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," where GPT-3.4 and GPT-4 adapt feature engineering based on dataset descriptions to replace human domain knowledge. We evaluate features using accuracy and ROC curve scores, only retaining those that enhance model performance. Unlike the referenced paper, we use Gemini-2.0-flash to aggregate all new features, regardless of immediate improvements, hypothesizing that some may be beneficial when combined with others. We ignore LLM suggestions to remove columns, keeping all possible features for comprehensive analysis. Our experimental approach includes three phases:

1. First Experiment: Apply only feature selection methods and then evaluate the model.
2. Second Experiment: Conduct causality analysis on the features, retaining only those with a causal estimate greater than 0.1. Then, apply feature selection methods to these selected features.
3. Third Experiment: First apply feature selection methods, and then conduct causality analysis on the selected features.

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 large language models (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.
- Traditional 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. I encountered difficulties running this older version, so I 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, I 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 **??**, 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. These features are:

- **Age_Education**: Wifes_age × Wifes_education
- **Education_Difference**: Husbands_education − Wifes_education
- **Children_Age**: Number_of_children_ever_born × Wifes_age
- **Working_Occupation**: Wifes_now_working? × Husbands_occupation
- **Religion_Education**: Wifes_religion × Wifes_education
- **Wifes_age_squared**: Wifes_age^2
- **Combined_education**: Wifes_education + Husbands_education
- **Occupation_living_interaction**: Husbands_occupation × Standard-of-living_index
- **Working_media_interaction**: Wifes_now_working? × Media_exposure

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.

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.

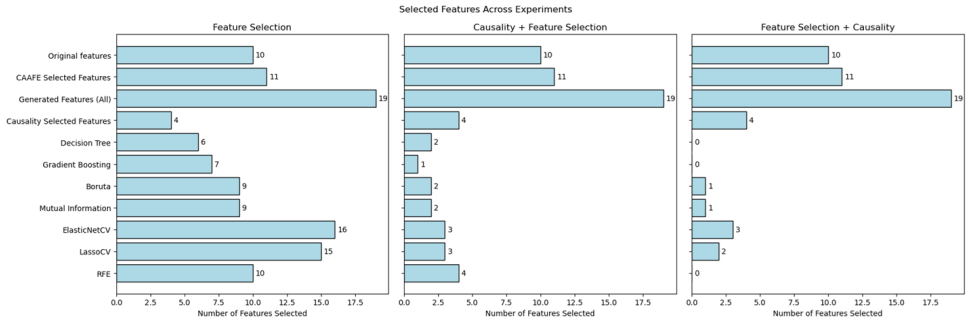

Figure 1: The three panels from left to right illustrate the number of features selected in CMC dataset, using different methods: the first panel shows feature selection only, the second combines causality analysis with feature selection, and the third applies feature selection followed by causality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

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):** Iteratively removes the least important features.
- **LassoCV:** Employs L1 regularization to eliminate unimportant features.
- **ElasticNetCV:** Combines L1 and L2 regularization to improve feature selection.
- **Mutual Information:** Measures the information gain between features and the target variable; it achieved the best accuracy across datasets.
- **Boruta:** A feature selection algorithm based on Random Forest importance scores.
- **Gradient Boosting:** Identifies the most relevant features based on boosting techniques.
- **Decision Tree:** Selects features based on their contribution to tree splits.

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 3 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.

