Innoplexus ADME Model Evaluation On TDC Benchmark ADMET Group Regression Datasets.

Rohit Yadav, Divya Nair, Prajakta Mate, Rushikesh Chaudhari, Anand Muglikar, Arpan Sheetal, Sudhir Mansukh, Amit Agarwal

Innoplexus Consulting Services Pvt. Ltd., A Partex Company

Introduction

Drug discovery and development is a complex and costly process that involves the identification and optimization of lead compounds with desirable absorption, distribution, metabolism, excretion, and toxicity (ADMET) properties. The Therapeutics Data Commons (TDC) provides a standardized platform for evaluating machine learning models on various ADMET benchmarks. This white paper presents the evaluation of four machine learning models LightGBM, XGBoost, Catboost and RandomForest on nine regression ADMET datasets from TDC benchmark group: caco2_wang, lipophilicity_astrazeneca, solubility_aqsoldb, ppbr_az, vdss_lombardo, half_life_obach, clearance_hepatocyte_az, clearance_microsome_az, ld50_zhu.

Machine learning techniques, such as Support Vector Machine (SVM) and Artificial Neural Network (ANN), have been used with high success in pharmacokinetic studies for the development of new drug candidates. These methods are known for offering relatively quick results and saving high laboratory costs. However, there is a gap in the literature regarding the application and comparison of other machine learning techniques, such as LightGBM, Random Forest, Catboost and XGboost, in ADMET property prediction.

This study aims to fill this gap by evaluating the performance of LightGBM, XGBoost, RandomForest, and CatBoost on the selected TDC regression ADMET datasets. The goal is to identify the most reliable model for predicting these properties and understand the underlying reasons for their performance.

Methodology

Data Preprocessing

The datasets were obtained from the TDC ADMET benchmark group. Each dataset consists of molecular structures represented by SMILES strings and their corresponding ADMET property values. The following preprocessing steps were applied:

1. Molecular Fingerprint Generation:

- Combined fingerprints were computed for each molecule using the RDKit Package. These fingerprints serve as feature vectors for the machine learning models.

2. Data Splitting:

- Each dataset was split into training, validation, and test sets. The training and validation sets were combined for model training, while the test set was reserved for final evaluation.
- A 5-fold cross-validation approach was employed to ensure robust model evaluation.

3. Feature Scaling:

- StandardScaler was used to normalize the feature vectors, ensuring that all features contribute equally to the model training process.

Model Training and Evaluation

Four machine learning models were evaluated: LightGBM, XGBoost, RandomForest, and CatBoost. These models were chosen due to their proven performance in regression tasks and their ability to handle high-dimensional data.

