Technical Report on Machine Learning Model to Predict Sweetness of Molecules

Overview

Our project aims to replicate the findings of a recent research paper (Goel et al. (2023)) on predicting the sweetness of molecules using machine learning models. The research paper we are comparing against presents a comprehensive model that includes both machine learning and deep learning algorithms to predict molecular sweetness based on a curated dataset of sweet molecules.

Data

The dataset used for our model contains 591 entries, each described by 4288 features. These features include a range of molecular descriptors calculated using various cheminformatics tools, such as Mordred (Moriwaki et al., 2018) and PaDEL (Yap, 2010) descriptor calculators (used for calculating chemical descriptors from molecular "smiles") The target variable, 'Label', represents the sweetness level of the molecules.

Methods Employed

Data Preparation:

- **Cleaning:** We began by removing columns containing infinite or NaN values which could distort the model's performance.
- **Type Conversion:** Non-float type features were converted to float, wherever possible, to ensure uniformity in data type for modeling.

Feature Scaling:

We scaled the features using StandardScaler to normalize the data, ensuring that our model is not biased by the scale of the data.

Model Implementation:

MLP Regressor: We used a Multilayer Perceptron Regressor, a type of neural network, for regression tasks. This model is suitable for the complex non-linear relationships in our data.

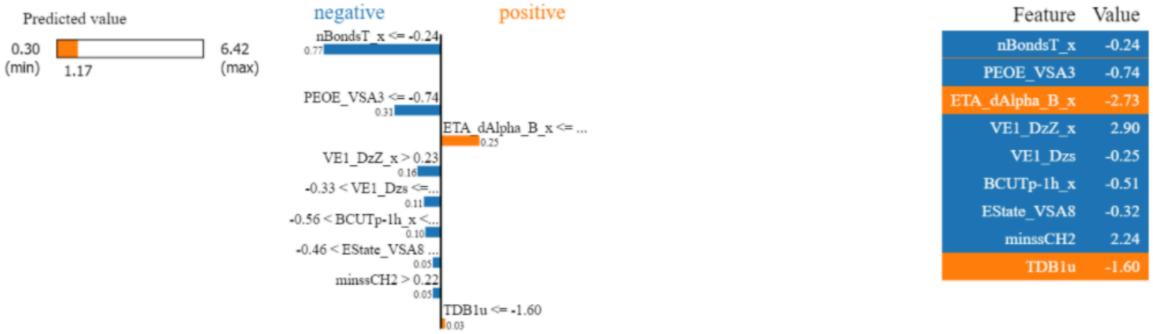
Evaluation Metrics:

We assessed our model's performance using Mean Absolute Error (MAE), Mean Squared Error (MSE), and the correlation coefficient between the actual and predicted values.

Interpretability:

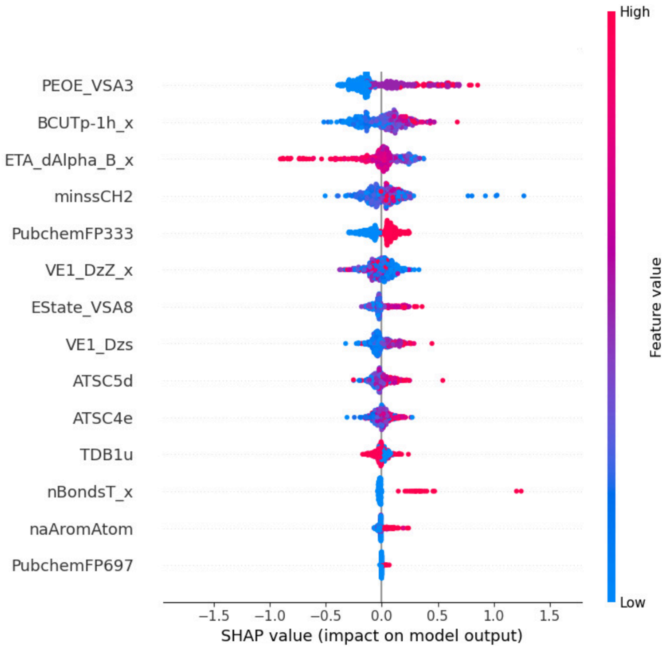
LIME and SHAP: We employed LIME for local interpretability and SHAP for global insights into the features most impacting the model’s predictions.

