Extreme Gradient Boosting Model to Predict Synergy Score of Drug Combinations

Name: Serhat BEYAZ

Immatriculation Number: 12495140

ABSTRACT

As a new way of cancer treatment targeted drug therapies have shown a significant increase in efficacy and reduced side effects at the same time. However, it has been observed that cancer cells are able to develop resistance against targeted therapies as well. In order to prevent this to happen, different drug combinations have been developed to be used at the same time. Nevertheless, finding an effective drug combination which works in synergy is a labor intensive, costly, and also time-consuming process. To reduce the consumption of these resources, machine learning algorithms have been developed to predict synergy between two different drugs that has not been used before. The development of these models relies on training the model with thousands of previously generated experimental data with the target value provided. In this study, I present a similar approach which uses Extreme Gradient Boosting (Xgboost) algorithm that was trained with monotherapy data of drugs and genetic data of each cell lines to predict synergy scores of unknown drug combinations. My model performed better than two dummy models, yet it has a significant space for further improvements.

INTRODUCTION

Cancer pharmacogenomics is the study of how changes in the genome affects the response of the different drug treatments. Since cancer cells are evolved from the patient's healthy cells, drug treatments have a lot of heavy side effects. With the advancement of next generation technologies, genetic and transcriptomic differences between cancer and healthy cells can be determined, which allows researchers to develop targeted therapies against cancer cells. However, it has been observed that cancer cells are able to develop resistance against those drugs as well(Sabnis & Bivona, 2019). For this reason, using several different drugs at the same time has been used as an effective way to fight against cancer.

However, drugs may interact with each other and may decrease or increase the overall efficiency. Drugs that display greatly enhanced effects when given in combination are called synergistic(Tallarida, 2011). Laboratory assays could easily determine if the drug combinations are antagonistic or synergistic, but the number of combinations are exponentially increase with

increasing number of drugs. To find the potential candidates of drug combinations, machine learning models has been developed which reduce overall cost and laboratory work. In short, these models have been trained by feeding with the features of previously combined drugs data including chemical structure of drugs, targeted pathway, monotherapy effects of drugs, genomic and transcriptomic data of the cell lines and many more features with the expectation that it can generalize the patterns and predict synergy score of unknown drug combinations.

In this study, I tried to predict synergy scores of the drug combinations by implementing Xgboost algorithm using DREAM challenge dataset(Menden et al., 2019), which includes drug monotherapy response and mutation states of different cancer cell lines.

RESULTS

In order to develop Xgboost model that predicts synergy score, I used monotherapy data of the drugs in addition to the mutation state of the individual cancer cell lines. Since there were 495 genes in the molecular data set, I thought giving all of them as an individual feature would make training difficult due to the limited number of observations in total. Thus, I performed Multiple Correspondence Analysis (MCA) to see if we can explain the variance with a few sets of genes.

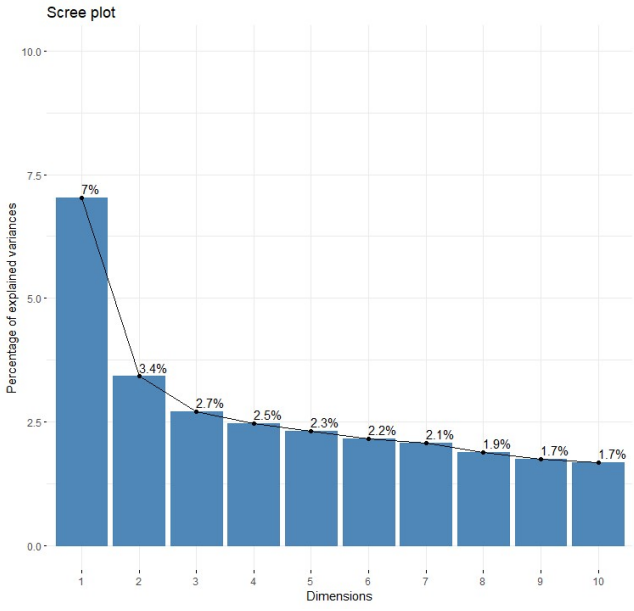

Figure 1: Multiple Correspondence Analysis of all gene matrix

As it can be seen in Figure 1, first 10 dimensions of the results of MCA only explains the %25 of all variation. Therefore, I decided to not use MCA outputs, and continued with all genes as an individual feature.

