# Modeling of ADME relevant endpoints from AstraZeneca data set: A Case study:

AstraZeneca has deposited ADME relevant endpoints into publically available resource called ChEMBL database (https://www.ebi.ac.uk/chembl/). Here below, I have attempted to model logD endpoint, and discussed how my findings can be generalized and extended in to some robust tool.

## Step-1: Data extraction

There were two ways to download data

1) By downloading tab delimitated file from https://www.ebi.ac.uk/chembl/bioactivity/results/1/cmpd_chemblid/asc/tab/display and then extract the logD data.
2) By downloading SQLite file of ChEMBL db from http://ftp.ebi.ac.uk/pub/databases/chembl/ChEMBLdb/releases/chembl_24_1/ and then extract the logD data for dataset with ASSAY_CHEMBLID of 'CHEMBL3301363'. From here onwards, this dataset has been referred as 'logD dataset' in this report.

## Step-2: Data analysis

There were total 4200 compounds in the logD dataset.

1) Duplicates: Total 12 compounds found to have duplicated canonical SMILES notations. All were discarded.
2) Check for salts and mixtures: There was single chemical in hydrate form, thus was discarded.
3) Missing values: No record with missing values
4) Data distribution: Data was normally distributed

## Step-3: Model building

The formatted logD dataset with 4187 compounds was used to construct different machine learning models.

1) Descriptor calculation: The PaDEL software was used to derive 729 1D and 2D numerical descriptors.
2) Descriptor filtering: Descriptors with missing values and those with near zero variance were removed. Final set was comprised of total 386 descriptors.
3) Data was scaled and mean centered.
4) For each pair of descriptors with a correlation coefficient higher than 90%, the one showing the largest pair correlation with all the other descriptors was excluded, which reduced descriptor set to only 201 descriptors.
5) Data set was randomly partitioned into training (80%) and test set (20%).
6) Outlier detection: 14 outliers were removed from training set

*Table 1: Composition of training and test sets.*

| Original data | Removed | Training set | Test set |
|---|---|---|---|
| 4200 compounds | 27 compounds | 3337 compounds | 836 compounds |

7) Model construction:

Four different types of methods were employed to construct four models using caret, randomforest and e1071 libraries in R.

a) Using simple linear regression method: lm
b) Using instance based learning method: knn
c) Using random subspace based method: rf
d) Using kernel based method: svm

8) Results:

a) Linear regression:

*Table 2: Model statistics for internal as well as external set predictions.*

| **Model name** | **CV fold** | **RMSE** | **R2** | **Avg RMSE** | **Avg R2** |
|---|---|---|---|---|---|
| Model_cv_1 | 1 | 0.82 | 0.55 | 0.78 | 0.58 |
| Model_cv_2 | 2 | 0.79 | 0.57 | | |
| Model_cv_3 | 3 | 0.78 | 0.57 | | |
| Model_cv_4 | 4 | 0.75 | 0.62 | | |
| Model_cv_5 | 5 | 0.78 | 0.58 | | |
| Test set prediction | - | 0.83 | 0.53 | - | - |

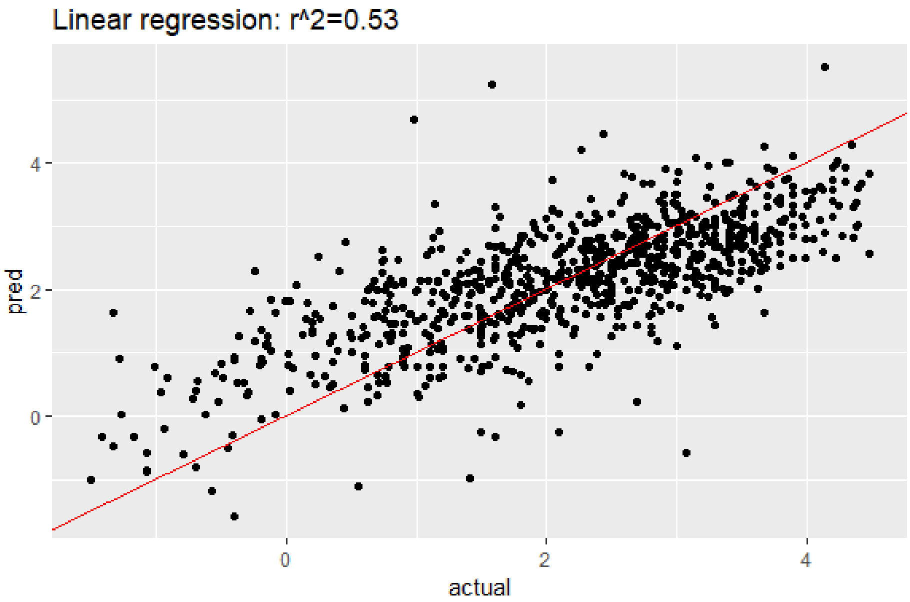

*Figure 1: Plot of actual Vs. predicted values of test set.*

b) *k*-nearest neighbor regression:

*Table 3: Model statistics for cross-validated training set models (k-values ranging from 1 to 10).*

| **Entry** | **K value** | **Avg R2 of CV** | **Avg RMSE of CV** |
|---|---|---|---|
| 1 | 1 | 0.46 | 0.99 |
| 2 | 2 | 0.53 | 0.87 |
| 3 | 3 | 0.54 | 0.85 |
| 4 | 4 | 0.55 | 0.84 |
| 5 | 5 | 0.54 | 0.85 |

| | | | |
|---|---|---|---|
| 6 | 6 | 0.54 | 0.85 |
| 7 | 7 | 0.53 | 0.86 |
| 8 | 8 | 0.53 | 0.87 |
| 9 | 9 | 0.52 | 0.87 |
| 10 | 10 | 0.52 | 0.87 |

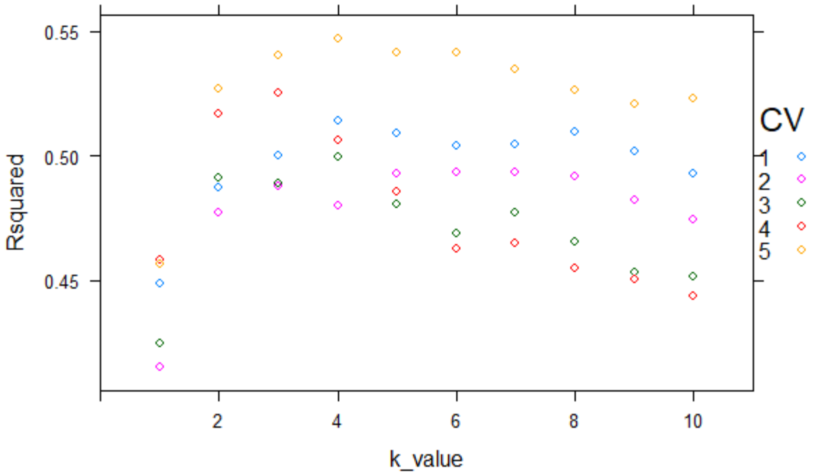

*Figure 2: Optimal k-value selection using principle of highest R-squared.*

The optimal k value was found to be 4. The model that gave highest R-squared at k=4 was chosen as best model. Test set was predicted with R-squared of 0.46 and RMSE of 0.88.

c) Random forest regression:

*Table 4: Random forest model statistics for training and test sets.*

| | R2 | RMSE |
|---|---|---|
| Train | 0.67 | 0.70 |
| Test | 0.66 | 0.69 |

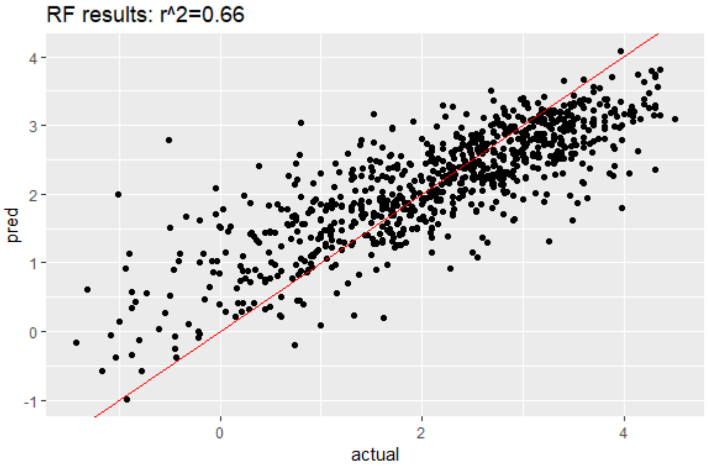

*Figure 3: Actual Vs. predicted values from random forest model*

d) Support vector regression:

*Table 5: SVM model statistics for training and test sets.*

| | RMSE | R-squared |
|---|---|---|
| Training | 0.44 | 0.87 |
| Test | 0.67 | 0.69 |

9) Limitations:

1. Models lack generalizability since they only trained with few thousands compounds.
2. Bias in the data can lead to inaccurate outcome.
3. Lack of 3D descriptors limits prediction power, and calculation of such descriptors are often time consuming.
4. Cannot be used to predict logD of compounds that are in salt form.

10) Summary:

Four different types of methods were adopted to construct regression model. Based on test set prediction statistics SVM>RF>LM>kNN. The difference between training and test set prediction in SVM model was very high and indicate the possibility of overfitting. Random forest model found to be more robust. Since only 729 descriptors were employed at first place, the sub-optimal outcome is expected. These models can be further improve by incorporating more numerical descriptors and by tuning different model parameters.

**Future perspective:**

The solubility, logD, plasma protein binding and intrinsic clearance are important endpoint for prioritization of compounds in a drug discovery project. ADME screening in tandem with toxicological screening will offer great values to the compound screening. My expertise in virtually screening diverse compounds for the off-target effects utilizing *in vitro* bioassay data, gene expression data, gene set enrichment/pathway analysis data, and single cell morphological profiling (high content imaging) data would add new dimensions to the compound prioritization project.