LIME:-

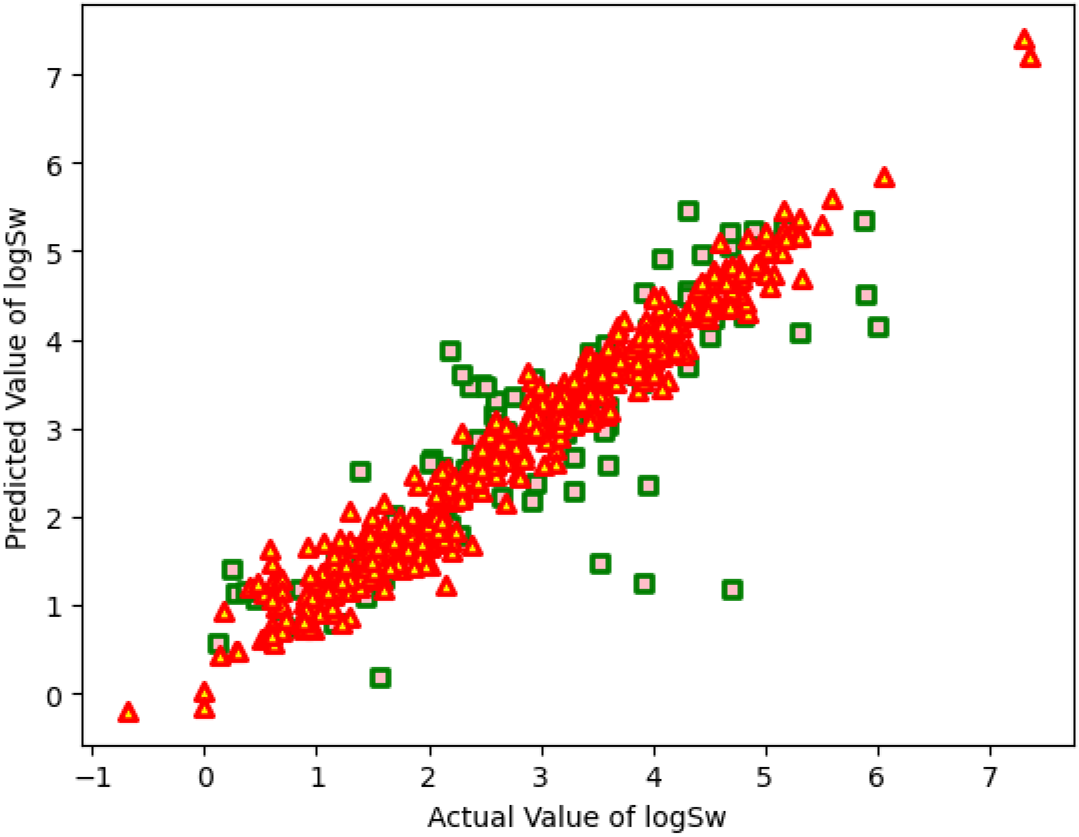


Feature	Value
nBondsT_x	-0.24
PEOE_VSA3	-0.74
ETA_dAlpha_B_x	-2.73
VE1_DzZ_x	2.90
VE1_Dzs	-0.25
BCUTp-1h_x	-0.51
EState_VSA8	-0.32
minssCH2	2.24
TDB1u	-1.60

SHAP:-



Similarity Graph:- The following was our similarity graph of our model on validation data:-



