CASE STUDY

CHEMICAL INDUSTRY

Sriya.AI
AI ANYWHERE ANYTIME

Objective

- To get the prediction accuracy of experiments with high solubility compounds by Auto-AI and SXI and compare.
- Precision AI^2 using Target SXI based Random Forest trees. Target increase in High solubility compounds is **20%** up from current levels.

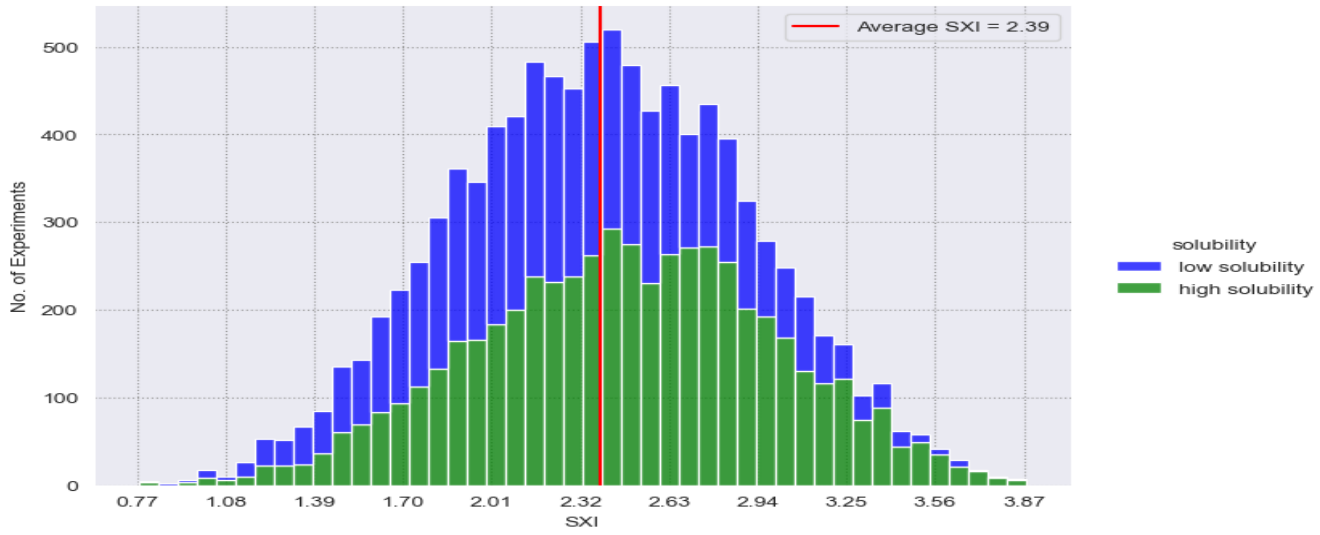
SXI Hypothesis

- SXI is a proxy/surrogate for all features responsible for ensuring high or low solubility compounds in an experiment. The higher the SXI, the better is the number of experiments with high solubility compound and hence increasing SXI score should lead to more experiments with higher solubility compounds.

SXI Definition

- **Sriya Expert Index (SXI):** Dynamic score/index obtained from a proprietary formula consisting of weights from 10 ML algorithms. SXI is a super feature and is a true weighted representative of all important features. Converts a multi-dimensional hard to solve problem into a simpler 2-dimensional solution (problem solved).
- **SCORE + CORRELATE = IMPROVE**

Discussion & Results



1. Exploratory Data Analysis

9982 experiments were distributed to **5505** good and **4477** bad. Good are high solubility compounds and Bad are low solubility compounds. So, **55.15%** is the current high solubility compounds and **44.85%** is solubility compounds.

2. SXI - Exploratory Data Analysis

Current Average SXI is **2.39**. No. of total experiments above 2.39 is **4985** and of these **3145** are of high solubility compounds and **1840** are low solubility compounds. So High solubility compounds (%) are **63.08**% and low solubility compounds are **36.91**%.

Correspondingly No. of total experiments below 2.39 is **4997** and of these **2360** are high and **2637** are low. So High solubility compounds (%) are **47.22**% and low is **52.77**%.

So SXI is a perfect proxy/surrogate for High Solubility Compounds and above average SXI ratio of good outcome is **1.14x** of the overall average and below average SXI ratio of good outcome is **0.85x** of the overall average. So, the increase in SXI leads to an increase in high solubility compounds.

