DAFEL:Domain-Aware Feature Engineering with LLMs for Supervised Learning: An Automated Framework for Classification and Regression Tasks

- Abhirup Ray - MDE2024004

Under the supervision of –
Dr. Sonali Agarwal

Table of contents

01 Introduction

02 Objectives

03 Research Gaps

04 Proposed Approach to fill the research gaps

05 Dataset Description

06 Overview of DAFEL Framework

07 Methodology

08 Comparison Based Results

09 Future Work

10 References Work

Introduction

- Feature engineering is the process of transforming raw data into meaningful and informative features that can be used to improve the performance of machine learning models.

- It involves selecting, creating, transforming, and extracting features from the data to make it more suitable for modeling.

- Feature engineering is a critical step in the machine learning pipeline because the quality and relevance of the features directly impact the model's ability to learn and make accurate predictions.

- Efficient feature engineering often requires in-depth domain knowledge because the process involves understanding the context, relationships, and nuances of the data.

Introduction

- With the advent of Large Language Models (LLMs), this process has become more powerful and efficient, as LLMs can extract rich semantic representations from unstructured data like text and images.

- By leveraging LLMs, we can generate embeddings, capture cross-modal relationships, and create unique feature transformations that improve model performance.

- By leveraging LLMs, practitioners can automate and enhance feature engineering tasks, reducing manual effort and improving model performance.

Objectives

- To introduce a benchmark that evaluates the ability of Large Language Models (LLMs) to perform feature engineering, a critical and knowledge-intensive task in data science, by generating code for feature transformation that transforms datasets to improve machine learning model performance.

- To show that LLMs can effectively assist in feature engineering tasks, reducing the time and expertise required for feature selection and feature transformation, thereby enhancing the efficiency of data science workflows.

Research Gaps

- Existing benchmarks (e.g., MMLU, HumanEval) often evaluate LLMs on **isolated skills** (e.g., language understanding, code generation, or mathematical reasoning) rather than assessing their ability to **integrate multiple skills** (e.g., domain knowledge, reasoning, and code generation) in real-world tasks.

- Many benchmarks do not adequately evaluate LLM's ability to apply domain-specific knowledge (e.g., healthcare, finance) to solve problems. Feature engineering, for example, often requires deep domain expertise to create meaningful features.

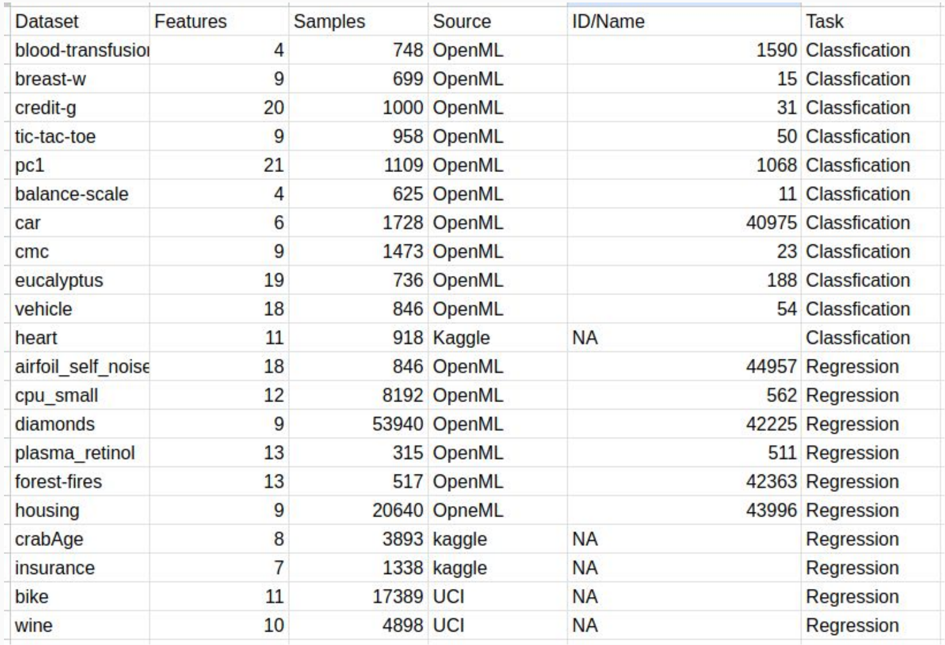
- Existing Benchmarks doesn't specifically focus on iterative methods of feature selection and transformation.

