Abstract:

Introduction:

Introduce reactions:

Methods:

Results:

Analysis:

Discussion:

Reactions used:

The reactions used were

Methods:

Data Overview

There were 381 total reactions in the dataset, collected from \_ sources. Each reaction included a substrate, solvent, catalyst, nucleophile, and imine. Numerical properties of the solvent (160 properties), catalyst (85 properties), nucleophile (15 properties), and imine (22 properties) were identified, as well as a DDG value of each reaction \_\_\_insert explanation\_\_\_. After data processing, we applied various machine learning algorithms to the dataset.

Development of Models

Three sets of relatively simple machine learning models were developed for various purposes.

The first set of models were machine learning regression models in which the properties of the solvent, catalyst, nucleophile, and imine of a reaction were used to predict the DDG value. Four separate machine learning models were tested – Lasso, Decision Tree, Boosting Tree, and Random Forest -- and the results of each were compared.

The second set of models were machine learning regression models in which the properties of the solvent, catalyst, and nucleophile of a reaction were used to predict the DDG value. Imine properties were excluded. Four separate machine learning models were tested – Lasso, Decision Tree, Boosting Tree, and Random Forest -- and the results of each were compared.

The third set of models were machine learning classification models in which the properties of the solvent, catalyst, and nucleophile were used to predict the imine transition state of a reaction (either E or Z). Six separate machine learning models were tested – K Nearest Neighbor, Decision Tree, Random Forest, Logistic Regression, and Linear Discriminant Analysis -- and the results of each were compared.

Training & Evaluation

We trained the models by randomly splitting the dataset 50:50 into a training set and testing set. The trained model was used to make predictions on the test data, and the predicted values were compared to actual test data values to measure accuracy. For regression models in which we predicted numerical values, r^2 was used as the metric of accuracy. The R^2 value measures how much variance in the dependent variable is explained by the model – the closer the r^2 value is to 1, the more accurate the model. For classification models, accuracy was measured simply by dividing the number of correct predictions by the total number of predictions made.

We also observed the feature importance of the models to identify chemical properties that were important and influential to the model’s predictions.

For every type of model we evaluated, we repeated the model creation/training process 100 times, each time with a different random 50:50 split of train/test data. This process ensured that \_\_\_ The mean accuracy/r^2 value over all 100 iterations was calculated and used to determine the accuracy of the model. Similarly, the mean feature importance of each chemical property across all 100 iterations was calculated.

Figure 1: Results of Random Forest Regression model from second set (excluding iminium properties). The predicted r^2 is 0.933, and the total r^2 is 0.953

Chart, scatter chart

Description automatically generatedChart, scatter chart

Description automatically generated

Figure 2: Results of Random Forest Regression model from first set (including iminium properties). The predicted r^2 is 0.926, and the total r^2 is 0.957.

Table 1: Results of models in the first set (including iminium properties)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Models | MSE | Test r2/STD | Train r2/STD | Total r2/STD |
| Lasso | 0.081529 | 0.857335/0.028593 | 0.939504/0.007948 | 0.898993/0.012574 |
| Decision Tree | 0.351993 | 0.883525/0.285952 | 0.997219/0.001423 | 0.940417/0.014141 |
| Boosting Tree | 0.229198 | 0.924112/0.011411 | 0.987517/0.001991 | 0.955898/0.005407 |
| Random Forest | 0.223459 | 0.926373/0.013612 | 0.987074/0.002056 | 0.956674/0.006035 |

Table 2: Results of models in the second set (excluding iminium properties)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Models | MSE | Test r2/STD | Train r2/STD | Total r2/STD |
| Lasso | 0.637415 | 0.788066/0.040520 | 0.873554/0.016922 | 0.831636/0.014616 |
| Decision Tree | 0.290518 | 0.903708/0.019526 | 0.976915/0.003481 | 0.940484/0.009601 |
| Boosting Tree | 0.233848 | 0.922961/0.011408 | 0.967222/0.004021 | 0.945171/0.005171 |
| Random Forest | 0.203178 | 0.932761/0.011805 | 0.972154/0.003513 | 0.952569/0.005606 |

Table 3: Results of models in the third set

|  |  |  |  |
| --- | --- | --- | --- |
| Model | Test acc | Train acc | Total acc |
| Knn | 0.970366 | 0.977737 | 0.974042 |
| Decision tree | 0.960000 | 0.993053 | 0.976483 |
| Random forest | 0.969581 | 0.993474 | 0.984196 |
| Logistic regression | 0.941571 | 0.973789 | 0.957638 |
| lda | 0.954555 | 0.983947 | 0.969213 |

Results:

Out of the first set of models, the best performing model was the Random Forest model with Hyperparameters [insert description].The model had a r^2 value on unseen test data of [insert score], and an overall r^2 value on all data of [insert score]. The performance of the other models developed can be found in [insert table].

The most important features in the Random Forest model were mostly properties of the imine and nucleophile of the reaction [reference to above table]. The Natural Bond Orbital C parameter was by far the most influential on the regressor, while other important features identified by our model include [insert features].

Similarly, the strongest model out of the second set of models was a Random Forest model with an r^2 value on unseen test data of [insert score], and an overall r^2 value of [insert score]. The performance of the other models developed can be found in [insert table].

The most important features in the Random Forest model were mostly properties of the nucleophile of the reaction [reference to above table]. The nucleophilic angles H-X-Nu and H-X-CNu were the most influential on the regressor.

The strongest model out of the third set of models was a K Nearest Neighbors Classifier with [insert accuracy] accuracy on unseen training data and [insert accuracy] accuracy on all data. The performance of the other models developed can be found in [insert table].

Analysis:

Firstly, our results indicate a strong correlation between the structural parameters of the various molecules involved in a reaction and the DDG value of the reaction. Our strongest models in both sets of regression models had r^2 values well over 0.9 when making predictions on test data, indicating strong performance, even on data it had not seen yet.

Another interesting result was the strong performance of the second set of models (which excluded imine properties). The Random Forest model (test r^2 = [insert score]) from this set was able to slightly outperform the Random Forest model from the first set (test r^2 = [insert score]), despite its disadvantage in not having information about the iminium involved with the reaction. This was an unexpected finding, as one would expect a decrease in performance after excluding imine properties since properties of the imine were very influential in making predictions in the first set of models. Rather, it was observed that nucleophile properties which were somewhat influential in the first set of models became the most important properties in the second set of models.

Additionally, the third set of models performed robustly in predicting the transition state of the imine. In addition, most of the important features identified were nucleophile properties that were also important features of the second set of models. This, along with the findings in the earlier paragraph, may signify that imine properties in its transition state can be predicted using properties of other substances involved in the chemical reaction, especially the nucleophilic reactant involved.

Another interesting finding was that generally, in all three sets of models developed, solvent and catalyst properties did not seem to play a major role in making predictions about the reaction.

that the properties of the iminium and nucleophile are clearly the most important in predicting DDG value of a reaction