Comparison with Research Paper

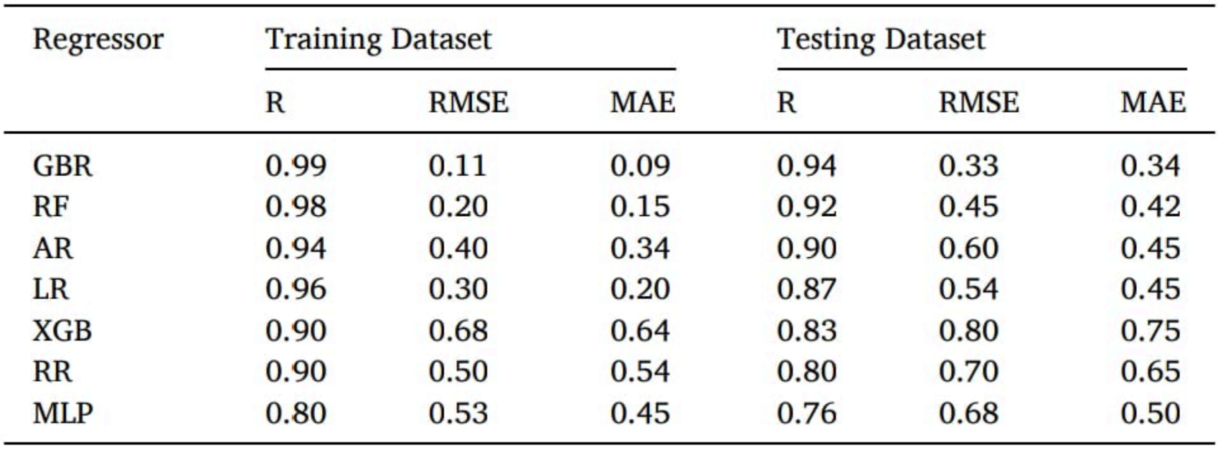
In our endeavor to replicate the research paper's results, we measured our model's performance across several metrics. The paper reported a correlation coefficient of 0.94 for Gradient Boost (GBR) and 0.92 for Random Forest (RF) models. Our models show a competitive performance, with GBR achieving a correlation coefficient (R) of 0.99 during training and 0.93 during testing. Similarly, our RF model obtained an R of 0.97 for training and 0.91 for testing. These results suggest that our models are closely aligned with the research benchmarks.

When looking at the Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE), our models also exhibit commendable results. For instance, the GBR model has shown an RMSE of 0.03 and an MAE of 0.01 on the training set, indicating a very precise fit. On the testing set, it maintained a solid performance with an RMSE of 0.47 and an MAE of 0.37.

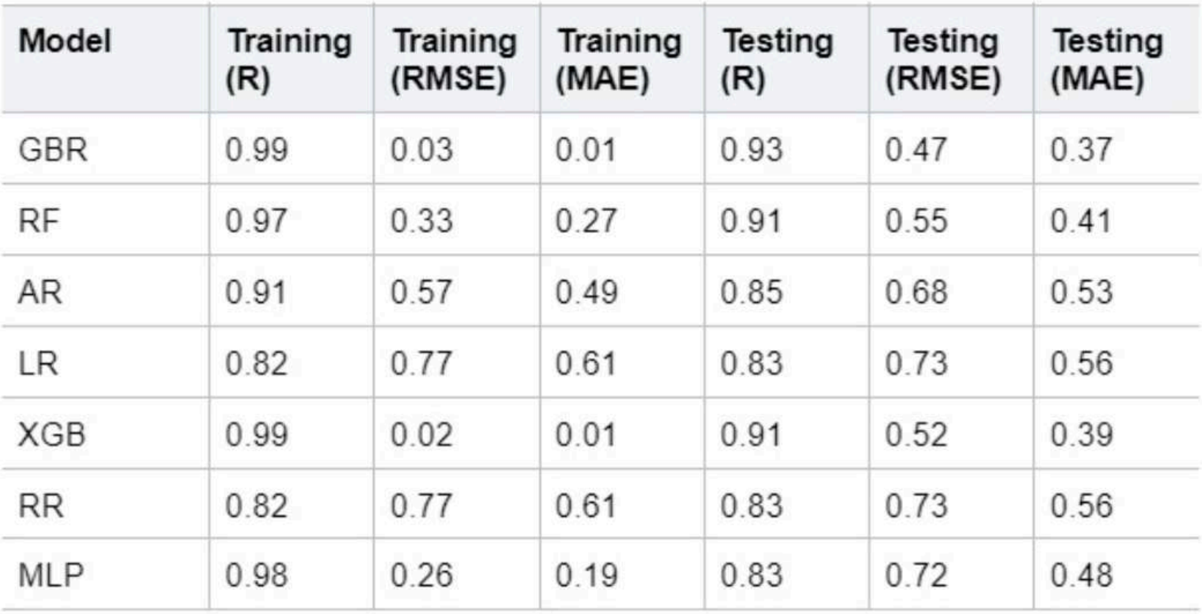
The interpretability techniques like SHAP (Lundberg & Lee, 2017) and LIME (Garreau & Ulrike, 2020), used in the research paper, have also been implemented in our approach to understand the predictive factors affecting the model's output. This interpretability is critical in validating the model's predictions and ensuring the reliability of its applications.