Then I checked the collinearity between the monotherapy features of the drugs to find a good feature set where all features are independent from each other.

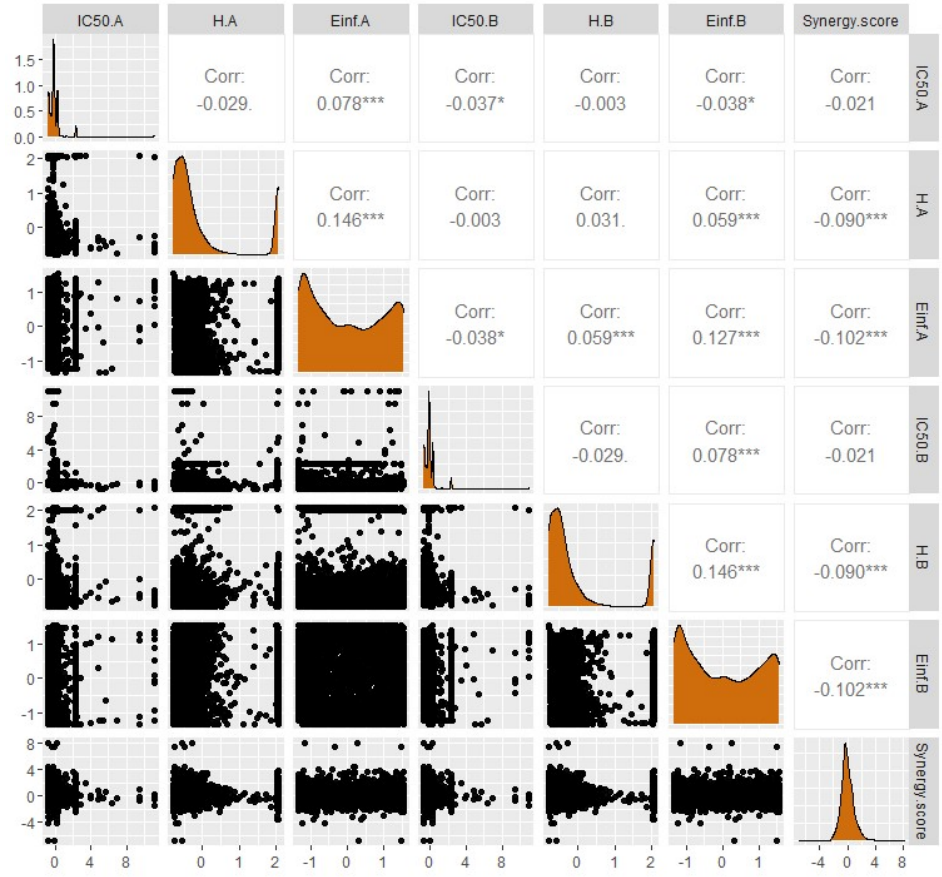

Figure 2: Correlations between monotherapy features in the training set.

As it can be seen in Figure 2, the value of Einf is correlated with both H and IC.50 values for each drug. Therefore, I tried to train the model without including H and IC.50 values and tested the performance of the model. However, in the end the performance has not changed significantly.

After having all the features and observations, I trained an Xgboost model with the default hyperparameter settings, and then trained several times with different hyperparameter

combinations. One example of hyperparameter tuning step can be seen in Figure 3, in which “eta” (learning rate) parameters has been tuned and in the end, I selected the one giving the lowest Root Mean Squared Error (RMSE) value which is 0.010.