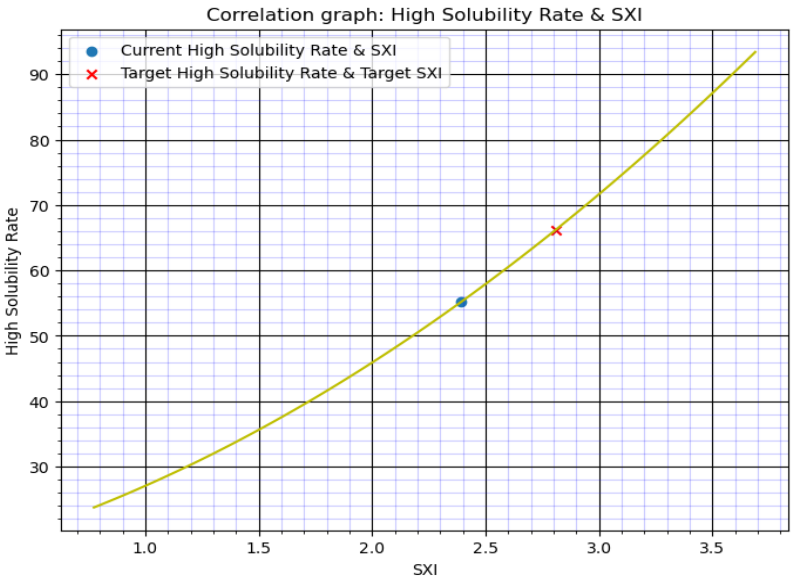
3. Predictive AI

- Auto-AI Prediction accuracy is **97.4%** and the best performing algorithm is **XGBoost**.
- SXI Prediction accuracy of high solubility rate is **97.8%**.
- Ratio of SXI/Auto-AI prediction accuracy is **1.**

4. Precision AI

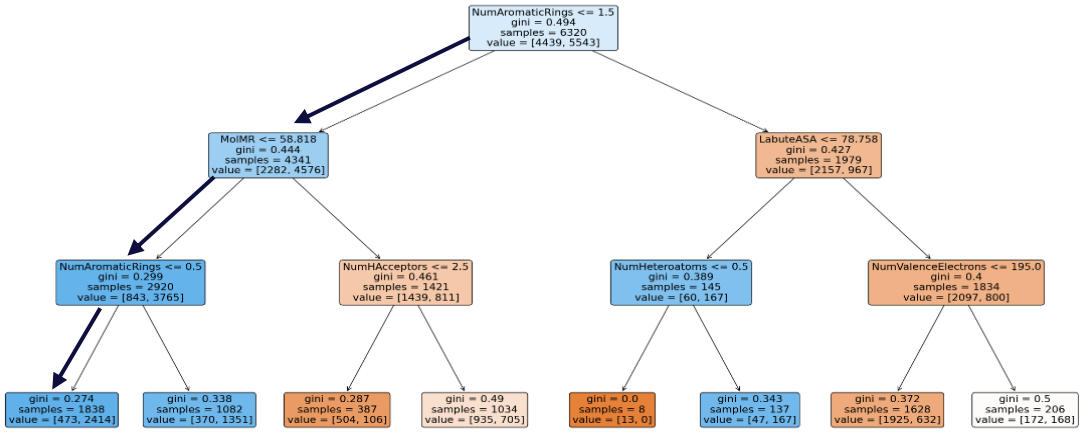
The desired increase in target outcome which is high solubility compounds is 20%. The original high solubility rate is **55.15**% so a **20**% increase should lead to a **66.18**% overall high solubility rate (**55.15*1.2**). Which means **6606** of the experiments from **9982** would become high solubility compounds rather than current **5505.**

The correlation between SXI and High Solubility Rate is **0.99**. This implies that SXI and High Solubility Rate are highly positively correlated to each other. Hence, an increase in SXI will result in increase in High solubility rate.



Current SXI and Target SXI Decision Trees

a. Current SXI Decision Tree



Interpretation

Node 1: Number of Aromatic Rings < 2 (Number of highly soluble compounds in parent node: 5543).

Left split: 4576 – majority positive class; gini: 0.44, **Right split:** 967; gini: 0.427.