Proposed Approach to fill the research gap

- Create feature engineering as a benchmark that evaluate LLM’s ability to integrate multiple skills (e.g., domain knowledge, reasoning, and code generation) in real-world tasks such as feature engineering.

- Evaluate Large Language Models on tabular data and how efficiently they can perform feature engineering for supervised learning.

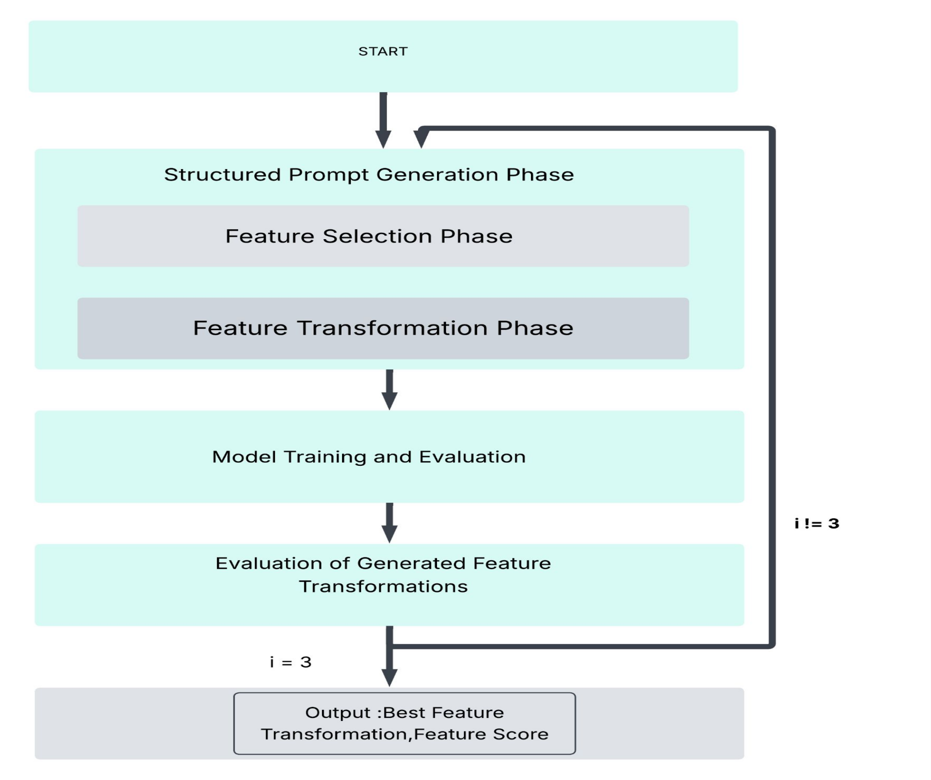
Dataset Description



Dataset	Features	Samples	Source	ID/Name	Task
blood-transfusion	4	748	OpenML	1590	Classfication
breast-w	9	699	OpenML	15	Classfication
credit-g	20	1000	OpenML	31	Classfication
tic-tac-toe	9	958	OpenML	50	Classfication
pc1	21	1109	OpenML	1068	Classfication
balance-scale	4	625	OpenML	11	Classfication
car	6	1728	OpenML	40975	Classfication
cmc	9	1473	OpenML	23	Classfication
eucalyptus	19	736	OpenML	188	Classfication
vehicle	18	846	OpenML	54	Classfication
heart	11	918	Kaggle	NA	Classfication
airfoil_self_noise	18	846	OpenML	44957	Regression
cpu_small	12	8192	OpenML	562	Regression
diamonds	9	53940	OpenML	42225	Regression
plasma_retinol	13	315	OpenML	511	Regression
forest-fires	13	517	OpenML	42363	Regression
housing	9	20640	OpneML	43996	Regression
crabAge	8	3893	kaggle	NA	Regression
insurance	7	1338	kaggle	NA	Regression
bike	11	17389	UCI	NA	Regression
wine	10	4898	UCI	NA	Regression

This table 1 describes the diverse collection of datasets spanning two major categories: (1)classification, (2) regression problems used in this study. The datasets were primarily sourced from established platforms, including OpenML (Vanschoren et al., 2014; Feurer et al., 2021), UCI (Asuncion et al., 2007), and Kaggle.Datasets with descriptive feature names are selected , excluding those with merely numerical identifiers.