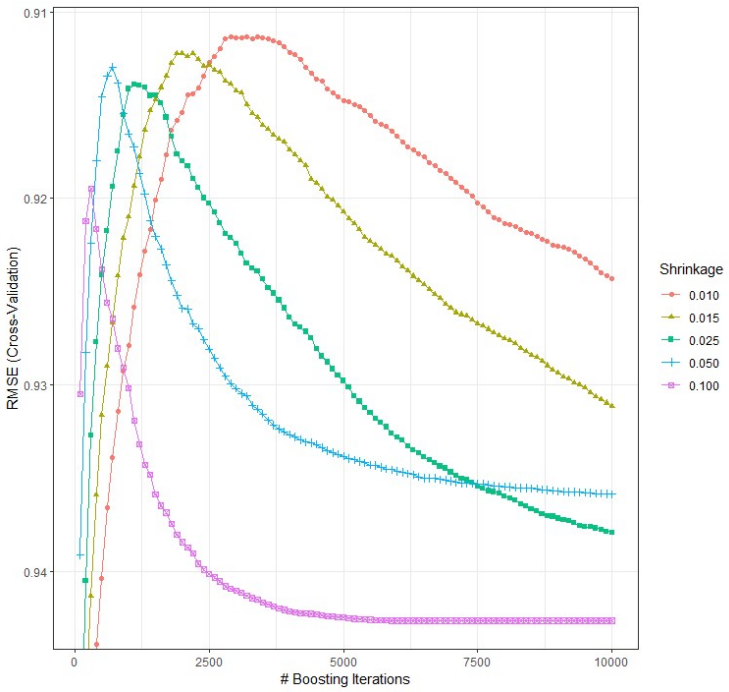

Figure 3: Example of Xgboost hyperparameter tuning. Each line represents performance of different learning rate during training.

After finding the best values for each hyperparameters, the model was trained again. The final model was tested with the test data provided with the train set. As a comparison, two dummy model were used. Details of the models can be found in the methods.

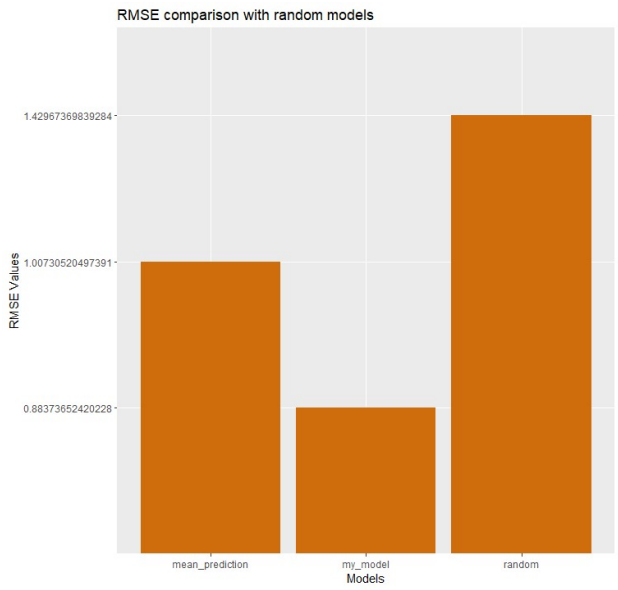

Figure 4: Comparison of the final model with 2 different dummy models.

The prediction of my model gave 0.88 RMSE value while the mean model 1 RMSE and the random one gave 1.42 (Figure 4). Note that all the numeric features in addition to the synergy score were normalized by centering 0 mean and scaling to 1 standard deviation.

