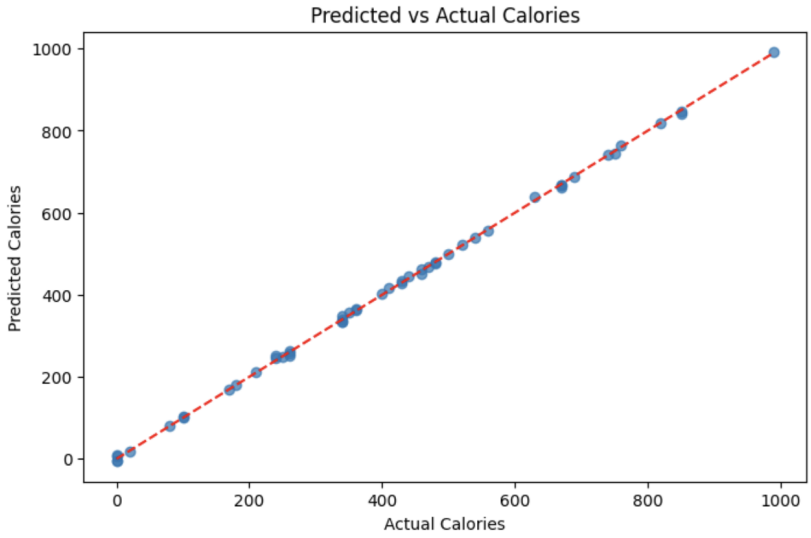


**Figure 5:** *Predicted vs Actual Calories: Initial Model*

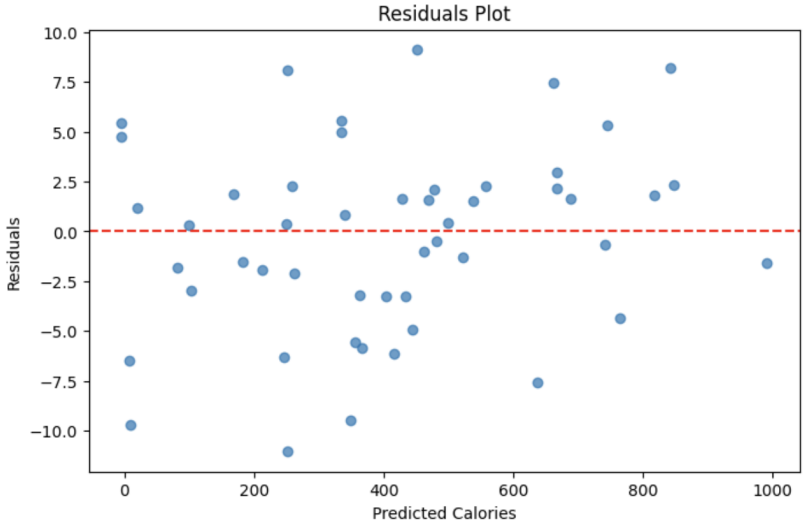


**Figure 6:** *Residuals Plot: Initial Model*

## 6.1 Code

The full code used for this analysis is available upon request. The code for this project is implemented in a Google Colab notebook. You can access the notebook using the following link:

Click here to view the Google Colab Notebook

# References


[1] J.H. Friedman. *Greedy Function Approximation: A Gradient Boosting Machine.* The Annals of Statistics, 29(5), 2001.

[2] T. Chen and C. Guestrin. *XGBoost: A Scalable Tree Boosting System.* Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pages 785 and 794, 2016.

[3] L. Breiman. *Random Forests.* Machine Learning, 45(1), 5-32, 2001.

[4] D.E. Rumelhart, G.E. Hinton, and R.J. Williams. *Learning representations by back-propagating errors.* Nature, 323, 533-536, 1986.

[5] M. Sokolova and G. Lapalme. *A systematic analysis of performance measures for classification tasks.* Information Processing and Management, 45(4), 427-437, 2009.

[6] R.J.A. Little and D.B. Rubin. *Statistical Analysis with Missing Data.* John Wiley & Sons, 1987.

[7] I. Goodfellow, Y. Bengio, and A. Courville. *Deep Learning.* MIT Press, 2016. Recuperado de `https://www.deeplearningbook.org`

[8] F. Pedregosa et al. *Scikit-learn: Machine Learning in Python.* Journal of Machine Learning Research, 12, 2825-2830, 2011.

[9] Kaggle. *Titanic - Machine Learning from Disaster.* Recuperado de `https://www.kaggle.com/competitions/titanic`

[10] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction.* Springer, 2009. Recuperado de `https://web.stanford.edu/~hastie/ElemStatLearn/`

[11] K.P. Murphy. *Machine Learning: A Probabilistic Perspective.* MIT Press, 2012.

[12] C.M. Bishop. *Pattern Recognition and Machine Learning.* Springer, 2006.

[13] Y. Zhang and Q. Yang. *A Survey on Multi-Task Learning.* IEEE Transactions on Knowledge and Data Engineering, 34(1), 2021.

[14] UCI Machine Learning Repository. *Automobile.* Recuperado de `https://archive.ics.uci.edu/dataset/10/automobile`

[15] R. Kohavi. *A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection.* International Joint Conference on Artificial Intelligence (IJCAI), 1995.

[16] F. Chollet. *Keras.* GitHub repository, 2015. Recuperado de `https://github.com/keras-team/keras`