Workshop 1: Sweet Regression Competition Analysis

Samuel Aljure Bernal- 20202020111
Carlos Alberto Barriga Gámez-20222020179
David Santiago Aldana Gonzalez-20222020158
Juan Diego Alvarez Cristancho-20221020076

September 2025

1 Overview

The objective of the competition is to develop a predictive system based on a regression model that estimates chocolate sales for a company expanding into new regions. The training dataset includes variables such as marketing campaign types, regional characteristics, consumer-related characteristics, and advertising exposure, measured using Gross Rating Points (GRP). The system must process these inputs to learn sales patterns and generate predictions for the test dataset.

Regarding the competition restrictions, according to the rules, students must implement their solution in R. The submitted result is evaluated against a reference model defined by the teaching team. Furthermore, there is a limit of 10 attempts to submit the team-defined solution, and only one of these solutions must be the final solution.

2 Systemic Analysis

The system is defined by four main components: inputs, processes, outputs, and constraints.

2.1 Inputs

Inputs consist of structured data artifacts (training dataset, test dataset, and sample submission file), the competition rules governing participation (use of R, maximum of ten attempts, one final graded submission), and participants' prior knowledge and skills.

2.2 Processes

Processes within the system include data preprocessing, selecting a regression model suited to the problem, developing models in R using the selected regression technique, validating them to ensure predictive accuracy, and generating submission files.

To better illustrate the logical sequence of activities, Figure 1 presents a flowchart of the entire process. This diagram shows the iterative nature of the competition workflow: starting with the reception of datasets and rules, continuing through preprocessing, model construction in R, validation, and prediction generation. Decision nodes show the key points where the process can loop back either because the data are not ready, the model is invalid, or the error rate is unacceptable. Finally, once predictions are generated, they are submitted as CSV files, evaluated, and scored. The loop closes depending on whether submission attempts remain available.

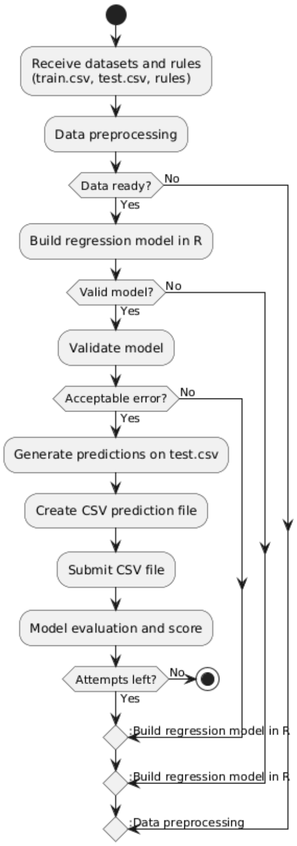
Receive datasets and rules
(train.csv, test.csv, rules)
Data preprocessing
Data ready?
No
Yes
Build regression model in R
Valid model?
No
Yes
Validate model
Acceptable error?
No
Yes
Generate predictions on test.csv
Create CSV prediction file
Submit CSV file
Model evaluation and score
Attempts left?
No
Yes
:Build regression model in R
:Build regression model in R
:Data preprocessing

Figure 1: System flowchart

2.3 Outputs

The system outputs are prediction files submitted to the Kaggle platform and corresponding evaluation scores, which measure performance against a benchmark model specified by the instructors. Feedback is incorporated through iterative submissions, allowing participants to improve their models within the defined constraints.

Figure 2 illustrates the Data Flow Diagram (DFD) of the Sweet Regression. The diagram details the interaction between external actors (lecturers, students, and the Kaggle platform) and the internal components of the R-based system. Lecturers provide the training dataset, test dataset, and competition rules, which are processed by the student through data preprocessing, model training, validation, and prediction generation. The resulting CSV file of predictions is submitted to the Kaggle platform, which in turn returns evaluation scores and feedback. This cycle of information emphasizes the iterative nature of the competition, where students refine their models based on received feedback until submission limits are reached.

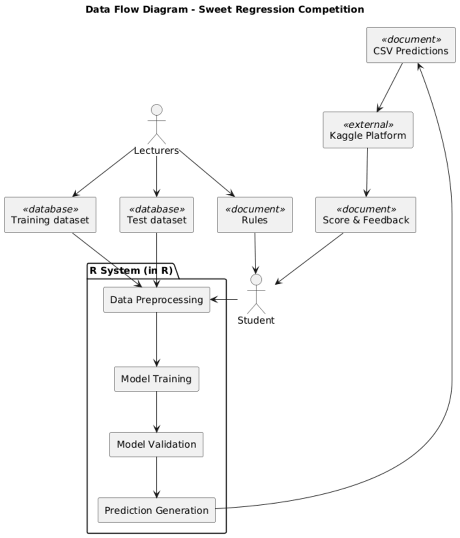

Figure 2: Data Flow Diagram

2.4 Constraints

The main system constraints include:

- Implementation of the solution in R.
- Limit of 10 attempts to submit the solution generated by the team.
- Evaluation of the solution proposed by the team against a reference model defined in the competition.

- Due to the number of variables, training the model can be slow and computationally expensive.

To complement the systemic analysis, Figure 3 presents a structural diagram of the ”Sweet Regression” competition. The diagram organizes the system into its main components inputs, processes, outputs, and constraints while explicitly showing the interactions between students and lecturers. Students interact directly with the datasets, the rules, and their own prior knowledge to carry out the processes of data preprocessing, model construction, validation, and prediction generation. These processes generate outputs in the form of submission files, evaluation scores, and feedback, which are then received both by the students and the lecturers responsible for assessment. Constraints such as the mandatory use of R, the maximum of ten attempts, and the requirement of a single final graded submission act as regulatory mechanisms, shaping the way processes unfold. This representation provides a holistic view of the system.