Overview of DAFEL Framework


START
Structured Prompt Generation Phase
Feature Selection Phase
Feature Transformation Phase
Model Training and Evaluation
Evaluation of Generated Feature Transformations
i != 3
i = 3
Output :Best Feature Transformation,Feature Score

Methodology

- **Prompt Generation Phase**
  - **Feature Selection Phase:** Feature Selection happens after analysing the following factors:

    a.Selection based on Correlation with target variable (keep if |r| > 0.15)

    b. Domain knowledge (e.g., "carat" matters for diamond prices)

    c. Variance (drop if >95% same value)

    d. Drop redundant features.
  - **Feature Transformation Phase:** Initially in the prompt generation phase, we give prompts for each of the benchmark datasets for feature transformation and give some example feature transformations for each of the datasets and generated feature transformations should occur following a certain type likewise :"**lambda df: df.assign(volume=df['x'] * df['y'] * df['z'])",**

    "lambda df: df.assign(carat_per_depth=df['carat'] / df['depth'])"
- **Model Training and Evaluation**
  - For Regression tasks, we trained the transformed dataset into a randomforest regressor and for classification tasks into a randomforest classifier.
  - Computed an accuracy score for classification datasets and NRMSE for regression datasets and then out of three best transformation in each iteration, we will choose the best transformation.

- **Evaluation of Generated Feature TransformaTions:** In each iteration,the LLM is generating three transformations and the transformed features are then appended to a copy of the original data and then each of the new copy of dataframe is evaluated using randomForestRegressor() and among the three transformations whichever has got least Normalised Mean Square Error score is stored in a memory buffer(in case of regression) and highest accuracy in case of classification. The transformed feature with least Normalised Mean Square Error is appended with the original training and testing data and thereby used for future iterations.After T iterations we are choosing the feature with the best score i.e the Normalised Mean Square Error or the accuracy along with the corresponding feature transformation .

Comparison Based Results

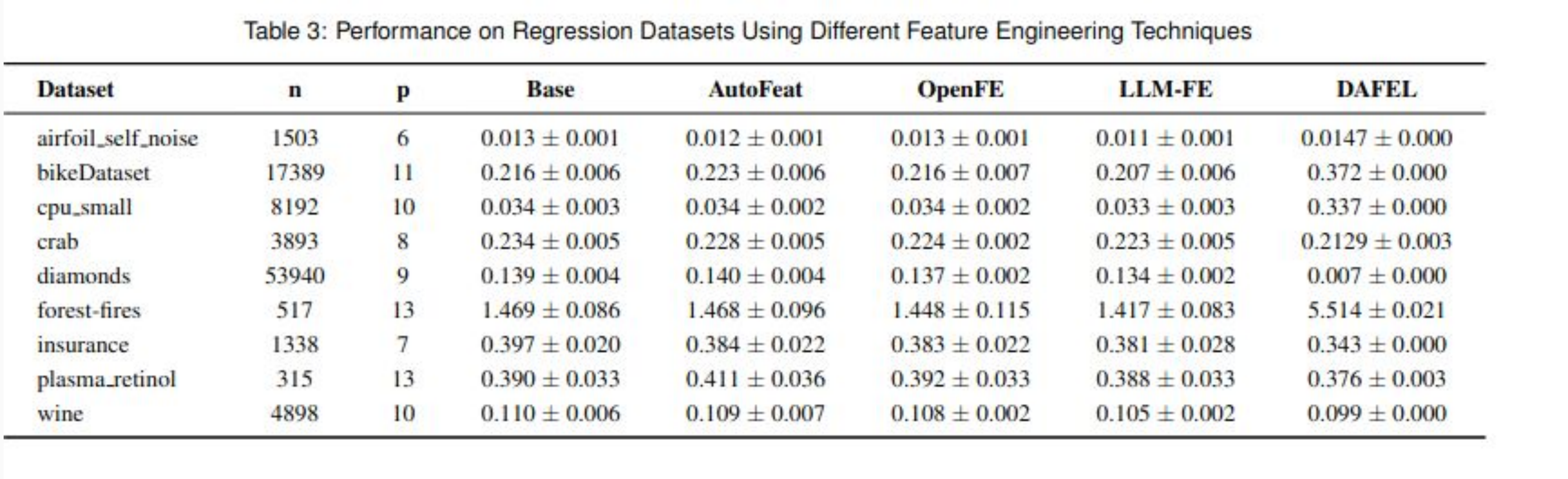


Table 3: Performance on Regression Datasets Using Different Feature Engineering Techniques