(Total value for the next split: 4576)

Node 2: Molar refractivity <= 58.82

Left split: 3765 – majority positive class; gini: 0.299, **Right split:** 811; gini: 0.461.

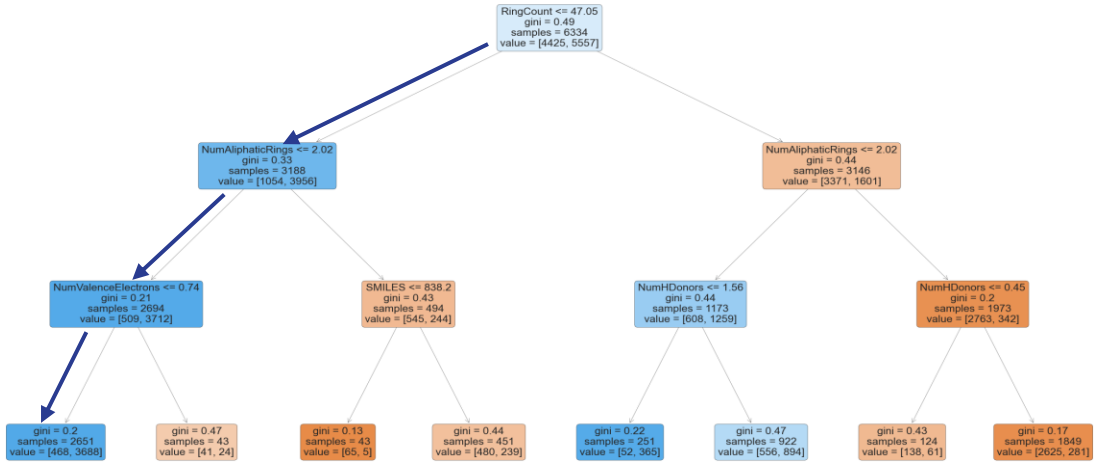
(Total value for the next split: 3765)

Node 3: Number of Aromatic Rings < 1

Left split: 2414 - majority positive class; gini:0.274, **Right Split**: 1351; gini:0.338 – Final Leaf Node

- ✓ **2414** compounds have high solubility.
- ✓ Success Ratio is **52.7%**. (2414/4576) *100 – (Total value of the positive class in the final leaf node/Total value of the positive class after first split) *100
- ✓ Compounds with high solubility / low solubility ratio is **5.1.**

b. Target SXI Decision Tree



Target SXI from correlation curve for 20% increase in target outcome of high solubility rate is **2.81**.

Interpretation

Node 1: Ring Count <= 47 (Number of highly soluble compounds in parent node: 5557).

Left split: 3956 – majority positive class; gini:0.33, **Right split:** 1601; gini:0.44.

(Total value for the next split: 3956)

Node 2: Number of Aliphatic Rings <= 2

Left split: 3712 – majority positive class; gini:0.21, **Right split:** 244; gini:0.43.

(Total value for the next split: 3712)

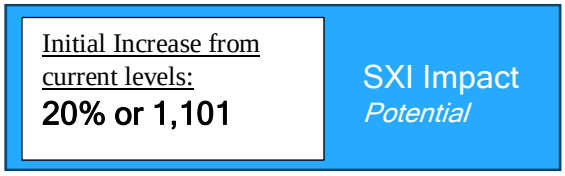
Node 3: Number of Valence Electrons <= 1

Left split: 3688 - majority positive class; gini:0.2, **Right Split**: 24; gini:0.47 – Final Leaf Node

- ✓ **3688** compounds have high solubility.
- ✓ Success Ratio is **93.2%**. (3688/3956) *100 – (Total value of the positive class in the final leaf node/Total value of the positive class after first split) *100
- ✓ Compounds with high solubility / low solubility ratio is **7.88**

Conclusion

- Experiments, whose SXI score is higher than current average SXI score of **2.39** have **14%** higher solubility rates than overall solubility rates average of all experiments.
- Target **20%** increase in high solubility compounds is achievable by raising target SXI to **2.81** from current **2.39** levels. This would result in **6606** high solubility compounds from current **5505** levels.



- Target Compounds with high solubility / low solubility ratio is **7.88** while the current ratio is **5.1.** This represents a **potential 54.51% compounded increase** if all recommendations in target SXI are completely implemented**.**