Final Project Report

Grace Lin (NetID: el3637)

5/10/2022

Introduction

Clinical trials are the central means by which preventive, diagnostic, and therapeutic strategies are evaluated, but the US clinical trials enterprise has been marked by debate regarding funding priorities for clinical researches. The National Institutes of Health (NIH) and the pharmaceutical industry have been major funders of trials. In general, the pharmaceutical industry funds trials that test their own products, whereas the NIH's funding strategies are not commercially motivated.

Until recently, however, we have lacked tools for comprehensively assessing trials across the broader US clinical trial enterprise. In September 2004, the International Committee of Medical Journal Editors (ICMJE) announced the policy of requiring registration of clinical trials as a pre-requisite for publication. The Food and Drug Administration Amendment Act expanded the mandate of ClinicalTrials.ov to include most non-phase1 interventional drug and device trials, with interventional trials defined as "studies" in human beings in which individuals are assigned by an investigator based on a protocol to receive specific interventions. The law obliges sponsors or their designees to register trials and record key data elements, report basic results, and report adverse events.

With the registry database, we are able to look for the funding strategies of different types of sponsors across different fields of clinical researches. In this study, we hypothesize that the clinical trials funded by NIH has a set of unique characteristics that leads to the NIH's sponsoring in this type of trials. A set of machine learning techniques will be applied to model the NIH's funding behavior in the field of Gastroenterology.

Related work

There had been studies using descriptive statistics to summarize the characteristics of clinical researches in different fields. Those studies would apply simple statistical tests, such as chi-squared test, to validate the differences between NIH-funded and non-NIH-funded trials. There is no existing publication applying more complex statistical methods, such as regression or machine learning algorithms,to analyse the funding trend of clinical trials data. This project, hence, seek to offer a new direction for trend analysis in the clinical trial registry data.

Methods

Data Overview and Objective

The raw data used was originally obtained from clinical trials registry's website, it includes all the clinical trials registration information in the field of Gastroenterology(GI), all of them have start dates between 2013 and 2019. The dataset version used for this project has 4182 subjects, and 23 types of information recorded, while majority of them being categorical. With this dataset, I aimed to find any behavioral pattern in federal/government fundings in this area of clinical trials. It is known that NIH is the most common type of federal funding in clinical trials, the objective of this project was finalized as: to model the funding behavior of NIH in GI clinical trials between year 2013-2019, which would be a binary classification machine learning problem.

Modelling Approaches

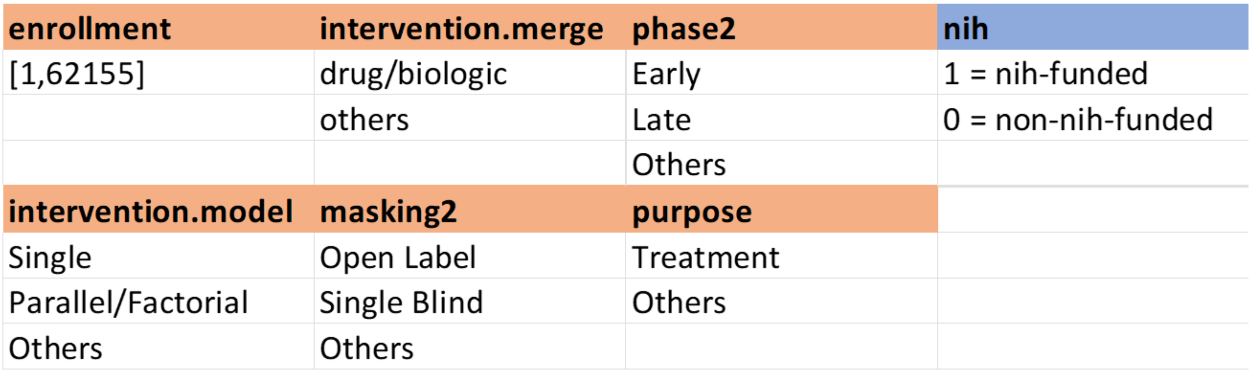
Without knowing the relationships between the variables in the final dataset used, the approaches chosen ranges from simple to more complex algorithms. Started with KNN and logistic regression, of which one is non-parametric and the other being parametric. The difference in output accuracy by the two simple models was of interest. 5-fold cross validation was used to boost the robustness of the two algorithms. Followed by conducting Random Forest, the non-parametric ensemble method chosen that is one of the most inclusive methods for a wide range of data structures. The method would also help revealing the importance of different features and utilize the information during the modelling. The modelling was concluded with applying Neural Network to the dataset, because it is the most robust algorithm in learning and modelling non-linear and complex relationships between the variables.