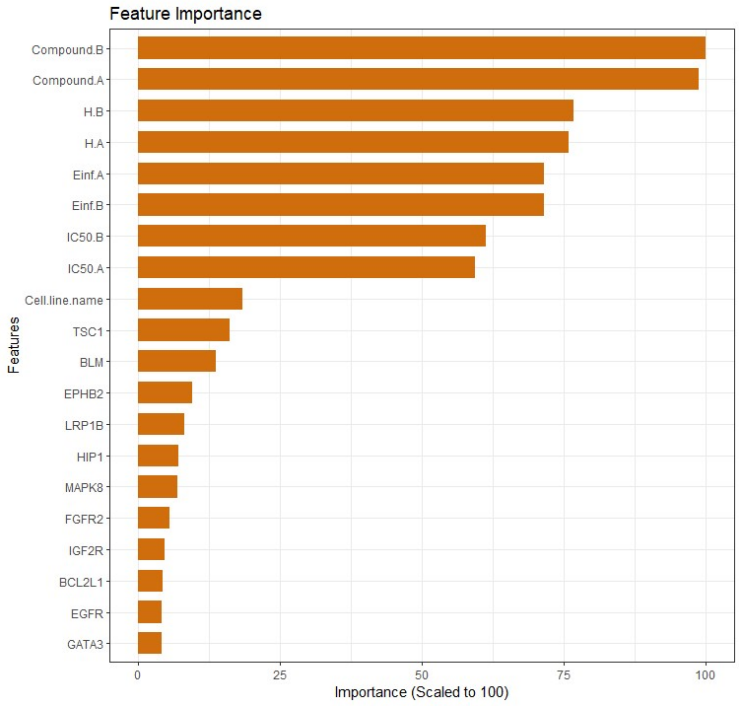

Figure 5: Top20 features that are determined as important by the model

After the RMSE calculation, I looked at the feature importance of the model to understand how my model predicts the synergy score. In Figure 5, you can see that monotherapy features of the drugs and the cell lines are determined as the most important variables by the final model. On the other hand, genes that are determined as important by the model are previously annotated as cancer-related genes.