Dataset	n	p	Base	AutoFeat	OpenFE	LLM-FE	DAFEL
airfoil_self_noise	1503	6	0.013 ± 0.001	0.012 ± 0.001	0.013 ± 0.001	0.011 ± 0.001	0.0147 ± 0.000
bikeDataset	17389	11	0.216 ± 0.006	0.223 ± 0.006	0.216 ± 0.007	0.207 ± 0.006	0.372 ± 0.000
cpu_small	8192	10	0.034 ± 0.003	0.034 ± 0.002	0.034 ± 0.002	0.033 ± 0.003	0.337 ± 0.000
crab	3893	8	0.234 ± 0.005	0.228 ± 0.005	0.224 ± 0.002	0.223 ± 0.005	0.2129 ± 0.003
diamonds	53940	9	0.139 ± 0.004	0.140 ± 0.004	0.137 ± 0.002	0.134 ± 0.002	0.007 ± 0.000
forest-fires	517	13	1.469 ± 0.086	1.468 ± 0.096	1.448 ± 0.115	1.417 ± 0.083	5.514 ± 0.021
insurance	1338	7	0.397 ± 0.020	0.384 ± 0.022	0.383 ± 0.022	0.381 ± 0.028	0.343 ± 0.000
plasma_retinol	315	13	0.390 ± 0.033	0.411 ± 0.036	0.392 ± 0.033	0.388 ± 0.033	0.376 ± 0.003
wine	4898	10	0.110 ± 0.006	0.109 ± 0.007	0.108 ± 0.002	0.105 ± 0.002	0.099 ± 0.000

