Enhancing Feature Engineering with Large Language Models through Causal Feature Engineering & Feature Selection

Yuval Saadaty

Abstract

Feature engineering is essential but challenging in data science, often requiring extensive domain knowledge. Our research aims to simplify this process by using large language models (LLMs) instead of human expertise, enhancing feature generation through automation and feature selection. Our method builds on the approach in "Large Language Models for Automated Data Science: Introducing CAAFE for Context-Aware Automated Feature Engineering". Differing from the referenced study, we employ LLM to aggregate all generated features, hypothesizing that some might prove beneficial collectively, even if not immediately apparent. We opt to retain all potential features, disregarding LLM suggestions to remove columns for a thorough analysis. Our experimental approach includes three phases:

1. First Experiment: Apply only feature selection methods.
2. Second Experiment: Conduct causality analysis on the features then, apply feature selection methods to these selected features.
3. Third Experiment: First apply feature selection methods, and then conduct causality analysis.

The first experiment achieved the highest accuracy, surpassing the results reported in the original paper. Subsequent experiments incorporating causality before or after feature selection resulted in a decrease in accuracy. We used the latest TabPFNClassifier for compatibility and conducted comparative analyses to validate our approach.

1 Problem Description

What Element in the DS Pipeline Are We Improving?

Feature engineering, the process of creating meaningful features from raw data, remains a bottleneck due to its reliance on domain expertise. Existing automated solutions have sought to alleviate these challenges. Notably, advancements in LLMs have paved the way for new methodologies. These models can analyze extensive datasets and generate innovative features without direct human oversight.

Challenges

- Feature engineering is often domain-dependent and requires expert intuition.
- Some feature selection methods can capture correlations but do not guarantee causality.

Table 1: Comparison of Baseline Results and Current Experiment Outcomes with and without feature engineering

Dataset	**Baseline Results**		**Current Experiment Results**	
	No Feat. Eng.	**CAAFE (Gemini-2.0-flash)**	**No Feat. Eng.**	**CAAFE (Gemini-2.0-flash)**
CMS	0.7375	0.7393	0.5962	0.5907
Diabetes	0.8427	0.8434	0.7917	0.7969
Eucalyptus	0.9319	0.9319	0.7228	0.7337
Airlines	0.6211	0.6203	0.6420	0.6203

2 Solution Overview

The classifier mentioned in the original paper utilized an older version of the TabPFNClassifier. We encountered difficulties running this older version, so we opted to use the latest version, which yielded different results from those shown in the paper. Consequently, my initial step was to evaluate the four datasets using the original code. Additionally, we was unable to use the GPT API, so the results presented here are derived from using the Gemini-2.0-flash model.

The second step in our research involved modifying the original CAFFE model code to aggregate all possible feature combinations suggested by the LLM output. Unlike the original approach, where only feature additions that improved the ROC curve score and accuracy were aggregated, our approach aggregated all potential new features after 5 epochs, disregarding the LLM's suggestions for feature removal. Subsequently, causality analysis and feature selection were applied to retain features with causal relationships and significant impact on the model.

Figure 1 displays the total number of features against the method applied to the CMC dataset. The x-axis represents the count of features, and the y-axis represents the applied method.

2.1 Feature Generation in CAFFE

In the original CAFFE model application, only one new feature was added: **Age_Children_Interaction**, calculated by multiplying the wife's age by the number of children she has ever given birth to.

Consequently, in Figure 1, the "CAFFE selected features" count is 11, reflecting the original 10 features plus one new feature.

2.2 Generated Features in Our Experiment

In our experiments, we generated 9 additional features from the LLM output over 5 epochs.Here are some new generated features:

- **Age_Education**: Wifes_age × Wifes_education
- **Education_Difference**: Husbands_education − Wifes_education
- **Children_Age**: Number_of_children_ever_born × Wifes_age

Therefore, in Figure 1, the "Generated Features (All)" bar shows a total of 19 features, comprising the original 10 plus the 9 new features created from the LLM in 5 epochs.