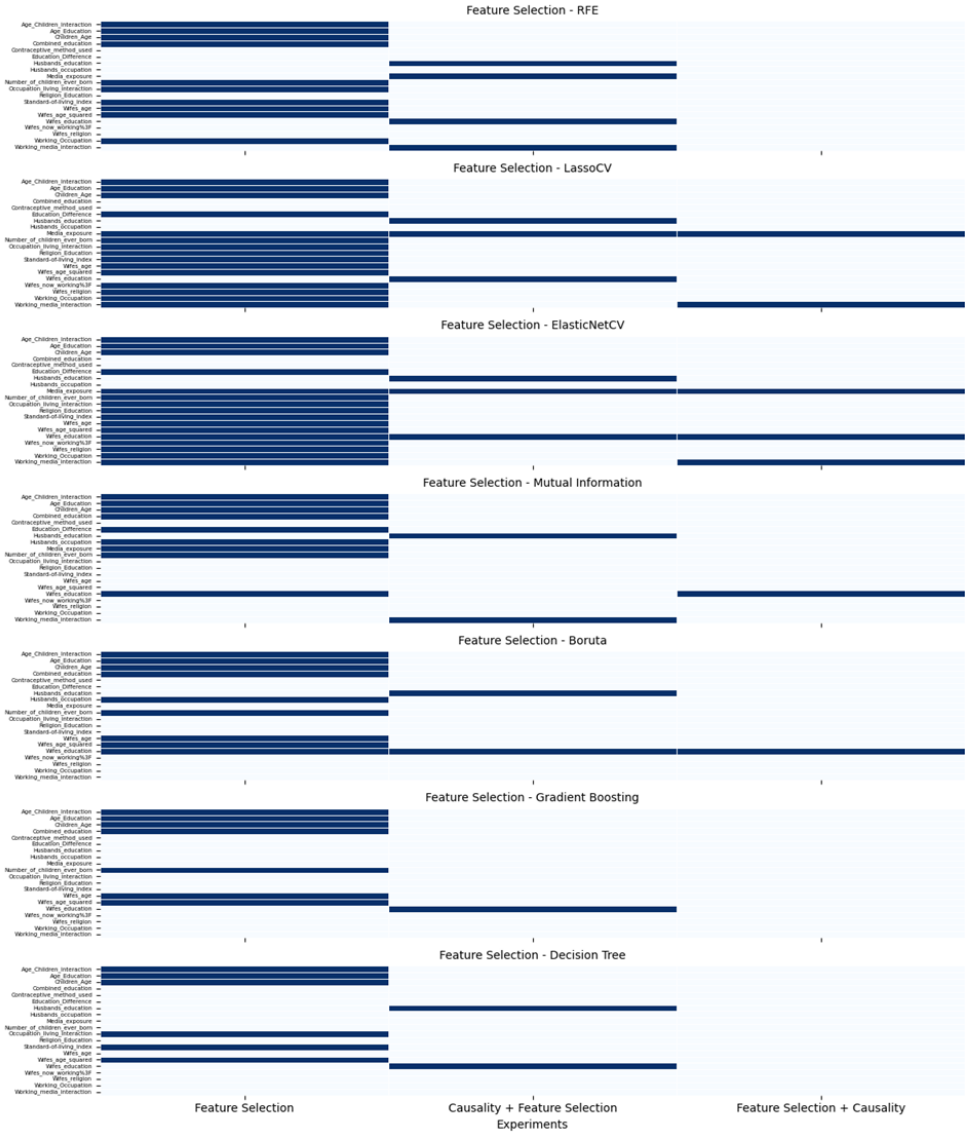

Figure 2: Results of various feature selection methods applied to a CMC dataset.

3 Experimental Evaluation

In Figure 4, we observe the results from each experiment on the CMC dataset, where we note that in Experiment 1 (the left plot), using Gradient Boosting, we achieved the highest accuracy (0.607), surpassing the original CAAFE method which recorded an accuracy of 0.5907. Figure 6 also illustrates differences in feature selection, highlighting that Gradient Boosting selected 7 features not considered by the original CAAFE approach, as evaluating them individually did not enhance accuracy. The results presented in Figures 3 and 4 are based on the CMC dataset.

The middle plot presents results from an experiment where causality analysis was applied first, followed by feature selection. The initial accuracy achieved solely through causality analysis

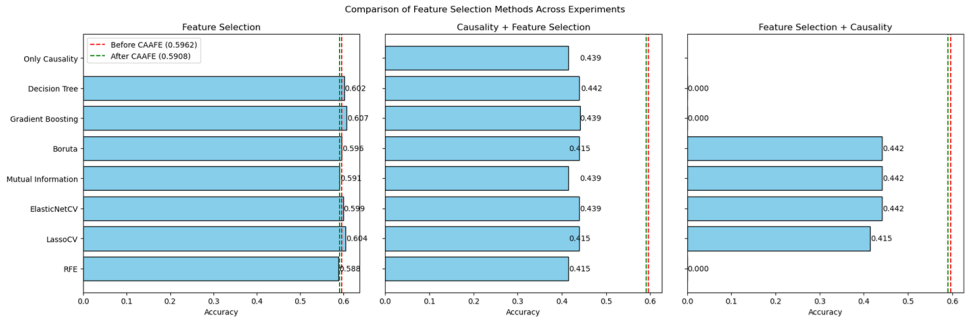

Figure 3: Comparison of accuracy scores across different feature selection methods in three experiments using the CMC dataset.

was 0.439, as indicated by the first bar labeled "Only Causality." This phase significantly constrains the feature space by retaining only causally relevant features, which inherently limits the potential accuracy improvements that subsequent feature selection methods can achieve.

In some benchmarks within this experimental setup, no features achieved a causal estimate greater than 0.1. Consequently, no features were selected, and the experiment had to be terminated prematurely because feature selection methods cannot be applied without any initial features. This resulted in cases where the accuracy could not be improved beyond the initial causality analysis, explaining the consistently lower accuracy scores across all methods in the middle plot compared to those in the left plot, where the average accuracy of the first experiment using various feature selection methods was approximately 0.5981.