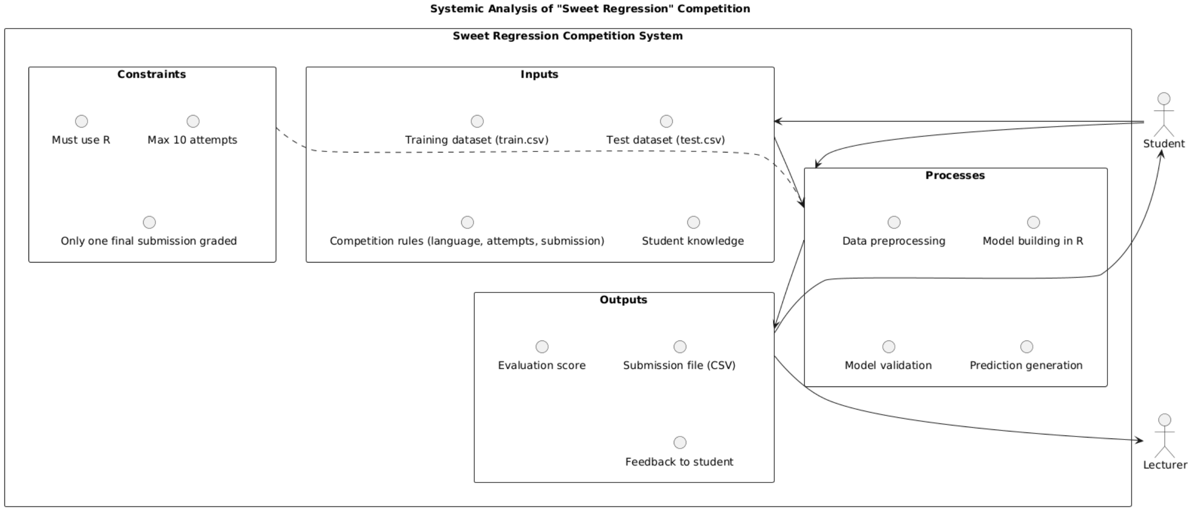

Figure 3: System architecture diagram

3 Complexity and Sensibility

3.1 Inputs

- Training dataset: high sensitivity, small changes in data quality (missing values, outliers, noise) strongly affect the model.
- Student knowledge: very high sensitivity, students with experience in R and machine learning achieve more stable models; inexperienced students produce models that are more sensitive to small changes.

3.2 Processes

- Data preprocessing: high sensitivity, normalization, imputation, or variable selection techniques drastically change the results.
- Model validation: medium sensitivity, splitting data into train/test in different ways can produce noticeable variations in the mean absolute error (MAE).
- Prediction generation: low sensitivity, once the model has chosen the prediction generation is mechanical.

3.3 Constraints

- Max 10 attempts: high sensitivity, less margin to adjust parameters and obtain a better solution.
- Only one file to evaluate: very high sensitivity, only one error can invalidate all work.

4 Chaos and Randomness

The Sweet Regression competition presents several systemic risks that introduce elements of uncertainty and randomness into its evaluation process. A primary source of concern arises from the reliance on a relative reference model, wherein the best-performing submission is assigned a perfect score, and all remaining models are evaluated against it. This approach transforms the assessment from an absolute measure of predictive accuracy into a relative comparison, thereby limiting the validity of the results as indicators of real-world performance.

The issue is further compounded by the variability in the quality of the models submitted by participants; if the strongest model within the competition is of limited accuracy, the resulting benchmark artificially inflates the scores of competing models, weakening the robustness of the evaluation. In addition, the dependency on the strategies of other participants introduces a competitive dynamic that affects fairness and reproducibility, since identical models may receive different scores under different competitive conditions.

The restriction of ten submissions further increases uncertainty, as outcomes may reflect submission strategy rather than the intrinsic quality of the predictive model, disproportionately penalizing participants who commit early errors.

Similarly, data variability can make it difficult to predict patterns even when all variables are available. Individual consumer decisions do not always follow logical patterns. Therefore, trends that work in one region may perform poorly in another.

Finally, the stochastic nature of many machine learning algorithms, including ensemble methods such as random forests and gradient boosting, introduces inherent variability unless random seeds are explicitly controlled, which diminishes the reproducibility of results.

Mitigation of these risks requires the integration of absolute error metrics in internal validation, the use of independent benchmarks beyond the competition dataset, and the systematic application of reproducibility practices such as seed control and rigorous documentation of experimental procedures.

5 Conclusions

The systemic analysis of the “Sweet Regression” competition allows for the identification of strengths and weaknesses in the system’s design and implementation:

5.1 Strengths

- There is a clear structure of the competition’s inputs (datasets, rules, platform) and the expected output values.
- The large amount of information contained in the training data provides important patterns that can improve the predictability of the regression model.
- The requirement to implement the system in R guarantees a common technical standard and promotes the learning of a language widely used in statistical analysis.

5.2 Weaknesses

- Minimal variations in the data can significantly alter the results, hindering system stability.
- Unpredictable consumer decisions introduce randomness that limits the accuracy of the regression models.
- The selected regression model can generate unwanted interactions with the training data.
- The large amount of data can make model training computationally expensive.
- The requirement for a single evaluated submission introduces great sensitivity: an error in that submission invalidates the work done.

These elements demonstrate that the system is attractive for analysis, as it is not a completely deterministic model, but rather incorporates randomness and even chaotic behavior.