Data and Experiment setup

To achieve the study objective, variable `fund_source_grp`, which stored the information of funding source for each clinical trial, was chosen as the outcome variable. 6 predictors of interest were picked from the raw data, which are the criteria of greatest concern in clinical research, namely the `enrollment`, `intervention.merge`, `phase2`, `intervention.model`, `masking2`, and `purpose`.

Among the predictors chosen, 5 of them are multi-class categorical variables, and 1 (`enrollment`) is continuous. In order to reduce the complexity of the data, categorical variables were recoded into fewer groups, so as to reduce the noise in the modelling process. Next, the only continuous variable `enrollment`, which is the number of enrolled candidates in the trial, is a highly skewed variable, the value ranged from 1 to 60,000. It was then decided to first standardize it before the modelling process. Last but not least, in order to fit models that address the objective of the project, the output variable was manipulated to a new binary variable called `nih`: with value 1 representing NIH-funded trials and 0 the non-NIH-funded trials.

The resulting variables and groups can be seen below.



enrollment	**intervention.merge**	**phase2**	**nih**
[1,62155]	drug/biologic	Early	1 = nih-funded
	others	Late	0 = non-nih-funded
		Others	
intervention.model	**masking2**	**purpose**	
Single	Open Label	Treatment	
Parallel/Factorial	Single Blind	Others	
Others	Others		

Figure 1: Reclassified variables

However, multi-class categorical variables still exist. Within them, each category does not have an ordinal relationship with each other, hence simply code them into integers is not enough to model their relationship with the outcome variables. To solve this problem, One-hot-encoding technique was used to recode the categorical predictors. This technique recodes each category of the variable into another new variable with binary values. By doing this, the machine learning will not rank one category higher than another unrelated category. At last, the resulting dataset has 14 predictors as shown below.

```
##                        colnames(dt1)[-15]
## 1       intervention.merge_drug/biologic
## 2              intervention.merge_others
## 3                            phase2_Early
## 4                             phase2_Late
## 5                           phase2_Others
## 6                              enrollment
## 7              intervention.model_Others
## 8  intervention.model_Parallel/Factorial
## 9              intervention.model_Single
## 10                 masking2_Open Label
## 11                     masking2_Others
## 12               masking2_Single Blind
## 13                      purpose_Others
## 14                   purpose_Treatment
```

Trainng and validation set

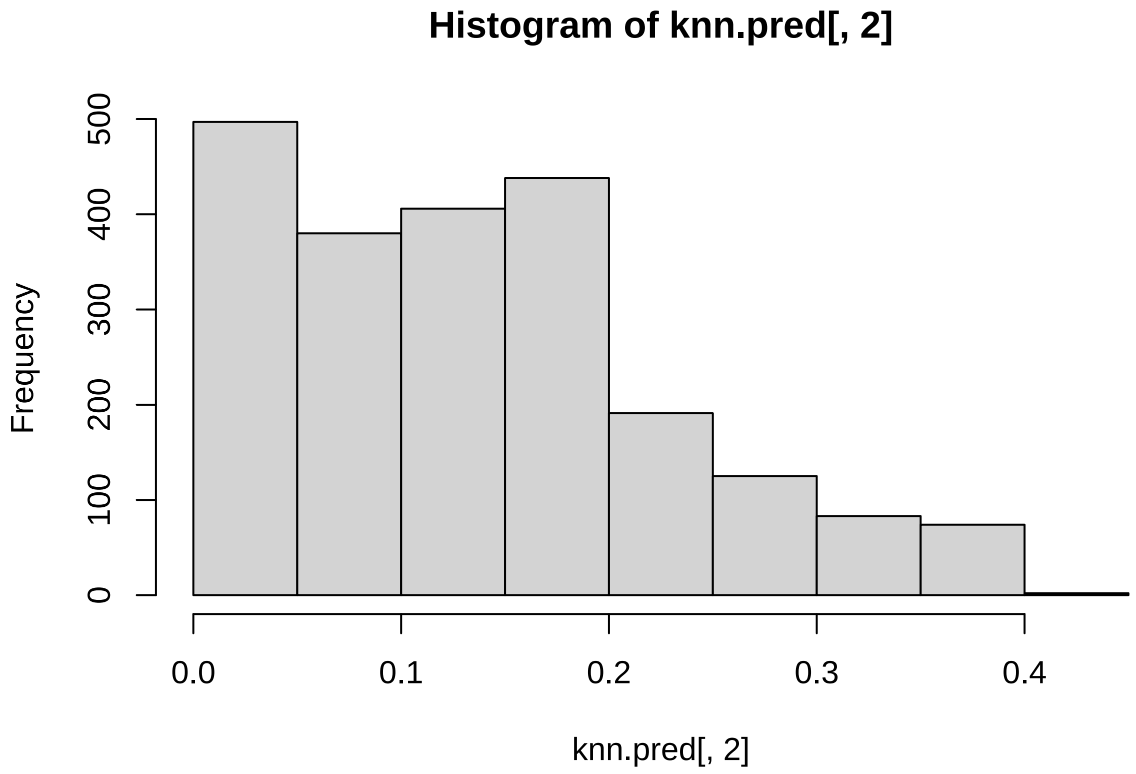
Results

K-Nearest Neighbour and Logistic Regression

The data was first fitted with KNN and Logit methods. It was found that the optimal K if 23 for KNN. Followed by checking the prediction results. Surprisingly, the two algorithms gave the same classification errors.

```
##      Train.Error Test.Error
## KNN    0.1411658   0.134471
## Logi   0.1411658   0.134471
```