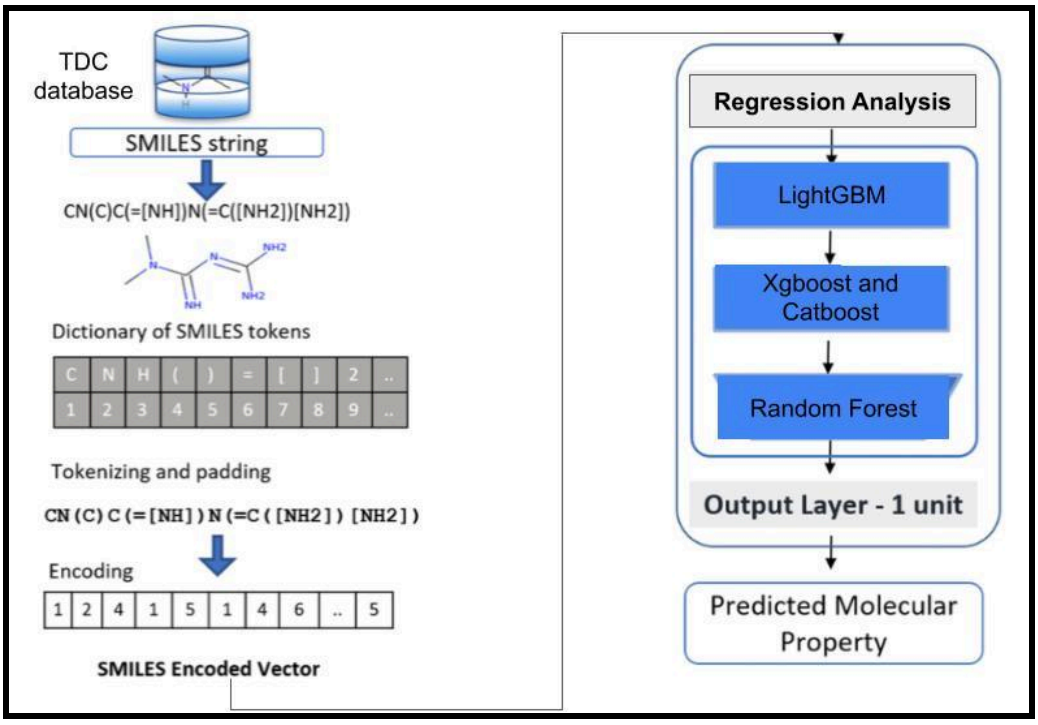

Figure1: Model Framework for ADME prediction

1. LightGBM:

LightGBM, short for Light Gradient Boosting Machine, is a highly efficient and fast implementation of the gradient boosting framework. It is specifically designed to be distributed and efficient, making it suitable for large-scale data processing and machine learning tasks.

2. XGBoost:

XGBoost, short for Extreme Gradient Boosting, is an optimized distributed gradient boosting library designed to be highly efficient, flexible, and portable. It implements machine learning algorithms under the Gradient Boosting framework. XGBoost provides a parallel tree boosting (also known as GBDT, GBM) that solves many data science problems in a fast and accurate way.

3. RandomForest:

RandomForest is an ensemble learning method primarily used for classification and regression tasks. It operates by constructing multiple decision trees during training and outputting the mode of the classes (classification) or mean prediction (regression) of the individual trees.

4. CatBoost:

CatBoost, short for Categorical Boosting, is a high-performance, open-source gradient boosting on decision trees library developed by Yandex. It is designed to handle categorical features automatically and efficiently, making it particularly well-suited for datasets with a mix of numerical and categorical data.

Performance Metrics

The models were evaluated using the following metrics:

1. Spearman Correlation:
- Measures the rank-order correlation between predicted and actual values. It is robust to outliers and does not assume a linear relationship.

2. Mean Absolute Error (MAE):
- Measures the average magnitude of errors in predictions.

3. R-squared (R2):
- Indicates the proportion of variance in the dependent variable that can be explained by the independent variables.

4. Mean Squared Error (MSE):
- Measures the average squared difference between predicted and actual values.

Results

Model Selection

Based on the average mean absolute error scores, a single model was selected as the best-performing model for each Benchmark Datasets from TDC.

Final Evaluation

The final evaluation was conducted on the test sets of each dataset using the best-performing model. The results are as follows:

TDC Benchmark Datasets	Metric	Score (SD)
caco2_wang	MAE	0.289, 0.005
lipophilicity_astrazeneca	MAE	0.499, 0.003
solubility_aqsoldb	MAE	0.771, 0.005
ppbr_az	MAE	8.582, 0.036

vdss_lombardo	Spearman	0.707, 0.006
half_life_obach	Spearman	0.511, 0.0
clearance_hepatocyte_az	Spearman	0.457, 0.013
clearance_microsome_az	Spearman	0.62, 0.007
ld50_zhu	MAE	0.588, 0.0

Discussion

The results suggest and align with previous research that the effectiveness of gradient boosting techniques in handling high-dimensional data and capturing complex patterns.

The success of these models can be attributed to several factors:

1. Feature Importance:
   - RandomForest, LightGBM, and CatBoost are capable of identifying and prioritizing important features, which enhances their predictive power.

2. Handling of Non-linearity:
   - Gradient boosting methods, such as LightGBM, XGBoost, and CatBoost, are effective in capturing non-linear relationships between features and target variables.

3. Robustness to Overfitting:
   - Ensemble methods, particularly RandomForest, are less prone to overfitting compared to single decision tree models.

Comparison with Other Techniques

While SVM and ANN have been widely used in ADMET property prediction, this study demonstrates that other machine learning techniques, such as gradient boosting and ensemble methods, can also achieve high predictive accuracy. The decision tree algorithm, for example, was found to be highly effective in previous studies due to its ability to handle small datasets and exclude irrelevant features.

Conclusion

This study demonstrates the efficacy of machine learning models in predicting ADMET properties. The Innoplexus ADME model, which integrates various machine learning techniques, works well across all categories, achieving high Spearman correlation scores and robust performance across multiple datasets. These results highlight the potential of machine

learning techniques in accelerating drug discovery by providing accurate predictions of ADMET properties.

Future work will focus on exploring more advanced models, such as deep learning architectures, and incorporating additional molecular descriptors to further improve prediction accuracy. Additionally, the integration of domain knowledge and experimental data will be essential for refining these models and ensuring their applicability in real-world drug discovery scenarios.