Estimating Aqueous Solubility Directly from Molecular Structure

By:
1.Grandhe Sreesai Vaishnavi (186119)
2.M.Sri Lokesh Reddy (186229)
3.Soumyajit Das (186246)
4.Renu Ravindhar (186143)

Under the guidance of Dr. Kola Anand Kishore

Motivation

- Pharmaceutical, material, physical, and environmental sciences.
- Pharmacokinetic properties
- Computational Models
- This intrigued us to develop 3 regression algorithms to understand the better relation between aqueous solubility and the considered parameters.

Introduction

After extensive literature review, four significant parameters that affect solubility are:

1. Molecular Weight
2. Rotatable bonds
3. Aromatic Proportion
4. clogP

What is logP?

- The logarithm of the 1-octanol/water partition coefficient.
- To assess biological properties relevant to drug action.
- 1-Octanol is a natural choice as the hydrophobic solvent in this respect because of its physicochemical similarity to lipids, its ready availability, and its ease of use.
- clogP is most widely used method for estimation of logP.

How to measure these properties?

- Using Simplified Molecular Input Line Entry System(SMILES) .

- SMILES is a chemical notation system designed for modern chemical information processing.

- SMILES which is simple to write allows rigorous structure specification by use of a very small and natural grammar.

Data preparation and exploratory analysis

Here we perform three operations :

1.Data sourcing: Data used in this project has been sourced from Open Access Database from ACS publications from the work by John S Delaney, Syngenta, Jealott's Hill International Research Centre, Bracknell, UK.

2.Data preparation: We receive data from a Comma Separated File(csv).We remove empty rows and columns . Chemical structures are encoded by SMILES.

To predict LogS (log of the aqueous solubility), we use molecular descriptors: cLogP (Octanol-water partition coefficient), Molecular weight, Number of rotatable bonds, Aromatic proportion.

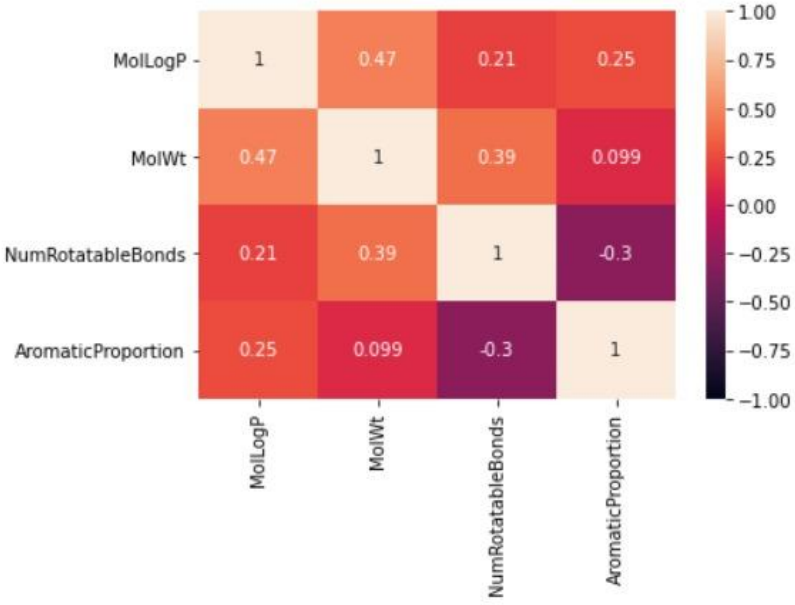
Final Training Dataset

	MolLogP	MolWt	NumRotatableBonds	AromaticProportion
0	2.59540	167.850	0.0	0.000000
1	2.37650	133.405	0.0	0.000000
2	2.59380	167.850	1.0	0.000000
3	2.02890	133.405	1.0	0.000000
4	2.91890	187.375	1.0	0.000000
...	...	...	...	...
1139	1.98820	287.343	8.0	0.000000
1140	3.42130	286.114	2.0	0.333333
1141	3.60960	308.333	4.0	0.695652
1142	2.56214	354.815	3.0	0.521739
1143	2.02164	179.219	1.0	0.461538

