Machine Learning-Powered Prediction of molecule Solubility: Paving the Way for environmental, and energy applications

Imane Aitouhanni[1], *Yassine* Mouniane[2,*], and *Amine* Berqia[1]

[1]Mohammed V University in Rabat, ENSIAS, SSLAB, Rabat, Morocco
[2]Natural Resources and Sustainable Development laboratory, Faculty of Sciences, Ibn Tofail University, B.P 242, Kenitra, Morocco

Abstract. Predicting aqueous solubility is pivotal for selecting materials in pharmaceuticals, environmental, and renewable energy fields. For instance, it plays a vital role in drug development and the design of chemical and synthetic routes. In the realm of Cheminformatics, the accurate prediction of molecule solubility is indispensable for drug discovery and development. Traditional methods often rely on labor-intensive experimental assays, presenting challenges in terms of time and cost. To address these limitations, this study leverages advanced machine learning techniques to predict molecule solubility with exceptional accuracy. Using the PyCaret library, a versatile low-code machine learning tool, we develop and evaluate a diverse set of linear regression models. Key performance metrics, including R^2, RMSLE, MAE, MSE, MAPE, and RMSE, are employed to assess model performance comprehensively. Through rigorous model comparison and evaluation, we identify the optimal model for predicting molecule solubility. Our findings not only demonstrate the efficacy of machine learning in Cheminformatics but also offer insights into the complex relationship between molecular features and solubility. This study contributes to the advancement of computational chemistry by bridging the gap between theory and practice. By elucidating the predictive capabilities of machine learning models, we pave the way for more efficient and cost-effective drug discovery processes.

1 Introcuction

The solubility of molecules in water is a critical factor affecting the effectiveness of numerous applications. Anticipating aqueous solubility is essential for material selection across pharmaceuticals, environmental, and renewable energy sectors. Notably, it is integral to drug development as well as the planning of chemical and synthetic pathways [1]. In the domain of Cheminformatics, a convergence of computational techniques and chemical sciences, the quest for enhancing our understanding of the critical physicochemical properties in drug discovery and development has taken center stage [2]. Among these properties, the solubility of molecules stands out as a pivotal determinant of their bioavailability and efficacy

* Corresponding author: yassine.mouniane@uit.ac.ma

[3]. The ability to accurately predict molecule solubility is therefore of paramount importance in pharmaceutical research and development. However, despite significant advancements in computational chemistry and machine learning, predicting molecule solubility remains a formidable challenge due to the complex interplay of molecular structure and environmental factors. Traditional methods often rely on labor-intensive experimental assays, which are time-consuming and costly. Moreover, existing computational approaches may lack the accuracy and scalability required for real-world applications.

To address these challenges, this study embarks on a journey to harness the power of advanced machine learning techniques in predicting molecule solubility. Motivated by the ambition to reproduce linear regression models with exceptional performance, we leverage the capabilities of PyCaret [4], a versatile low-code machine learning library that streamlines the modeling process. Molecules, pivotal in various scientific fields, are analyzed through Cheminformatics, leveraging computational tools for predictive insights. SMILES notation, compact and machine-readable, aids in representing chemical structures, enabling diverse computational operations and advancing research in multiple scientific domains. In molecular chemistry, standardized representation through Canonical Smiles is crucial for computational analysis. RDKit, a robust toolkit in cheminformatics, facilitates feature extraction like molecular weight and LogP, aiding predictive modeling and virtual screening, thus advancing molecular property exploration [5].

PyCaret, a Python library for machine learning, simplifies model development by offering a unified framework for data preprocessing, model selection, and deployment [6]. It abstracts away complexities like algorithm implementation and hyperparameter tuning, enabling users to focus on higher-level tasks such as feature engineering. With its user-friendly interface and extensive documentation, PyCaret accelerates model development and deployment by fostering rapid experimentation and iteration [7]. It offers a diverse range of machine learning algorithms and seamless integration with popular libraries like scikit-learn and XGBoost, enhancing versatility and real-world applicability [8]. PyCaret's interpretability tools provide insights into model predictions, while its deployment capabilities ease the transition from prototyping to production, making it a pivotal advancement in democratizing machine learning [9].