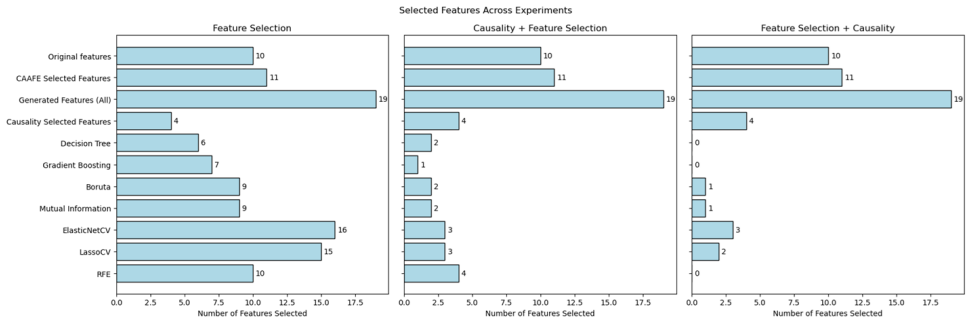

Figure 1: The three panels from left to right illustrate the number of features selected in CMC dataset, using different methods: the first panel shows feature selection only, the second combines causality analysis with feature selection, and the third applies feature selection followed by causality analysis.

Experiment 1: Evaluating Feature Selection

Hypothesis: Iterative refinement using feature selection methods improves model performance.

Process:

1. Extract dataset metadata.
2. Generate features using LLMs.
3. Filter features using feature selection methods.
4. Train TabPFNClassifier on selected features.

5. Evaluate using Accuracy.

Feature Selection Methods Used: Recursive Feature Elimination (RFE), LassoCV, ElasticNetCV, Mutual Information, Boruta, Gradient Boosting, Decision Tree.

Observations: In Figure 1, the left plot shows the number of features selected by each method. RFE selected 10 features from the total of 19; LassoCV selected 15; ElasticNetCV selected 16; Mutual Information and Boruta each selected 9; Gradient Boosting selected 7; and Decision Tree selected 6. Figure 2 plots the names of the features selected by each method in each experiment.

Experiment 2: Evaluating Causal and then Feature Selection

Hypothesis: Features identified through causal inference will enhance model generalization and accuracy.

Process:

1. Extract dataset metadata.
2. Generate features using LLMs.
3. Filter features using causal inference techniques. Retaining only those features with a causal estimate greater than 0.1
4. Filter features using feature selection methods.
5. Train TabPFNClassifier on selected features.
6. Evaluate using Accuracy.

Experiment 3: Evaluating Feature Selection and then Causal

Hypothesis: Features identified through causal inference will enhance model generalization and accuracy.

Process:

1. Extract dataset metadata.
2. Generate features using LLMs.
3. Filter features using feature selection methods.

4. Filter features using causal inference techniques.Retaining only those features with a causal estimate greater than 0.1

5. Train TabPFNClassifier on selected features.

6. Evaluate using Accuracy.

Observations: Figure 1 displays the number of selected features after each experiment and method. In the second experiment, the middle plot shows a bar named "Causality Selected Features," representing the number of features selected after causality analysis and before feature selection. Therefore, all bars below this represent features that underwent additional feature selection, likely resulting in the same or a reduced number of selected features.