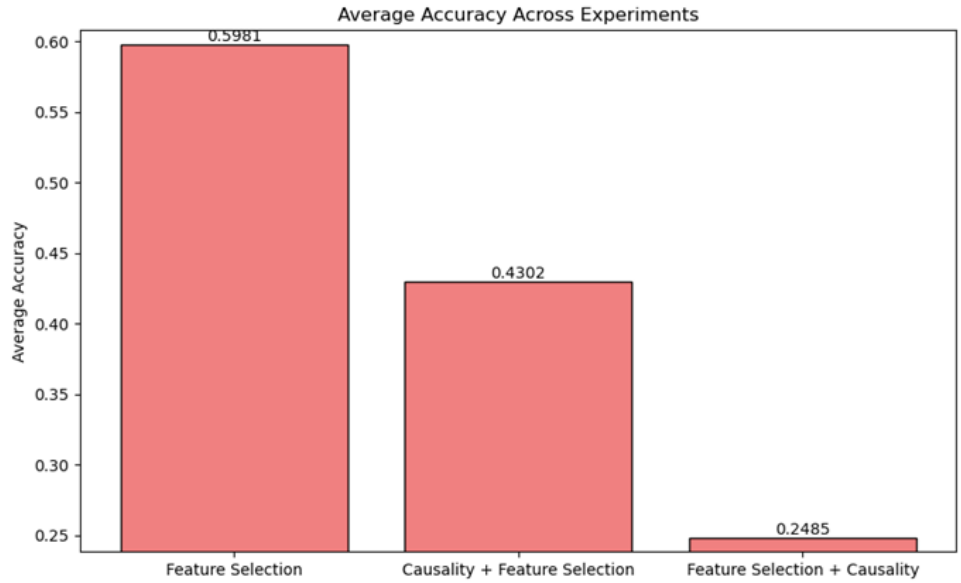

Figure 4: The average accuracy across three different experiments conducted on the CMC dataset.

This highlights the challenge of relying solely on causality to determine feature relevance, especially when stringent thresholds lead to no features being selected, thus restricting the potential for any subsequent analytical enhancements through feature selection methods.

The right plot displays the third experiment, which involved applying feature selection methods first and then causality analysis. This sequence led to lower accuracy compared to the first experiment. Additionally, in some cases after applying causality, no features achieved a causal estimate greater than 0.1, resulting in an empty bar on the plot because the evaluation was halted; it's not possible to evaluate the accuracy of zero features.

The diminished results observed in the second and third experiments, where causality was used either before or after feature selection, were consistent across all four benchmarks tested. Since the incorporation of causality significantly reduced accuracy across all benchmarks, I will focus exclusively on examining the accuracy of the first experiment across all benchmarks.

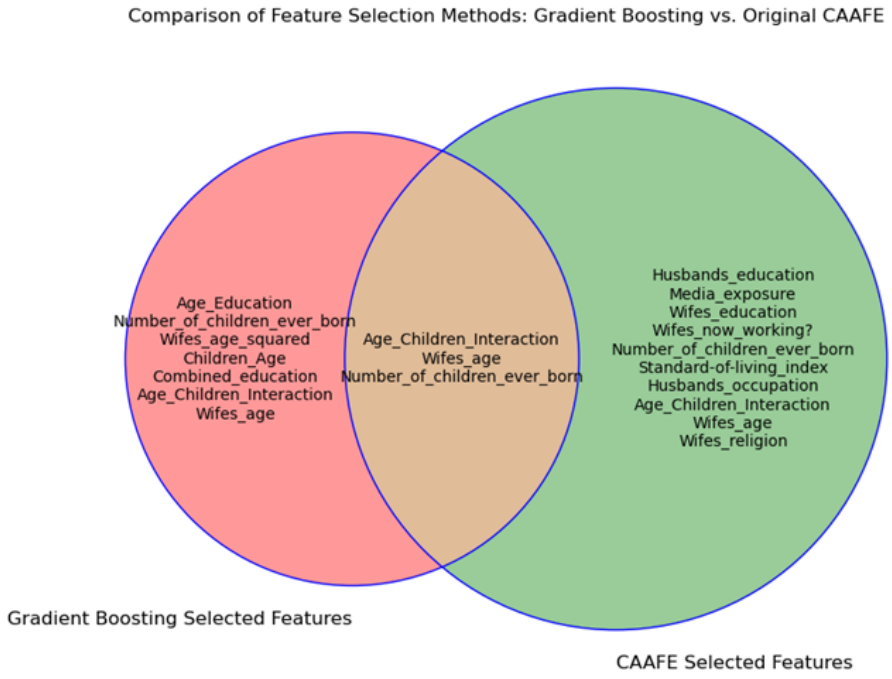

Figure 5: Compares the features selected by Gradient Boosting and Original CAAFE, illustrating both unique and shared feature selections between the two methods for feature selection.