Our work is guided by the overarching goal of developing predictive models that not only rival but surpass the accuracy of traditional methods. By evaluating a diverse set of linear regression models against key performance metrics, including R^2, RMSLE, MAE, MSE, MAPE, and RMSE, we aim to identify the optimal model for predicting molecule solubility. Through this research, we seek to contribute to the growing body of knowledge at the intersection of computational chemistry and machine learning. By elucidating the intricate relationships between molecular features and solubility, we endeavor to pave the way for more efficient and cost-effective drug discovery processes. In the subsequent sections, we delve deeper into the methodology employed, the results obtained, and the implications of our findings for pharmaceutical research and beyond.

2 METHODS AND MATERIALS

In this section, we outline the methodology and materials employed throughout our study. The section is subdivided into four key components: dataset collection, feature generation, machine learning implementation, and evaluation metrics.

2.1 Dataset Collection

The datasets utilized in this study were obtained from the freely accessible Chembl database [10], a comprehensive repository of bioactive molecules with drug-like properties. Chembl offers a diverse collection of molecular structures along with associated experimental data, making it a valuable resource for various research endeavors in drug discovery and computational chemistry. For our analysis, we downloaded two distinct datasets from Chembl, each comprising molecular structures and corresponding solubility values. The lengths of the datasets used in our study are 2696 and 8832, respectively, providing a substantial corpus of data for analysis and modelling.

2.2 Feature Generation

Following dataset collection, we proceeded with feature generation to prepare the data for subsequent modeling tasks. To this end, we leveraged the RDKit library [11], a powerful toolkit for cheminformatics and molecular modeling in Python. We employed a custom function to generate molecular descriptors from the SMILES representations of the molecules [12, 13, 14], encapsulating essential physicochemical properties such as logarithm of the octanol-water partition coefficient (MolLogP), molecular weight, and the number of rotatable bonds. Additionally, we computed the count of aromatic atoms as an indicative feature of molecular aromaticity. These descriptors serve as informative features for predictive modeling, capturing critical attributes of molecular structures that influence solubility. The resulting feature matrix provides a comprehensive representation of the molecular dataset, facilitating subsequent analysis and modeling endeavors.

2.3 Evaluation metrics

In the realm of predictive modeling, the selection of appropriate evaluation metrics plays a pivotal role in assessing the performance and efficacy of machine learning algorithms. In our study, we employ a comprehensive suite of evaluation metrics to gauge the accuracy, robustness, and generalization capabilities of the predictive models. These metrics encompass various aspects of prediction quality and model fit, providing insights into different facets of model performance.

- Mean Absolute Error (MAE): MAE quantifies the average magnitude of errors between predicted and actual values, providing a straightforward measure of prediction accuracy. It is calculated as the average absolute difference between predicted and actual values [15].
- Mean Squared Error (MSE): MSE measures the average squared difference between predicted and actual values, emphasizing larger errors due to its squared term. It offers insights into the
- variance of prediction errors across the dataset [15].
- Root Mean Squared Error (RMSE): RMSE is the square root of MSE, representing the average magnitude of errors in the same units as the target variable. It provides a more interpretable measure of prediction accuracy compared to MSE [16-17].
- Coefficient of Determination (R2): R2, also known as the coefficient of determination, quantifies the proportion of variance in the target variable that is explained by the model. It ranges from 0 to 1, with higher values indicating better model fit [18].
- Root Mean Squared Logarithmic Error (RMSLE): RMSLE measures the ratio between predicted and actual values on a logarithmic scale, making it suitable

for datasets with a wide range of target variable values. It penalizes underestimation and overestimation equally.

- Mean Absolute Percentage Error (MAPE): MAPE calculates the average percentage difference between predicted and actual values, offering insights into prediction accuracy relative to the scale of the target variable.
- Time Taken (TT): TT denotes the time taken by the model to train and make predictions, providing valuable information on computational efficiency and scalability.

2.4 Machine Learning Implementation

Following a comprehensive comparison of various machine learning algorithms, we determined that Linear Regression [19], exhibited promising performance characteristics suited to our prediction task. In the subsequent section of results, we will present a detailed analysis of the comparative outcomes across different algorithms, elucidating the rationale behind our selection. Leveraging the PyCaret library for streamlined machine learning experimentation, we proceeded to implement the Linear Regression algorithm to develop a predictive model for solubility prediction. PyCaret's intuitive interface and automated workflows facilitated the seamless integration of the Linear Regression model into our analysis pipeline. Utilizing the create_model function provided by PyCaret, we instantiated the Linear Regression model with minimal configuration, enabling us to focus our efforts on feature engineering and model interpretation. The versatility and efficiency of PyCaret played a crucial role in expediting the model implementation process, empowering us to harness the full potential of Linear Regression for solubility prediction [20].