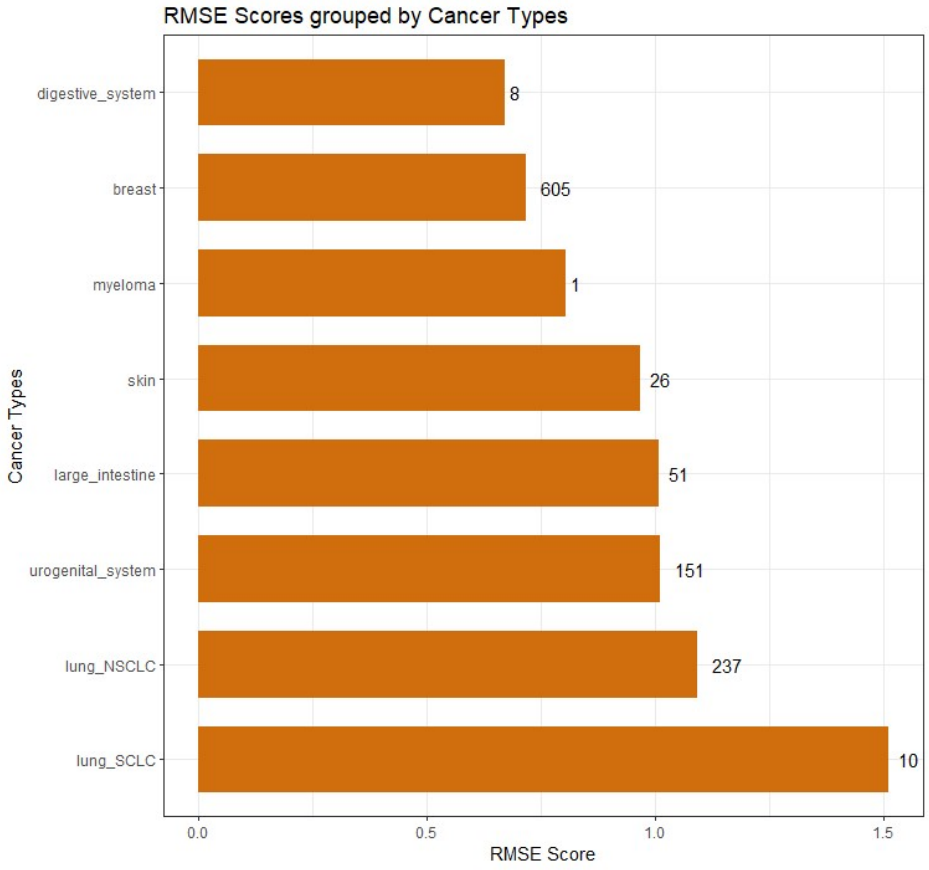

Figure 6: RMSE scores of the final model when the predictions are grouped by cancer type. The numbers at the tip of the bars represents the sample size for a given type.

Finally, I also looked at the performance of my model for a given cancer type to see if there is any difference between them (Figure 6). My model predicted best the digestive cancer types, whereas the worst one is the lung_SCLC. No statistical test was done to see if the difference in prediction errors is significant depending on either sample size or cancer type.

DISCUSSION

In this study, I hypothesized that the synergy of unknown drug combinations could be predicted by machine learning models by training them with previously known drug combinations and genetic features. I used Xgboost algorithm to test the hypothesis. For the training of the model, I used previously provided monotherapy and mutation data from DREAM Challenge(Menden et al., 2019).

Before training I tried to decrease the number of features in the whole dataset by implementing MCA which reduces the number of features by using the most variable dimension in the dataset. However, even if I use the first 10 dimensions which explain 25% of all the variance, I would have lost 75% variation in the dataset, which is not desirable, thus I abandoned this approach. I also checked the collinearity between monotherapy results but excluding the dependent variables have not increased the performance, indeed it increased the overall training time. I also used several feature extraction algorithms to make the model less complex, but it decreased the performance in the end.

In the end my final model performed better than two different dummy models. However, when you consider that synergy score is centered to 0 mean and scaled to 1 standard deviation, resulted RMSE is still high, thus, my model has a significant space for further improvements. After calculating the model performance, I checked the relative importance of the features in the dataset according to my model (Figure 5). My model determined that monotherapy data is the most important one, indeed, the importance of Drug A and Drug B is the same which is what we should expect because combination of A-B and B-A should gave same response. In addition, the important genes my model used are previously determined as cancer-related genes such as TSC1, EGFR, MAPK8 (Slattery et al., 2012). I also checked the RMSE for each cancer types (Figure 6). I found that my model performance differs with the cancer type: the worst predicted one's RMSE (lung_SCLC) is almost two times higher than the best one's (digestive). But further statistical test is needed to find if this difference is significant. In addition to cancer type, I could not find any correlation between sample size and the model performance for each cancer type, which was unexpected for me.

There are several reasons why I could not make a better model than this one. One of them is related to the limitations that I had. Before implementing the Xgboost algorithm I tried Deep

Learning model for the same task. However, computational power of my laptop is very low, and I lost significant amount of time to train different models with different hyperparameters, and network structure, or with same parameters but different features. Then, I switched from Deep Learning model to the Xgboost model, which is more efficient in terms of training time, but I had less time to optimize it in the end. Another thing that I realized in the end is that only using mutation data of the cell lines was not enough to give a good prediction. Indeed, after searching in literature I found that models which use gene expression, drug structures and targeted pathway data in addition to monotherapy data performed much better than the mutation data(Celebi et al., 2019).

If I would have more time, I would try to implement more features in the training data. In addition, I would try to design some autoencoder for each feature which takes data and reduce the dimensionality of it and gave an output to train another big model for prediction. In addition, generation of more drug combination data will allow researchers to train more accurate and more generalized models. In the foreseeable future, we will have a pretty good models with maybe more efficient algorithms which could predict synergy scores of unknown drug combinations and will reduce both laboratory work for synergy assays and the cost of drug development. However, we have to acknowledge that current models perform pretty well too.

