ИТМО

Laboratory 3.

Authors:
Jorge Sosa
Elizabetty Guseva
Siraeva Gulnara

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

Course:
Methods and Models for Multivariate Data Analysis

December 12, 2024

Contents

1 Introduction

The effectiveness of machine learning models heavily depends on data preprocessing techniques. This study aims to evaluate and compare the effects of univariate and multivariate data analysis methods on linear regression performance. We utilize both classical performance metrics and statistical tests to identify the most effective data analysis approach for improving predictive quality.

2 Methodology

2.1 Dataset

The study utilizes the McDonald's menu dataset, which contains nutritional information for various menu items. The dataset includes features such as calories, fat content, protein, and other nutritional metrics. Our target variable is the calorie content of menu items.

2.2 Preprocessing Methods

Three different approaches were implemented:

- Baseline Model: Raw data without preprocessing
- Univariate Analysis: Individual feature scaling and encoding
- Multivariate Analysis: Feature interactions and dimensional reduction

2.3 Evaluation Metrics

The following metrics were used to evaluate model performance:

- Root Mean Squared Error (RMSE)
- Mean Absolute Error (MAE)
- RÂš Score
- Adjusted RÂš Score

3 Statistical Analysis

3.1 Friedman Test

The Friedman test was conducted to evaluate overall differences in model performance:

- RMSE: Statistic = 8.400, p-value = 0.015
- MAE: Statistic = 7.600, p-value = 0.022
- RÂš Score: Statistic = 8.400, p-value = 0.015
- Adjusted RÂš Score: Statistic = 10.000, p-value = 0.007

3.2 Wilcoxon Signed-Rank Test

For RMSE:

- Baseline vs Multivariate: p-value = 0.062
- Baseline vs Univariate: p-value = 0.062
- Multivariate vs Univariate: p-value = 0.125

3.3 Nemenyi Test

Critical Difference = 1.625

- RMSE: Significant difference between Baseline and Multivariate (Rank Difference = 1.800)
- MAE: No significant differences
- RÂš Score: Significant difference between Baseline and Univariate (Rank Difference = 1.800)
- Adjusted RÂš Score: Significant difference between Multivariate and Univariate (Rank Difference = 2.000)

3.4 Paired T-Tests

Results showed significant differences between:

- Baseline vs Multivariate: All metrics ($p < 0.05$)
- Baseline vs Univariate: All metrics ($p < 0.05$)
- Multivariate vs Univariate: Only significant for Adjusted RÂš Score

4 Model Selection

Bayesian Information Criterion (BIC) analysis:

- Univariate Model BIC: -15725.17
- Multivariate Model BIC: -1065.70
- Bayes Factor: <1, indicating strong evidence for the univariate model

5 Results and Discussion

The statistical analyses reveal several key findings:

1. Both preprocessing methods significantly outperform the baseline model across all metrics
2. The univariate model demonstrates superior performance when considering model complexity (BIC)
3. The multivariate approach shows particular strength in adjusted RÂš performance
4. Statistical tests consistently indicate significant differences between preprocessed and baseline models

6 Conclusion

While both preprocessing methods demonstrate significant improvements over the baseline, the univariate approach emerges as the preferred choice when considering the balance between model complexity and performance. This finding suggests that simpler preprocessing techniques may be sufficient for this particular dataset and modeling task.

References

[1] AdministraciÃşn Nacional de Seguridad del TrÃąfico en Carreteras. Recuperado de `https://www.nhtsa.gov/standards`

[2] AdministraciÃşn Nacional de Seguridad del TrÃąfico en Carreteras. Recuperado de `https://www.nhtsa.gov/cafe`

[3] Agencia de ProtecciÃşn Ambiental. Recuperado de `https://www.epa.gov/clean-air-act-overview`

[4] Agencia de ProtecciÃşn Ambiental. Recuperado de https://www.epa.gov/greenvehicles/greenhouse-gas-emissions-standards

[5] Agencia de ProtecciÃşn Ambiental. Recuperado de https://www.epa.gov/fuel-economy-trends/fuel-economy-and-greenhouse-gas-standards

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