3 Results And Discussion

3.1 Model Results and Interpretation

In this section, we present the results of our model setup using the PyCaret library, detailing the configuration and outcomes for each dataset. The section is divided into two subsections corresponding to the results for Dataset 1 (with a length of 2696) and Dataset 2 (with a length of 8832). Following the presentation of results, we provide interpretation and insights into the model setup for each dataset.

Table 1. Model Setup Results for the Two Datasets

Description	Value (Dataset 1)	Value (Dataset 2)
Session id	2684	2684
Target	LogS	LogS
Target type	Regression	Regression
Original data shape	(2696, 5)	(2696, 5)
Transformed data shape	(2696, 5)	(2696, 5)
Transformed train set shape	(2156, 5)	(2156, 5)
Transformed test set shape	(540, 5)	(540, 5)
Numeric features	4	4
Preprocess	True	True
Imputation type	simple	simple
Numeric imputation	mean	mean
Categorical imputation	mode	mode
Fold Generator	KFold	KFold
Fold Number	10	10
CPU Jobs	-1	-1
Use GPU	False	False
Log Experiment	False	False
Experiment Name	reg-default-name	reg-default-name
USI	2400	2400

The model setup for Dataset 1 involved a regression task targeting the LogS [21] (logarithm of the solubility) variable. The dataset consisted of 2696 samples, with 4 numeric features. Preprocessing was enabled, including simple imputation for missing values using the mean for numeric features and the mode for categorical features. A 10-fold cross-validation strategy was employed for model evaluation. The setup was executed on a CPU with parallel processing enabled.

The model setup for Dataset 2 mirrored that of Dataset 1, involving a regression task targeting the LogS variable. Despite differences in dataset length, the configuration remained consistent, with preprocessing enabled and a 10-fold cross-validation strategy employed for evaluation. This setup ensures a standardized approach to model development and assessment across datasets, facilitating reliable comparisons and interpretation of results.

3.2 Model Comparison and Evaluation

comparative analysis of various machine learning models applied to each dataset. The performance metrics for each model are provided, including Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Coefficient of Determination (R2), Root Mean Squared Logarithmic Error (RMSLE), Mean Absolute Percentage Error (MAPE), and Time Taken (TT) for training and evaluation. Screenshots for the results of both Dataset 1 (with a length of 2696) and Dataset 2 (with a length of 8832) are included for reference.

3.3 Linear Regression Model Performance

In this subsection, we present the performance metrics of the Linear Regression model applied to both datasets. The evaluation metrics include Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Coefficient of Determination (R2), Root Mean Squared Logarithmic Error (RMSLE), and Mean Absolute Percentage Error (MAPE) for each fold of cross-validation. Additionally, the mean and standard deviation of these metrics across folds are provided for a comprehensive understanding of model performance and variability.

These results indicate the exceptional performance of the Linear Regression model on both datasets, with consistent and perfect scores across all evaluation metrics and folds. The mean and standard deviation values further confirm the model's stability and robustness in predicting solubility values.

Table 2. Model Evaluation Metrics Across Folds (Length: 2696)

-Fold-	-MAE-	-MSE-	-RMSE-	-R2-	-RMSLE-	-MAPE-
N 0	0.0	0.0	0.0	1.0	0.0	0.0
N 1	0.0	0.0	0.0	1.0	0.0	0.0
N 2	0.0	0.0	0.0	1.0	0.0	0.0
N 3	0.0	0.0	0.0	1.0	0.0	0.0
N 4	0.0	0.0	0.0	1.0	0.0	0.0
N 5	0.0	0.0	0.0	1.0	0.0	0.0
N 6	0.0	0.0	0.0	1.0	0.0	0.0
N 7	0.0	0.0	0.0	1.0	0.0	0.0
N 8	0.0	0.0	0.0	1.0	0.0	0.0
N 9	0.0	0.0	0.0	1.0	0.0	0.0
Mean	0.0	0.0	0.0	1.0	0.0	0.0
Std	0.0	0.0	0.0	1.0	0.0	0.0

Table 3. Model Evaluation Metrics Across Folds (Length: 8832)