Performance of DAFEL on Regression Tasks with existing models

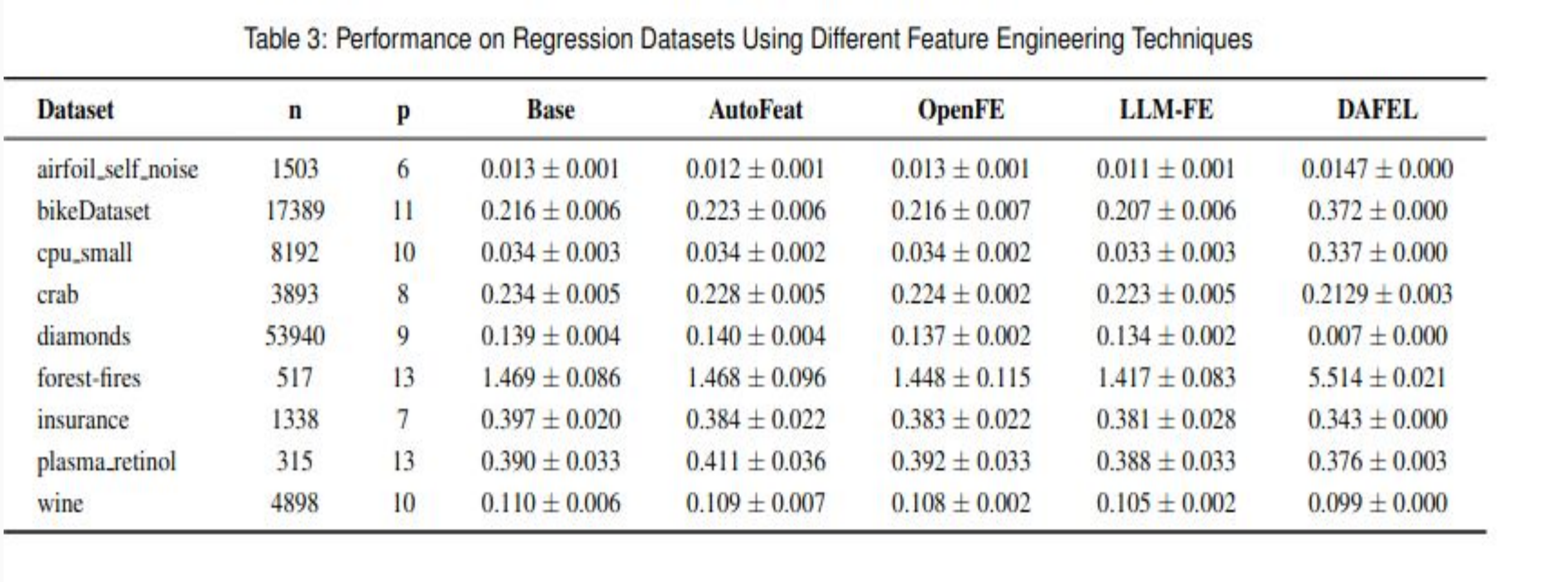


Table 3: Performance on Regression Datasets Using Different Feature Engineering Techniques

Dataset	n	p	Base	AutoFeat	OpenFE	LLM-FE	DAFEL
airfoil_self_noise	1503	6	0.013 ± 0.001	0.012 ± 0.001	0.013 ± 0.001	0.011 ± 0.001	0.0147 ± 0.000
bikeDataset	17389	11	0.216 ± 0.006	0.223 ± 0.006	0.216 ± 0.007	0.207 ± 0.006	0.372 ± 0.000
cpu_small	8192	10	0.034 ± 0.003	0.034 ± 0.002	0.034 ± 0.002	0.033 ± 0.003	0.337 ± 0.000
crab	3893	8	0.234 ± 0.005	0.228 ± 0.005	0.224 ± 0.002	0.223 ± 0.005	0.2129 ± 0.003
diamonds	53940	9	0.139 ± 0.004	0.140 ± 0.004	0.137 ± 0.002	0.134 ± 0.002	0.007 ± 0.000
forest-fires	517	13	1.469 ± 0.086	1.468 ± 0.096	1.448 ± 0.115	1.417 ± 0.083	5.514 ± 0.021
insurance	1338	7	0.397 ± 0.020	0.384 ± 0.022	0.383 ± 0.022	0.381 ± 0.028	0.343 ± 0.000
plasma_retinol	315	13	0.390 ± 0.033	0.411 ± 0.036	0.392 ± 0.033	0.388 ± 0.033	0.376 ± 0.003
wine	4898	10	0.110 ± 0.006	0.109 ± 0.007	0.108 ± 0.002	0.105 ± 0.002	0.099 ± 0.000

Performance of DAFEL on Classification Tasks with existing models

Future Work

- The proposed method can be evaluated using different classification and regression algorithms.

- Explainable AI can be applied to analyse the codes generated from the feature engineering done by Large Language Models and evaluate the trustworthiness of the codes generated by LLM's,

References

- Khurana, U., Turaga, D., Samulowitz, H., Parthasrathy, S.: Cognito: automated feature engineering for supervised learning. In: 2016 IEEE 16th International Conference on Data Mining Workshops (ICDMW), pp. 1304–1307. IEEE (2016)

- Malberg, S., Mosca, E., Groh, G. (2024). FELIX: Automatic and Interpretable Feature Engineering Using LLMs. In: Bifet, A., Davis, J., Krilavičius, T., Kull, M., Ntoutsi, E., Žliobaitė, I. (eds) Machine Learning and Knowledge Discovery in Databases. Research Track. ECML PKDD 2024. Lecture Notes in Computer Science(), vol 14944. Springer, Cham. https://doi.org/10.1007/978-3-031-70359-1_14

- Hollmann, N., Müller, S., & Hutter, F. (2023). Context-Aware Automated Feature Engineering. In *Advances in Neural Information Processing Systems* (Vol. 36, pp. 44753-44775). Curran Associates, Inc.

References

- Hendrycks, D., Burns, C., Basart, S., Zou, A., Mazeika, M., Song, D., & Steinhardt, J. (2021). Measuring Massive Multitask Language Understanding. *Proceedings of the International Conference on Learning Representations (ICLR)*.

- Chen, M., Tworek, J., Jun, H., Yuan, Q., Pinto, H. P. d. O., Kaplan, J., Edwards, H., Burda, Y., Joseph, N., Brockman, G., Ray, A., Puri, R., Krueger, G., Petrov, M., Khlaaf, H., Sastry, G., Mishkin, P., Chan, B., Gray, S., ... Zaremba, W. (2021). Evaluating Large Language Models Trained on Code. *arXiv preprint arXiv:2107.03374*.

Thank You!