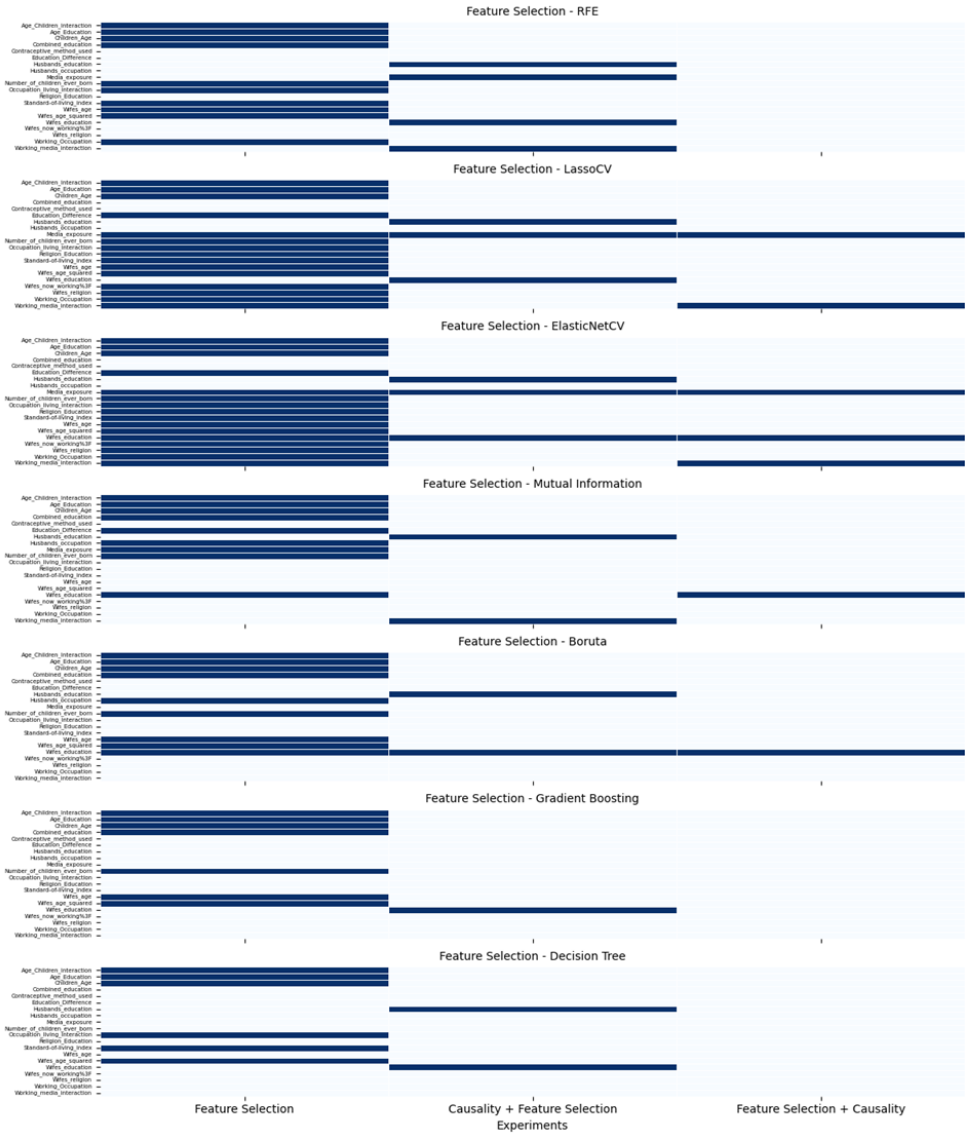

Figure 2: Results of various feature selection methods applied to a CMC dataset.

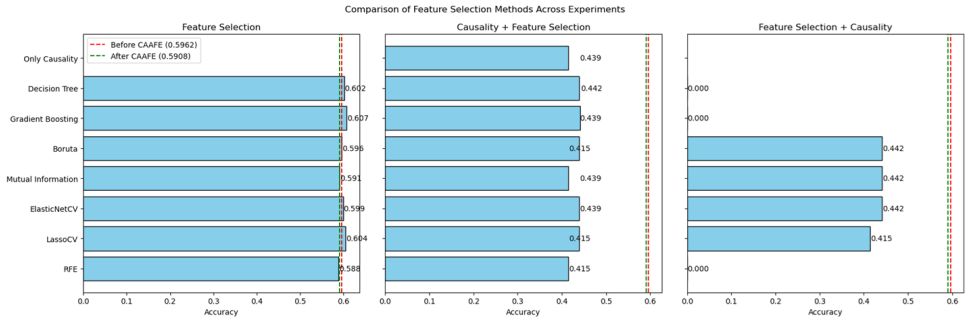

Figure 3: Comparison of accuracy scores across different feature selection methods in three experiments using the CMC dataset.

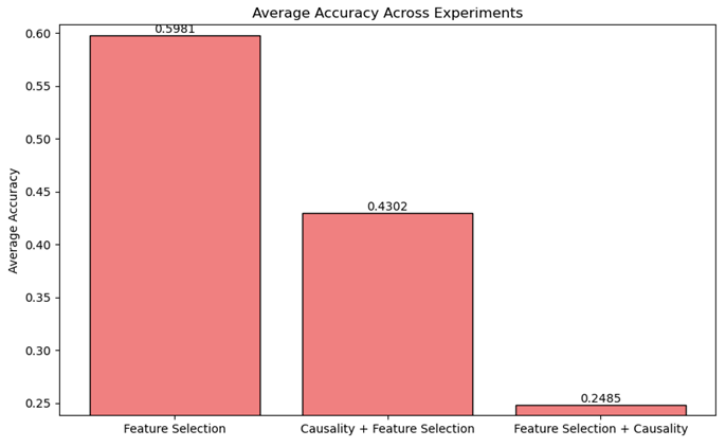

Figure 4: The average accuracy across three different experiments conducted on the CMC dataset.