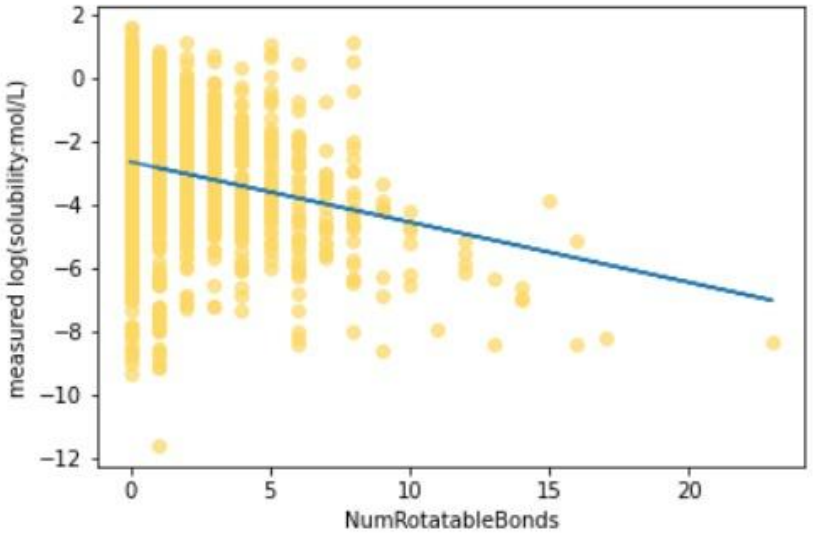
1144 rows × 4 columns

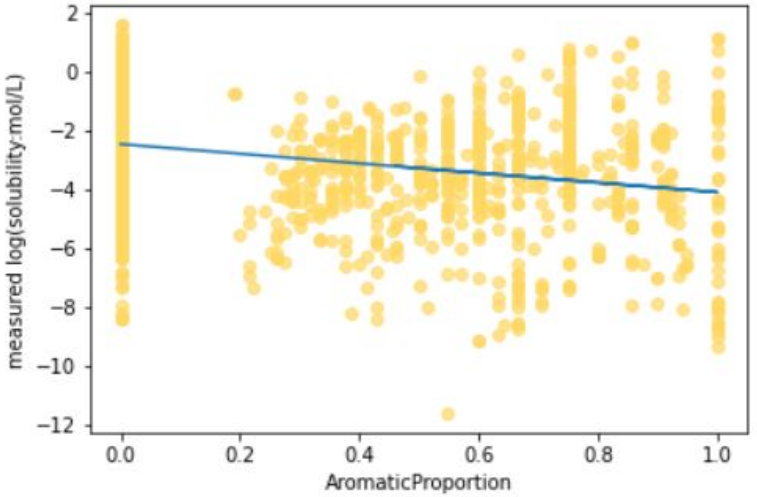
Exploratory Analysis

- For the exploratory analysis we plot some graphs to see correlations among certain descriptors:

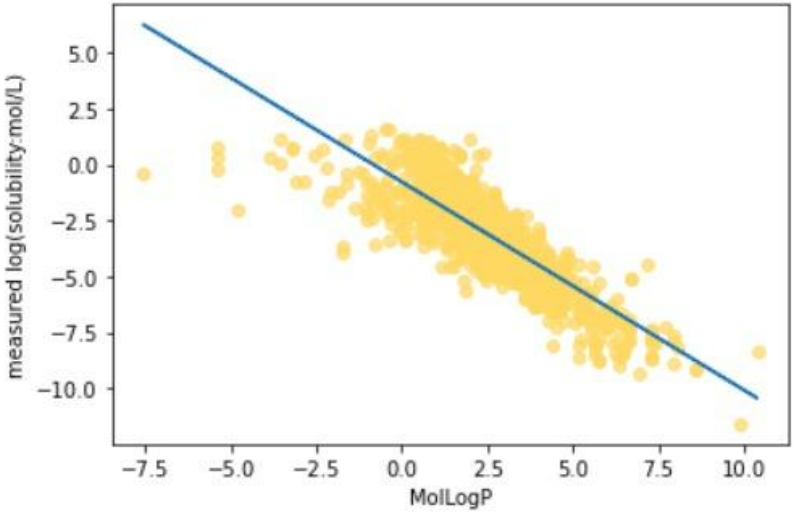


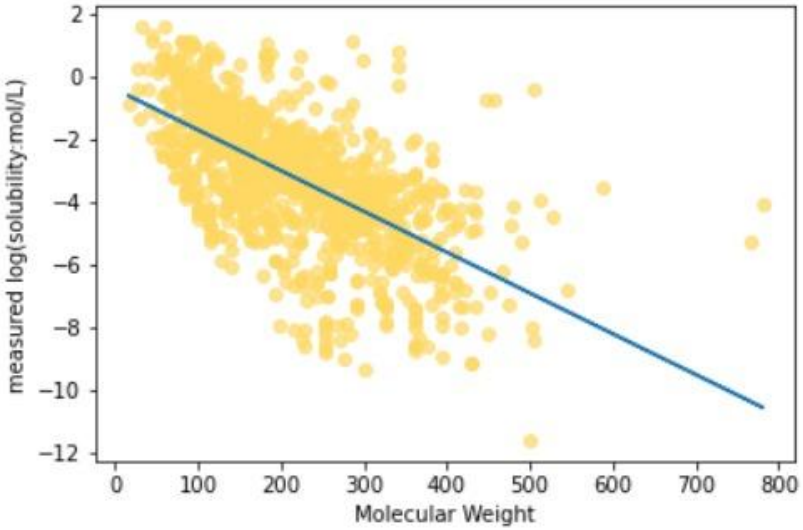
Exploratory Analysis





Exploratory Analysis





Regression: Introduction

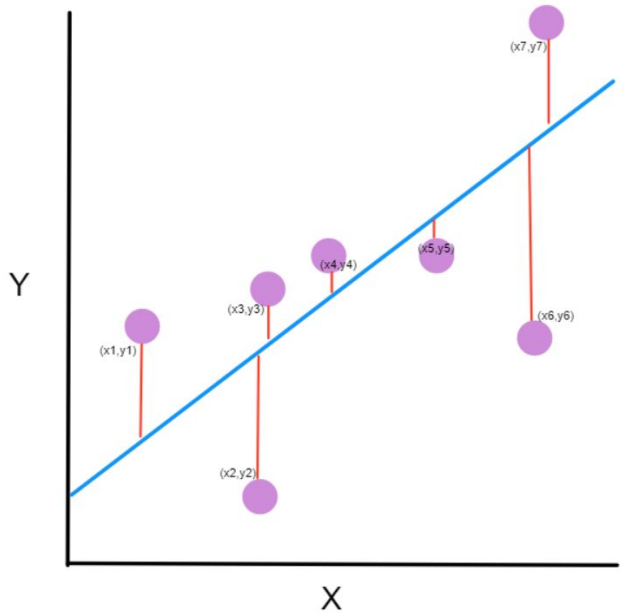
- Regression is a statistical method used that attempts to determine the strength and character of the relationship between one dependent variable (usually denoted by Y) and a series of other variables (known as independent variables).
- Python language

Code- Jupyter IDE

$$y_i = f(p,q,r,s) = w_1 p + w_2 q + w_3 r + w_4 s + c + \epsilon$$

Mean Squared Error(MSE)

- Let us take a simple function $y_i = f(x)$ as shown
- The purple dots - The points on the graph.
- The blue line - Prediction line.
- The red line between each purple point and the prediction line are the errors.
- For each point, we take the y-coordinate of the point, and the y'-coordinate. The y-coordinate is our purple dot. The y' point sits on the line we created.
- $MSE = (1/n)\sum(y_i - y'_i)^2$

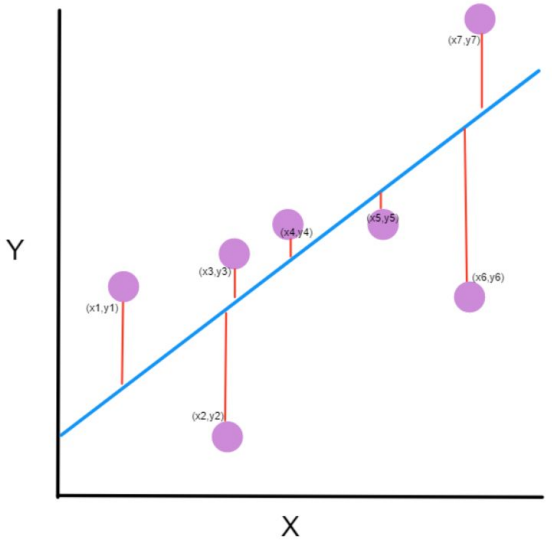


Coefficient of determination (R^2)