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	Best Feature Selection Method	Feature Selection Accuracy	CAAFE (Gemini-2.0-flash) A
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

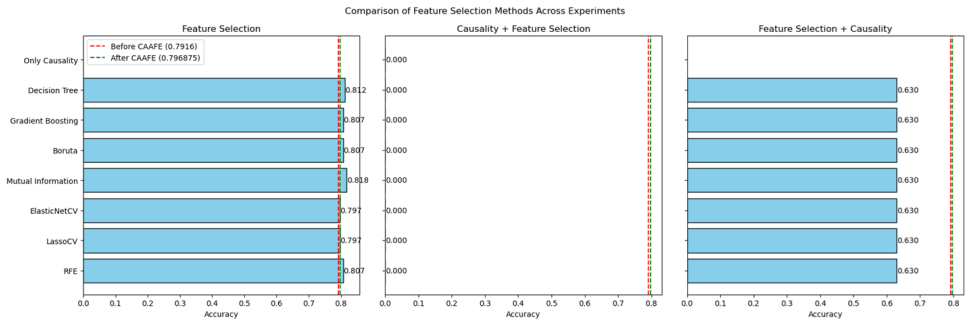

Figure 6: The three panels from left to right illustrate the number of features selected in Diabetes dataset, using different methods.

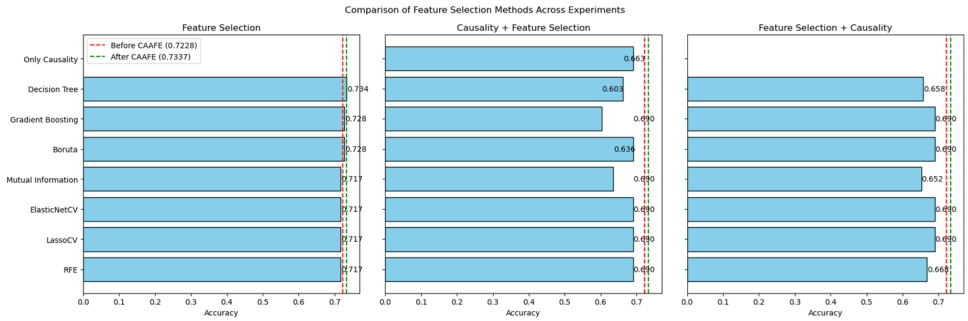

Figure 7: TThe three panels from left to right illustrate the number of features selected in Eucalyptus dataset, using different methods.

4 Related Work

4.1 Large Language Models for Automated Data Science: Introducing CAAFE for Context-Aware Automated Feature Engineering

This work introduced the use of Large Language Models (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

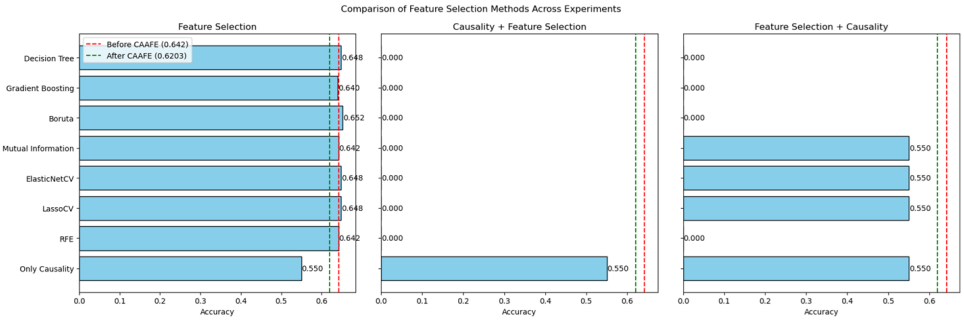

Figure 8: TThe three panels from left to right illustrate the number of features selected in Airlines dataset, using different methods.

they demonstrate accuracy improvements. We extend their framework by applying both feature selection and causality analysis to all generated features. Our methodology was indeed inspired by their initial use of LLMs for feature creation.

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. In contrast, our method uses a modified version of the CAAFE code for feature generation and introduces an iterative process where feature improvement is continuously sought from the LLM, utilizing more advanced models. This iterative enhancement inspired us to manipulate the generated features more extensively.

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, thereby enriching the potential for uncovering meaningful patterns. 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 Large Language Models (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.

This raises an important question encountered during the research: Should we prioritize features that ensure high model accuracy, or should we opt for features with high causal relevance that might lead to better generalization on new or larger datasets?

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.