**ALY 6020 Final Project**

**Prediction of Mechanism of Action of Drugs**

**Northeastern University**

**Professor: Justin Grosz**

**Project Team： Parveen A & Yuxin Deng**

**Oct 19th, 2020**

**Executive Summary**

We are doing Mechanism of Action (MoA) Prediction which uses drug signatures data set which is an active Kaggle competition. MoA means a drug’s mechanism of action and how it affects a specific target protein in a cell, such as an enzyme, or a cell function, such as cell growth. Knowing the mechanism of action of a drug may help provide information about the safety of the drug and how it affects the body. It may also help identify the right dose of a drug and which patients are most likely to respond to treatment. Scientists seek to identify a protein target associated with a disease and develop a molecule that can modulate that protein target. We can develop an algorithm to predict a compound’s MoA given its cellular signature, thus helping scientists advance the drug discovery process.

We do the supervised machine learning with Python and chose the most appropriate classification model to predict if the drug has human cell reaction or not. Gradient Boosting model is chosen for the prediction through cross validation method and the accuracy of the model is up to 70%. We also find top 5 import features, which are cp\_type\_trt\_cp; g-385; g-100; g-175; g-414. These are the important variables for the MoA prediction.

**Introduction**

We are supposed to identify the Mechanism of Action (MoA)of a new drug based on the available information of cell viability and gene expressions and their target MoA. In this problem scientists seek to identify a protein target associated with the disease and develop a molecule that can modulate that protein target. As a shorthand to describe the biological activity of a given molecule, scientists assign a label referred to as mechanism-of-action or MoA for short. Here our target variable is MoA and features used to predict MoA are cell viability and gene expressions. We have been provided the information about human cell responses to drug within a pool of 100 cell types and 772 gene expressions in addition we have access to MoA annotations of 5000 more drugs. Each drug can have more than one MoA, so this is an interesting part where we need to perform multi label classification on the data.

The data set will help us to analyze big data where we will be doing Multi Label Classification Model which is one step ahead of Logistic Regression model. We will be familiar with the challenges that we face in IT jobs in future. We are getting an opportunity to work on real world data set. This will give us an occasion to participate in Kaggle challenge.

We have used Logistic Regression, Random Forest Classifier, Gradient Boosting Classifier, Boost Classifier, Adaptive Boosting and Neural Network in order to observe the accuracy of each model on the training and test data set. The best model will help scientists to predict a drugs MoA given its cellular signature, thus helping them advance the drug discovery process.

**Implementation**

**Step 1 Exploratory Data Analysis**

A picture containing icon

Description automatically generatedWe have imported below 3 files from the Kaggle to perform our data analysis, including Train\_features.csv, Test\_features.csv, and Train\_targets\_scored.csv. Then we do the exploratory data analysis. The first step is to understand the shape of testing, training and target datasets by using *data.shape()* function. To better understand the dataset, we check the training features based on the data types. There are 3 categorical features in the train data set such as cp\_time, cp\_type and cp\_dose. We have mapped the values with integers using map function in pandas. For numerical variables, we specially check the gene expressions. From the distribution plot of gene expression features we can see that most of the samples are normally distributed with mean 0 and values range from -10 to +10.

**Step 2 Feature Engineering**

Chart, bar chart

Description automatically generatedAll the target columns are binary in nature, indicating whether a cell type responds to the drug or not. Since, this is a multi-label classification problem drug samples can have multiple MoA i.e. more than one target variable class can be active. We draw the bar plot to show multi label distribution in target variables. From the plot we can see that nearly 40% of the data has no MoA annotations and majority of the cases show single annotation nearly 52%. Therefore, we try to use *cell\_reaction* column as a binary variable - ‘0’ indicates no cells reacted and ‘1’ indicates one or more cells reacted.

**Step 3 Splitting Dataset**

We have huge dataset, including 26000 rows and 876 columns, we have sampled 80% of it Training and 20% of data for testing.

**Step 4 Data Modelling**

**Part 1 Logistic Regression**

Since we have modified our target variable into binary values ‘0’ indicating no cells reacted and ‘1’ indicating one or more cells reacted we used Logistic Regression to model our dataset. The logistic regression output shows the performance of the model on testing data. **The accuracy is 0.67.**

Cells reacted precision recall f1-score Support

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 0 | 0.64 | 0.40 | 0.49 | 1928 |
| 1 | 0.67 | 0.85 | 0.75 | 2835 |

The accuracy of logistic regression is not good, and we consider that there are several reasons. First, number of features are more and logistic regression will fail to accommodate for nonlinear relations between variables. Also, step wise selection of variable could be tedious process. Hence, we used random forest regression to find the most important features for the model. Second, we did check on the **multicollinearity** which greatly increases bias in the model.

**Part 2 Random Forest Model**

We performed RF classification with n\_estimators (number of decision trees) = 200, min\_sample\_leaf= 4 so as to not overfit our model and performed multiple iterations by changing input parameters to optimize our model. **The random forest model accuracy is 0.68**.