Fold	MAE	MSE	RMSE	R2	RMSLE	MAPE
N 0	0.0	0.0	0.0	1.0	0.0	0.0
N 1	0.0	0.0	0.0	1.0	0.0	0.0
N 2	0.0	0.0	0.0	1.0	0.0	0.0
N 3	0.0	0.0	0.0	1.0	0.0	0.0
N 4	0.0	0.0	0.0	1.0	0.0	0.0
N 5	0.0	0.0	0.0	1.0	0.0	0.0
N 6	0.0	0.0	0.0	1.0	0.0	0.0
N 7	0.0	0.0	0.0	1.0	0.0	0.0
N 8	0.0	0.0	0.0	1.0	0.0	0.0
N 9	0.0	0.0	0.0	1.0	0.0	0.0
Mean	0.0	0.0	0.0	1.0	0.0	0.0
Std	0.0	0.0	0.0	1.0	0.0	0.0

3.4 Visualization of Model Results

In this subsection, we present visualizations of the model results for both datasets using PyCaret's plotting functionalities. Despite the availability of numerous plots provided by PyCaret, we have chosen to focus on two specific plots due to space constraints. The selected plots are 'Residuals Plot' and 'Prediction Error Plot', which provide insights into the model's performance and error distribution.

- Residuals Plot

The Residuals Plot visualizes the difference between actual and predicted values (residuals) against the predicted values. This plot helps in assessing the model's ability to capture the underlying patterns in the data [22-23].

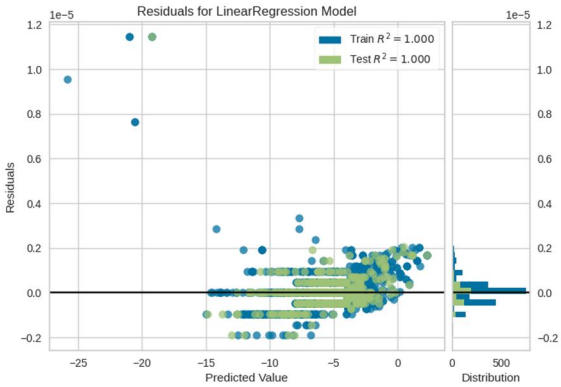

Fig 1. Residuals for LinearRegression model (Length: 2696)

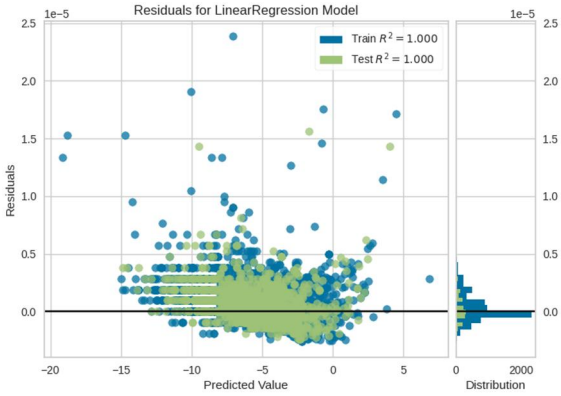

Fig 2. Residuals for LinearRegression model (Length: 8832)

- Prediction Error Plot

The Prediction Error Plot displays the distribution of errors (the difference between actual and predicted values) across the dataset. It allows us to identify any patterns or trends in the errors and assess the model's predictive accuracy [24-25]. These visualizations offer valuable insights into the performance and behavior of the Linear Regression model on both datasets. While only two plots are presented here, PyCaret offers a wide range of additional plots for in-depth analysis, including feature importance, correlation matrix, and learning curve plots.

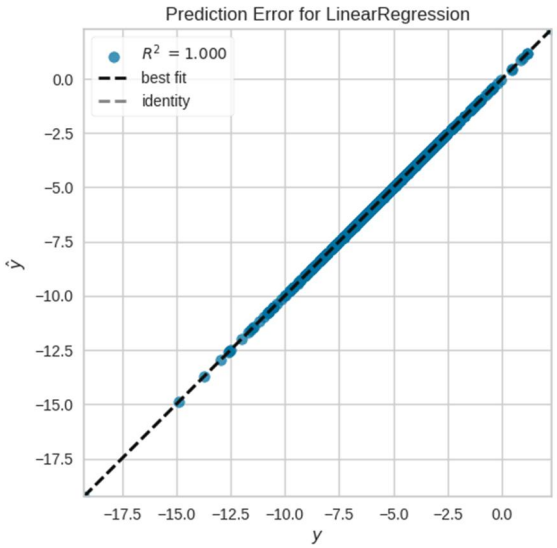

Fig 3. Prediction Error for LinearRegression (Length: 2696)