- y_m is the mean of the observed data: $y_m = (1/n)\sum y_i$
- The total sum of squares (proportional to the variance of the data):

  $SS_T = \sum(y_i - y_m)^2$
- Now if we consider residuals as $e_i = y_i - y'_i$
- The sum of squares of residuals, also called the residual sum of squares:

  $SS_R = \sum e_I^2$
- Thus we get $R^2 = 1-(SS_R/SS_T)$



Multivariate Linear Regression

- Used to predict values within a continuous range.
- Predicted output is continuous and has a constant slope.
- We make a train-test split- (Training 80%: Testing 20%).

Multivariate Linear Regression

- Tested it on the out-of-sample test set and R2 and MSE values are observed

R^2	MSE
0.75	1.13

- Regression Equation from the test data:

logS= 0.235096 -0.76774926 logP -0.00640718 MW +0.01230929 RB -0.38766774 AP

Ridge Regression

- Type of linear regression that includes an L2 penalty.

- L2 causes shrinking the coefficients for those input variables that do not contribute much to the prediction task. Prevents Overfitting

- We make a train-test split- (Training 80%: Testing 20%).

- Ridge Regression consists of various hyperparameters. For the model we have selected default values.

Ridge Regression

- Tested it on the out-of-sample test set and R2 and MSE values are observed

R^2	MSE
0.74	1.17

- Regression Equation from the test data:

logS= 0.234446 -0.76768984 logP -0.00640935 MW +0.01240782 RB -0.38568153 AP

- Compared to Multivariate Linear Regression model, we find a marginal reduction in R^2 and an increase in MSE values, both of which are undesirable.

Adaboost Regression

- Boosting refers to a class of machine learning ensemble algorithms where models are added sequentially and later models in the sequence correct the predictions made by earlier models.

- Adaboost, expanded to Adaptive Boosting, is called so as it responds adaptively to the errors of the weak hypotheses.

- For the model, here we use Cross Validation techniques to determine the optimal values for the hyperparameters that we need for a better R^2 score.

Adaboost Regression

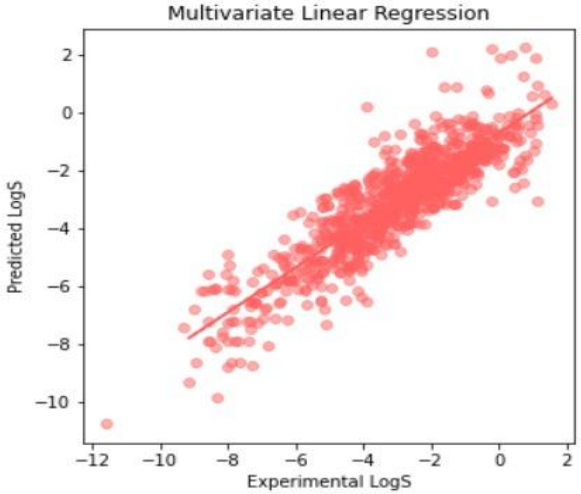
- Tested it on the out-of-sample test set and R^2 and MSE values are observed