METHODS

Data that are used in this study includes monotherapy data for each drug and binarized mutation matrix of 495 genes of given cancer cell lines. Monotherapy data consist of Drug names, Max concentrations of each drugs, IC50 values (conc. required to reduce cell viability by half) of each drugs, slope of the dose-response curve (H), and maximum cells killed with drugs in percentage (Menden et al., 2019). In addition, for each observation we have synergy scores to train the model with the above features.

Extreme gradient boosting (Xgboost) algorithm is used in this study to train the model and predict the synergy scores for an unknown drug combination. Xgboost is a tree learning algorithm used in classification and regression tasks. Gradient boosting algorithms typically use decision tress as weak learners. The idea behind is to train weak learners sequentially, each trying to correct its predecessor. It creates new trees until there will be no improvement in prediction power. It is a

very efficient algorithm which is not computationally demanding like deep learning algorithms(Chen & Guestrin, n.d.).

Before training the model, categorical features were converted into numerical factors, and the numerical features were normalized. For the normalization, several strategies were used. I used min-max scaling and standardization separately and found that standardization is a better choice. Then, I used tangent hyperbolic function after standardization to reduce the effect of extreme values. However, after testing I found that it decreases the performance and not needed. I also tried to implement several feature selection strategies like MCA and Neuralnet + Xgboost combination, but none of them gave significant increase in the performance.

For the training of the model, I used grid search and sequentially train models with different hyperparameters to find the best tune. The tuning trainings was carried out with 3-fold cross validation to reduce the training time. After tuning hyperparameters set as follows: [nrounds: 2900, max_depth = 4, eta = 0.01, gamma = 0, colsample_bytree = 0.4, min_child_weight=2, subsample =1]. With these parameters final model was trained with 5-fold cross validation.

Root mean squared error was used as a performance evaluation metric. Two different dummy models were created to compare with my model. Always_mean model gives the same value for each prediction which is the mean of the actual vector. On the other hand, random model sample random number from the distribution of the actual vector.

REFERENCES

Celebi, R., Bear Don't Walk, O., Movva, R., Alpsoy, S., & Dumontier, M. (2019). In-silico Prediction of Synergistic Anti-Cancer Drug Combinations Using Multi-omics Data. *Scientific Reports*, *9*(1). https://doi.org/10.1038/S41598-019-45236-6

Chen, T., & Guestrin, C. (n.d.). *XGBoost: A Scalable Tree Boosting System*. https://doi.org/10.1145/2939672.2939785

Menden, M. P., Wang, D., Mason, M. J., Szalai, B., Bulusu, K. C., Guan, Y., Yu, T., Kang, J., Jeon, M., Wolfinger, R., Nguyen, T., Zaslavskiy, M., Abante, J., Abecassis, B. S., Aben, N., Aghamirzaie, D., Aittokallio, T., Akhtari, F. S., Al-lazikani, B., … Saez-Rodriguez, J. (2019). Community assessment to advance computational prediction of cancer drug combinations in a pharmacogenomic screen. *Nature Communications 2019 10:1*, *10*(1), 1–17. https://doi.org/10.1038/s41467-019-09799-2

Sabnis, A. J., & Bivona, T. G. (2019). Principles of resistance to targeted cancer therapy: lessons from basic and translational cancer biology. *Trends in Molecular Medicine*, *25*(3), 185. https://doi.org/10.1016/J.MOLMED.2018.12.009

Slattery, M. L., Lundgreen, A., & Wolff, R. K. (2012). *MAP kinase genes and colon and rectal cancer*. https://doi.org/10.1093/carcin/bgs305

Tallarida, R. J. (2011). Quantitative methods for assessing drug synergism. *Genes & Cancer*, *2*(11), 1003–1008. https://doi.org/10.1177/1947601912440575