Cells reacted precision recall f1-score support

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 0 | 0.94 | 0.24 | 0.38 | 1928 |
| 1 | 0.66 | 0.98 | 0.79 | 2835 |

From the above results we can see that our precision rate is highly increased for prediction of no cells reacted and recall rate increased for correctly identifying one or more cells reacted.

That means our model is correctly identifying ‘0’ cell reaction. So, in order to improve the precision rate of the model to correctly identify one or more cell reaction we will perform under sampling technique to fine tune our model.

**Part 3 Gradient Boosting Classifier**

We also performed ensemble modeling using GB classifier where our model tries to improve the error got the previous models in the future models. We adjusted the number of boosting stages and learning rate together because there is a trade-off between these two parameters. We set leaning rate equal 0.1 and change the n\_estimators from 20 to 300. We found that when n\_estimators =200 there are high test score. Then we change max\_depth, which limits the number of nodes in the tree. And we can see, the accuracy didn’t change to much after tuning the model. **Accuracy of the model is 69%**.

Cells reacted precision recall f1-score Support

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 0 | 0.98 | 0.22 | 0.36 | 1928 |
| 1 | 0.65 | 1 | 0.79 | 2835 |

**Part 4 XGBoost Classifier**

We also performed ensemble modeling using XGBoost classifier where our model tries to improve the error got the previous models in the future models. In this model, the target variable we used is cell\_reaction, which is the binary variable, including value “0” and “1”. Therefore, in the model, we use objective="binary: logit raw" for binary classification problem. Then, we iterated the model by changing n\_estimators’ value (100, 200, 500), learning\_rate (0.2, 0.1, 0.01), max\_depth (3, 5, 6). In the end, we found that, when n\_estimators=500, learning\_rate=0.2(default), and max\_depth=5 may have the better result.

**Accuracy of the model is 65.59%**.

Cells reacted precision recall f1-score Support

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 0 | 0.61 | 0.40 | 0.48 | 1928 |
| 1 | 0.67 | 0.83 | 0.74 | 2835 |

Here also our model correctly identifies ‘0’ cell reaction.

**Part 5 Cross Validation**

That means our model is correctly identifying ‘0’ cell reaction. So, in order to improve the precision rate of the model to correctly identify one or more cell reaction we performed kfold cross validation by changing the input parameters to all the models and found the below result:

Kfold cross validation for all the models:

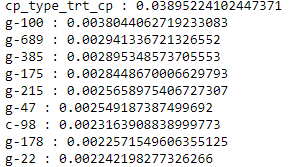
Table

Description automatically generated

From the above result we can see that Gradient Boost gives the highest accuracy nearly 70% when we performed 3-fold sampling of the data.

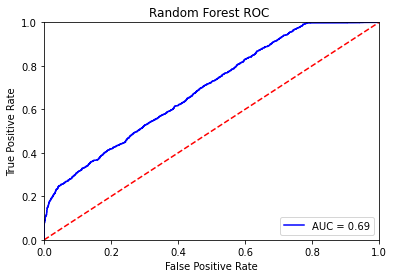
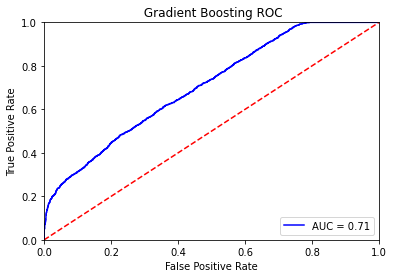
**Part 6 Feature Importance**

We performed random forest classification using all the features to identify which features helps to correctly identify the target MoA. Below screenshot shows the top 10 important features and among them cp\_type with compound treatment highly influences in the detection of target MoA.

**A picture containing graphical user interface

Description automatically generated**

**Part 7 ROC Plot for comparing models**

****

**From the above plots we can see that area under the curve is highest for Gradient Boosting which again proves that it is the best model to perform MoA prediction.**

**Conclusion**

Based on the kfolds cross validation report we can see that Gradient Boosting gives good precision and recall score to correctly identify the target MoA for test data which tries to compare similar patterns from the training set given cell and gene expressions of the test set. In addition, we find the Sample treated with Compound is an important feature to predict MoA of new drugs. We can provide this model to the scientists so that they will be able to predict the target MoA of a new drug which further helps them in their drug discovery.

**References**

Mechanism of action. Retrieved from <https://en.wikipedia.org/wiki/Mechanism_of_action#:~:text=In%20pharmacology%2C%20the%20term%20mechanism,as%20an%20enzyme%20or%20receptor>

Alison. L. (2008.05). Factors Affecting the Production of l-Phenylacetyl carbinol by Yeast: A Case Study. Retrieved form <https://www.sciencedirect.com/science/article/abs/pii/S0065291108601642>

Data Souse: Mechanisms of Action (MoA) Prediction. Retrieved from: <https://www.kaggle.com/c/lish-moa>