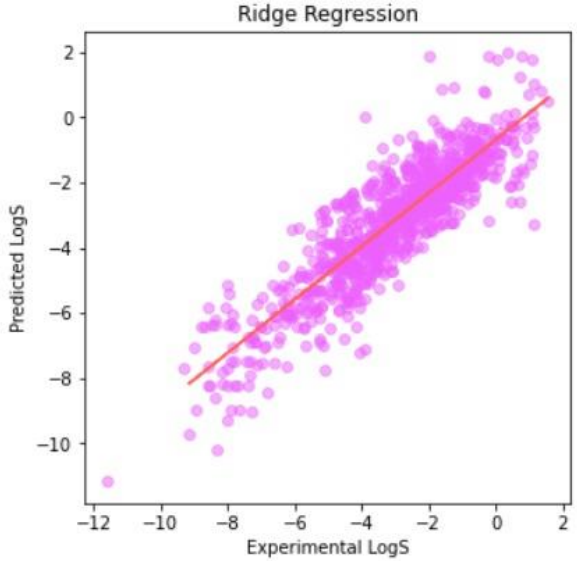
R^2	MSE
0.79	0.95

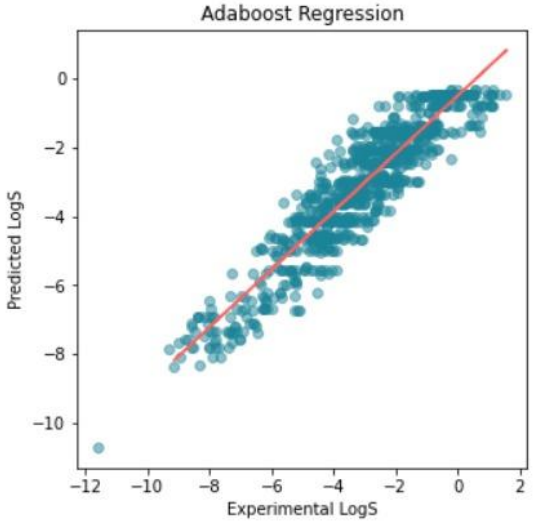
- We get significantly better results with the Adaboosting model with improvements in both out-of-sample R2 and MSE.

Results & Comparison

- We visualise the correlation of the Experimental LogS values with those of the Predicted LogS values in the testing sets by means of the scatter plots for each Regression Model. The Adaboost model is significantly better, with all the plot points much closer to the regression line than in other models indicating higher R^2 and lower MSE values.







Scope for Improvement and Conclusion

- Solubility depends significantly on the Octanol-water partition coefficient.
- We receive satisfactory results from the Adaboost model, when the Hyperparameters are optimized by Cross Validation.
- Moving forward, one can look to further tune the hyperparameters or look to develop models based on other Machine Learning algorithms like XG Boosting.
- Further research may also be done into looking at more molecular properties that may influence the solubility and incorporate them in models.

References

[1]Cheng, A.; Merz, K.M. Prediction of Aqueous Solubility of a Diverse Set of Compounds Using Quantitative Structure-Property Relationships. J. Med. Chem. 2003, 46, 3572-3580.
[2]Butina, D.; Gola, J. M. R. Modeling Aqueous Solubility. J. Chem.Inf. Comput. Sci. 2003, 43, 837-841.
[3]Ghose, A. K.; Viswanadhan, V. N.; Wendoloski, J. J. Prediction of Hydrophobic (Lipophilic) Properties of Small Organic Molecules Using Fragmental Methods: An Analysis of LOGP and CLOGP Methods. J. Phys. Chem. 1998, 102, 3762-3772.
[4]John S. Delaney ESOL: Estimating Aqueous Solubility Directly from Molecular Structure J. Chem. Inf. Comput. Sci. 2004, 44, 1000-1005
[5] DAVID WEININGER SMILES, a Chemical Language and Information System. 1. Introduction to Methodology and Encoding Rules J. Chem. In$ Comput. Sci., Vol. 28, No. I , I988
[6]DAVID WEININGER, ARTHUR WEININGER, and JOSEPH L. WEININGER SMILES. 2.Algorithm for Generation of Unique SMILES Notation J. Chem. Inf. Comput. Sci., Vol. 29, No. 2, 1989
[7]https://www.rdkit.org/docs/Overview.html
[8]Gibbs Y. Kanyongo, Janine Certo, Brown I. Launcelot, Using regression analysis to establish the e environment and reading achievement: A case of Zimbabwe, International Education Journal, 2006, 7(5), 632-641
[9]Yoav Freund, Robert E. Schapire, A decision-theoretic generalization of on-line learning and an application to boosting, In: Vitányi P. (eds) Computational Learning Theory. EuroCOLT 1995. Lecture Notes in Computer Science (Lecture Notes in Artificial Intelligence), vol 904. Springer, Berlin, Heidelberg.
[10]Trevor Hastie (Department of Statistics, Stanford University, Stanford, Calif., U.S.A.), Saharon Rosset (Department of Statistics, Tel Aviv University, Tel Aviv, Israel), Ji Zhu (Department of Statistics, University of Michigan, Ann Arbor, Mich., U.S.A.), Hui Zou (School of Statistics, University of Minnesota, Minneapolis, Minn., U.S.A.), Multi-class AdaBoost, Statistics and Its Interface Volume 2 (2009) Number3, 349-360