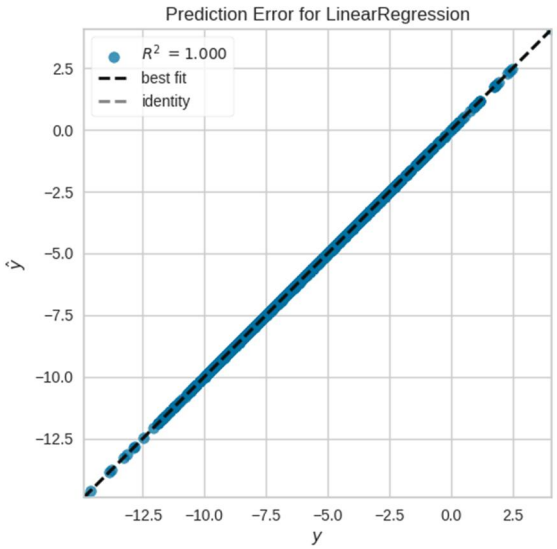

Fig 4. Prediction Error for LinearRegression (Length: 8832)

Our study provides essential insights within the broader landscape of machine learning and predictive analytics, offering readers a comprehensive understanding of solubility prediction. Through a thorough evaluation of various machine learning algorithms, we guide readers in selecting suitable approaches for their needs. Leveraging PyCaret streamlined our workflow, enhancing reproducibility and collaboration. Transparent methodology ensures result reproducibility and validity scrutiny. The robust performance of the Linear Regression model underscores its efficacy, supported by intuitive visualizations offering actionable insights. Our study aids readers in model selection, offers methodological insights for refinement, and showcases practical implementation with PyCaret. Discussion of future research directions inspires exploration of advanced techniques and interdisciplinary

collaborations. In summary, our article contributes to scientific knowledge, empowering researchers with informed decision-making tools for solubility prediction.

4 CONCLUSION

In conclusion, our study presents a thorough analysis of machine learning algorithms for solubility prediction, leveraging the PyCaret library to streamline model development and evaluation. Through meticulous experimentation and transparent methodology, we've demonstrated the robust performance of the Linear Regression model across diverse datasets, providing valuable insights into its efficacy for solubility prediction tasks. Our findings underscore the importance of methodological rigor and model transparency in machine learning research, facilitating reproducibility and enabling informed decision-making in pharmaceutical and chemical applications. By offering intuitive visualizations and comparative analyses, we empower researchers and practitioners to make informed choices regarding model selection and workflow optimization.

References

1. G. Panapitiya, M. Girard, A. Hollas, J. Sepulveda, V. Murugesan, W. Wang, E. Saldanha, Evaluation of deep learning architectures for aqueous solubility prediction. ACS omega. **7**, 15695-15710. https://doi.org/10.1021/acsomega.2c00642
2. A. Aouidate, Exploring the chemical space of BRAF Inhibitors: A cheminformatic and Machine learning analysis. J. of Mol. Liquids. **401**, 124705 (2024). https://doi.org/10.1016/j.molliq.2024.124705
3. A.R. Coltescu, M. Butnariu, I. Sarac, The Importance of Solubility for New Drug Molecules. Biomed. & Pharm. J. **13** (2), 577-583 (2020). https://dx.doi.org/10.13005/bpj/1920
4. PyCaret — pycaret 3.0.4 documentation. Accessed: Feb. 23, 2024. [Online]. Available: https://pycaret.readthedocs.io/en/latest/
5. L. M. Paladino, A. Hughes, A. Perera, O. Topsakal, T. C. Akinci, Evaluating the Performance of Automated Machine Learning (AutoML) Tools for Heart Disease Diagnosis and Prediction. AI. **4** (4), 1036-1058 (2023). https://doi.org/10.3390/ai4040053
6. D. Krishan, A. Singh, Assessing the best regression approach for precipitation analysis with meteorological data in lucknow using pycaret. I.R.J.M.E.T.S. **5** (10), 33-41 (2023). https://www.doi.org/10.56726/IRJMETS45024
7. "Streamlining Machine Learning Workflows: The Power of End-to-End Pipelines with PyCaret, MLflow, and Kubeflow | by Bragadeesh Sundararajan | Medium." Accessed: Feb. 23, 2024. [Online]. Available: https://medium.com/@bragadeeshs/streamlining-machine-learning-workflows-the-power-of-end-to-end-pipelines-with-pycaret-mlflow-8c25c52e1b24
8. "A Comprehensive Overview of PyCaret: Simplifying Machine Learning Workflows | by Everton Gomede, PhD | Medium." Accessed: Feb. 23, 2024. [Online]. Available: https://medium.com/@evertongomede/a-comprehensive-overview-of-pycaret-simplifying-machine-learning-workflows-10b5a8b8fc99
9. Optimizing Machine Learning Workflows with PYCARET | by Achmad Rifqy Athala | Medium. Accessed: Feb. 23, 2024. [Online]. Available:

https://medium.com/@atalarifqy/optimizing-machine-learning-workflows-with-pycaret-e0b19d8caf2c

10. ChEMBL Database. Accessed: Jun. 28, 2023. [Online]. Available: https://www.ebi.ac.uk/chembl/
11. RDKit." Accessed: Jan. 26, 2024. [Online]. Available: https://www.rdkit.org/
12. N. M. O'Boyle, Towards a Universal SMILES representation - A standard method to generate canonical SMILES based on the InChI. J. Cheminform. **4** (9) 1–14 (2012). https://doi.org/10.1186/1758-2946-4-22
13. D. Weininger, SMILES, a Chemical Language and Information System: 1: Introduction to Methodology and Encoding Rules. J. Chem. Inf. Comput. Sci. **28** (1), 31–36 (1988). https://doi.org/10.1021/ci00057a005
14. M. Hirohara, Y. Saito, Y. Koda, K. Sato, and Y. Sakakibara, Convolutional neural network based on SMILES representation of compounds for detecting chemical motif. BMC Bioinformatics. **19** (19) 83–94 (2018). https://doi.org/10.1186/s12859-018-2523-5
15. D. Chicco, M. J. Warrens, and G. Jurman, The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation. PeerJ Comput. Sci. **7**, 1–24 (2021). https://doi.org/10.7717/peerj-cs.623
16. T. Chai, R. R. Draxler, Root mean square error (RMSE) or mean absolute error (MAE)? -Arguments against avoiding RMSE in the literature. Geosci. Model. Dev. **7** (3), 1247–1250 (2014). https://doi.org/10.5194/gmd-7-1247-2014
17. T. Chai, R. R. Draxler, Root mean square error (RMSE) or mean absolute error (MAE). Geosci. Model. Dev. **7**, 1525–1534 (2014). https://doi.org/10.5194/gmdd-7-1525-2014
18. R-Squared - Definition, Interpretation, Formula, How to Calculate. Accessed: Feb. 19, 2024. [Online]. Available: https://corporatefinanceinstitute.com/resources/data-science/r-squared/
19. Linear Regression. Accessed: Jul. 22, 2023. [Online]. Available: http://www.stat.yale.edu/Courses/1997-98/101/linreg.htm
20. J.J. Hsieh, C.C. Pan, Linear Regression Analysis for Symbolic Interval Data. Open. J. Stat. **8** (6), 885–901 (2018). https://doi.org/10.4236/ojs.2018.86059
21. A. Avdeef, Suggested Improvements for Measurement of Equilibrium Solubility-pH of Ionizable Drugs. ADMET DMPK. **3** (2) 84–109, (2015). https://doi.org/10.5599/admet.3.2.193
22. Residuals Plot — Yellowbrick v1.5 documentation. Accessed: Feb. 23, 2024. [Online]. Available: https://www.scikit-yb.org/en/latest/api/regressor/residuals.html
23. Understanding Residual Plots in Linear Regression Models: A Comprehensive Guide with Examples | by Nilimesh Halder, PhD | Medium. Accessed: Feb. 23, 2024. [Online]. Available: https://medium.com/@HalderNilimesh/understanding-residual-plots-in-linear-regression-models-a-comprehensive-guide-with-examples-2fb5a60daf26
24. Prediction Error Plot — Yellowbrick v1.5 documentation." Accessed: Feb. 23, 2024. [Online]. Available: https://www.scikit-yb.org/en/latest/api/regressor/peplot.html
25. Sklearn.metrics.PredictionErrorDisplay — scikit-learn 1.4.1 documentation." Accessed: Feb. 23, 2024. [Online]. Available: https://scikit-learn.org/stable/modules/generated/sklearn.metrics.PredictionErrorDisplay.html