3 Experimental Evaluation

In Figure 3, the left plot shows Experiment 1 results on the CMC dataset, where Gradient Boosting achieved the highest accuracy of 0.607, surpassing the original CAAFE method's accuracy of 0.5907. Figure 5 further highlights that Gradient Boosting selected 7 features not recognized by the original CAAFE, which individually did not enhance accuracy.

The middle plot of Figure 3 details an experiment that first applied causality analysis, followed by feature selection. The initial causality-only accuracy was 0.439 ("Only Causality"). This stringent filtering left no features with a causal estimate over 0.1, leading to the termination of the experiment as subsequent feature selection was unfeasible. This explains the notably lower accuracies in this plot compared to Experiment 1, where the average accuracy was about 0.5981 as shown in Figure 4.

The right plot illustrates Experiment 3, where feature selection preceded causality analysis. This order also resulted in lower accuracies, with some cases yielding no features post-causality

analysis, represented by an empty bar in the plot.

Given these results, the diminished accuracies in Experiments 2 and 3, where causality analysis constrained the feature space, are consistent across all tested benchmarks. The focus will therefore remain on the superior accuracies from Experiment 1 across all benchmarks.

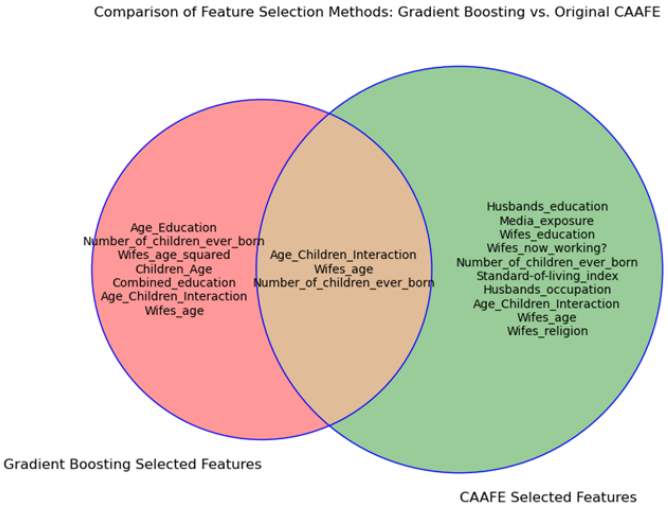

Figure 5: Compares the features selected by Gradient Boosting and Original CAAFE, illustrating both unique and shared feature selections between the two methods for feature selection.

Conclusion of Experiments

Overall, we observe that utilizing all potential new features generated by the LLM and then applying feature selection methods resulted in better outcomes compared to the approach suggested in the CAAFE model. As shown in Table 2, across all four benchmarks, our method achieved the same or better results than the original CAAFE approach.

Table 2: Comparison of Best Feature Selection Methods vs. CAAFE (Gemini-2.0-flash) Accuracy

Dataset	**Method**	**Feature Selection Accuracy**	**CAAFE (Gemini-2.0-flash) Accuracy**
CMC	Gradient Boosting	0.607	0.5907
Diabetes	Mutual Information	0.8177	0.7969
Eucalyptus	Decision Tree	0.734	0.734
Airlines	Boruta	0.652	0.642

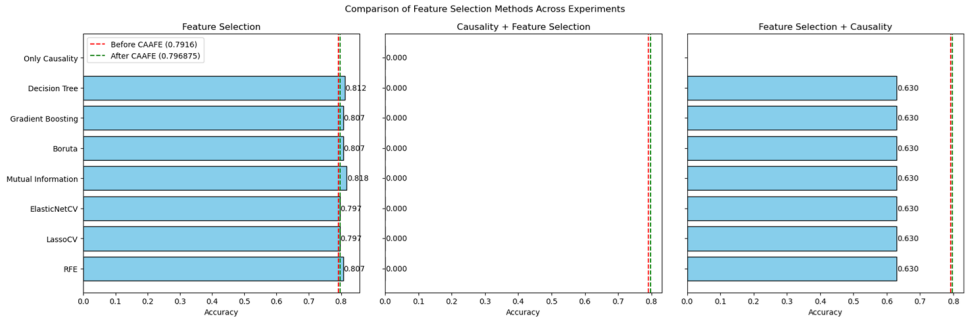

Figure 6: Comparison of accuracy scores across different feature selection methods in three experiments using the Diabetes dataset.