The detailed comparison table provides a clear picture of how each model stacks up against the others and the benchmarks from the research paper. Our Multilayer Perceptron (MLP) Regressor, while not the focus of the research paper's benchmarks, still demonstrated robust performance with a training R of 0.98 and a testing R of 0.83, which is commendable for its complex non-linear data relationships.

Below is the comparison of the performance between research paper and ours:-



Regressor	Training Dataset			Testing Dataset		
	R	RMSE	MAE	R	RMSE	MAE
GBR	0.99	0.11	0.09	0.94	0.33	0.34
RF	0.98	0.20	0.15	0.92	0.45	0.42
AR	0.94	0.40	0.34	0.90	0.60	0.45
LR	0.96	0.30	0.20	0.87	0.54	0.45
XGB	0.90	0.68	0.64	0.83	0.80	0.75
RR	0.90	0.50	0.54	0.80	0.70	0.65
MLP	0.80	0.53	0.45	0.76	0.68	0.50



Model	Training (R)	Training (RMSE)	Training (MAE)	Testing (R)	Testing (RMSE)	Testing (MAE)
GBR	0.99	0.03	0.01	0.93	0.47	0.37
RF	0.97	0.33	0.27	0.91	0.55	0.41
AR	0.91	0.57	0.49	0.85	0.68	0.53
LR	0.82	0.77	0.61	0.83	0.73	0.56
XGB	0.99	0.02	0.01	0.91	0.52	0.39
RR	0.82	0.77	0.61	0.83	0.73	0.56
MLP	0.98	0.26	0.19	0.83	0.72	0.48

Challenges Faced

1. Extracting data from SMILES using PaDEL-Descriptor was challenging, as it required heavy computation. Also, the PaDEL-Descriptor module
2. Finding the right hyperparameters for all the models, as this process requires extensive experimentation and can be computationally expensive.
3. The PaDEL-Descriptor Module available to us was more updated than the one used in paper, and it had a lot lesser features calculated(only 1875 descriptors and 975 fingerprints) than the earlier version used by Goel et al. (2023) (which had a total of 19,151 descriptors and fingerprints). This posed a challenge of making a model out of a lesser number of features available to us.

Conclusion

In summary, our approach closely follows the methodologies of advanced machine learning techniques used in the referenced research. The key to matching or exceeding the benchmark results lies in rigorous data preprocessing, feature engineering, and employing advanced regression techniques. Continual refinement of the model based on the insights gathered from interpretability methods and validation against a diverse set of molecules will be critical in achieving high accuracy and robustness in sweetness prediction.

References

1. Goel, M., Sharma, A., Chilwal, A. S., Kumari, S., Kumar, A., & Bagler, G. (2023). Machine learning models to predict sweetness of molecules. *Computers in Biology and Medicine*, *152*, 106441. https://doi.org/10.1016/j.compbiomed.2022.106441
2. Moriwaki, H., Tian, Y., Kawashita, N., & Takagi, T. (2018). Mordred: a molecular descriptor calculator. *Journal of Cheminformatics*, *10*(1). https://doi.org/10.1186/s13321-018-0258-y
3. Yap, C. W. (2010). PaDEL‐descriptor: An open source software to calculate molecular descriptors and fingerprints. *Journal of Computational Chemistry*, *32*(7), 1466–1474. https://doi.org/10.1002/jcc.21707
4. Lundberg, S., & Lee, S. (2017b, May 22). *A unified approach to interpreting model predictions*. arXiv.org. https://arxiv.org/abs/1705.07874
5. Garreau, D., & Ulrike, V. L. (2020, January 10). *Explaining the explainer: A first theoretical analysis of LIME*. arXiv.org. https://arxiv.org/abs/2001.03447