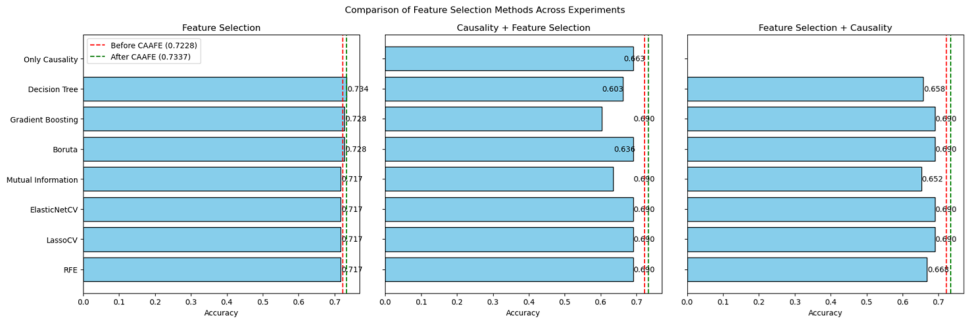

Figure 7: Comparison of accuracy scores across different feature selection methods in three experiments using the Eucalyptus dataset.

4 Related Work

4.1 LLMs for Automated Data Science: Introducing CAAFE for Context-Aware Automated Feature Engineering

This work introduced the use of LLMs (LLMs) for automated feature generation, but it did not incorporate causal filtering. Our approach builds upon this by utilizing all features generated by the LLM, as opposed to their method of selectively adding features only if they demonstrate accuracy improvements. We extend their framework by applying both feature selection and causality analysis to all generated features.

4.2 Optimized Feature Generation for Tabular Data via LLMs with Decision Tree Reasoning

This study employs decision trees to evaluate features generated by LLMs but does not incorporate causality in its analysis. This iterative enhancement inspired us to manipulate the generated features more extensively.

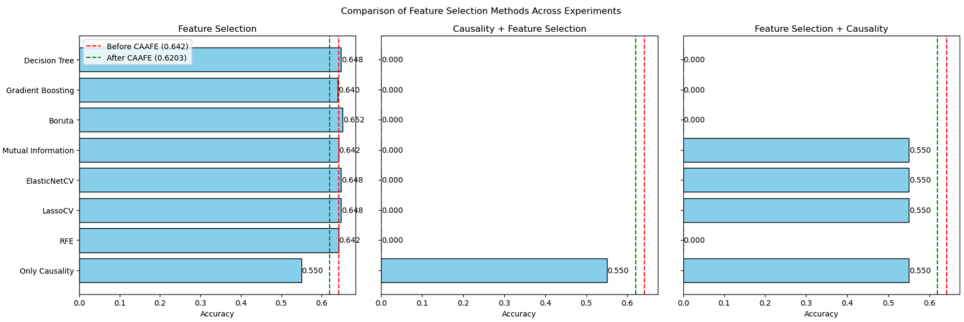

Figure 8: TThe three panels from left to right illustrate the number of features selected in Airlines dataset, using different methods.

4.3 Automated Feature Engineering Using Causal Inference

While this research focuses on applying causality in feature engineering, it does not leverage LLMs for the generation of features. Our solution differs by integrating causality with LLM-driven feature generation. We drew inspiration from their causal approach to apply a similar methodology to the LLM-generated features.

5 Conclusion

Throughout this research, we examined three experiments to assess different feature selection methodologies integrated with causal analysis. Our findings indicate that the first experiment, which solely utilized feature selection methods, achieved the highest results. This outcome underscores the effectiveness of direct feature selection strategies over combined approaches with causality in enhancing model performance.

While attempting to improve the CAAFE approach by integrating feature selection and causality, we observed that causality analysis often decreased model accuracy. This reduction in performance highlights a potential trade-off between employing causally relevant features and achieving optimal accuracy. In future studies, alternative methods to incorporate causality into the model development process should be explored, potentially leading to more robust models that maintain high accuracy while incorporating causal insights.

This project provided valuable insights into using LLMs as a domain knowledge substitute for human-driven feature engineering. The results confirmed that LLMs could effectively enhance model outcomes, underscoring their potential in automating and improving feature engineering processes. However, the application of causality, while ensuring features have a reasoning relationship with the labels rather than mere correlations, significantly impacted the model's

accuracy.

Given that the datasets used from CAAFE were limited to small samples, up to 2,000 in total, it would be prudent to test the impact of causal features on larger datasets in future studies. There is a possibility that features with genuine causal relationships could demonstrate improved generalization capabilities on broader datasets, potentially validating the causal approach's efficacy in realistic settings.