This could be due to the imbalanced binary outcome variable we have here, this is almost a 1 to 6 ratio for NIH-funded vs. non-NIH funded clinical trials. The prediction made in both algorithm largely favored the class of non-NIH output, they predicted all the subjects as non-NIH funded, and hence we got the identical errors. By looking at the calibrated probabilities from the prediction, the probabilities are apparently also not normally distributed. An example of the distribution of the prediction probability is shown below.



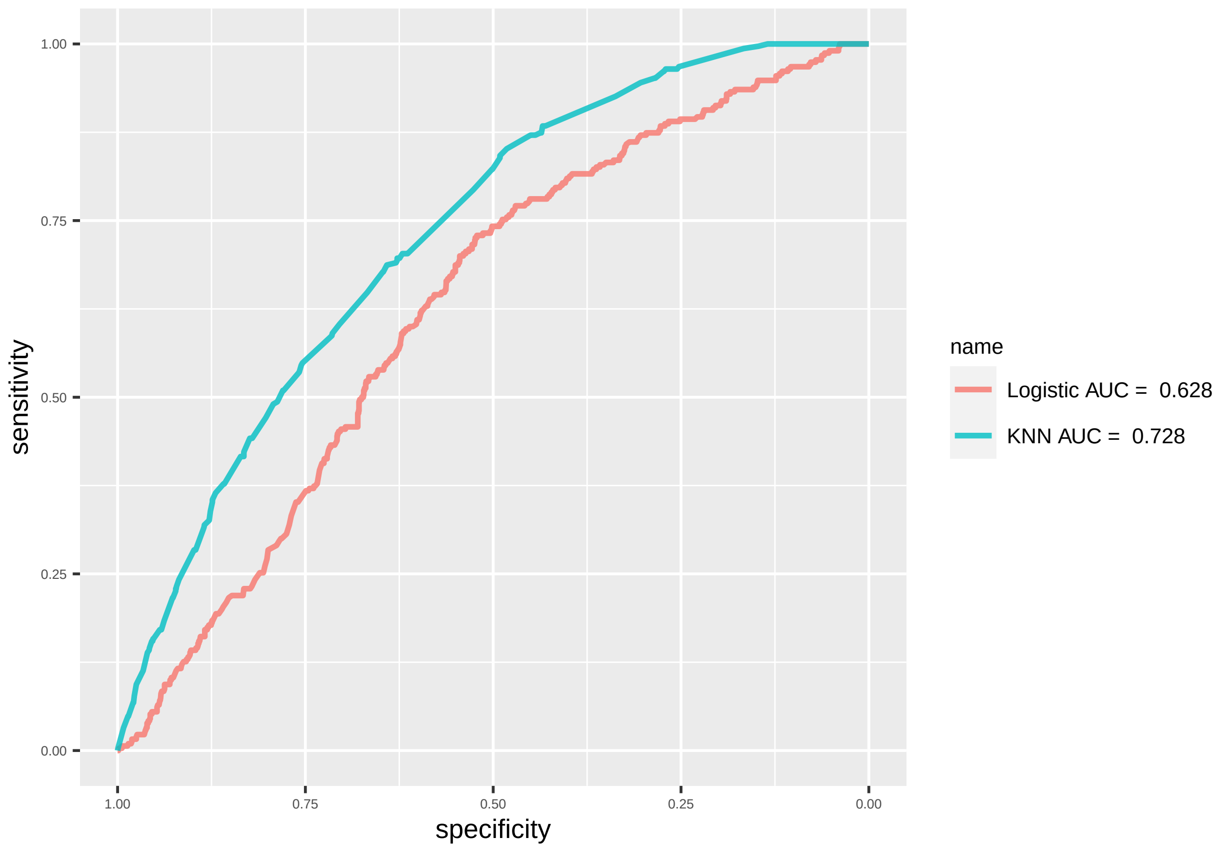
Hence the threshold of 0.5 in determining the class of the output is not desired. In order to improve the model, ROC curves were plotted and the optimal threshold was obtained.

```
## Setting levels: control = 0, case = 1
```

```
## Setting direction: controls < cases
```

```
## Setting levels: control = 0, case = 1
```

```
## Setting direction: controls < cases
```



```
##   KNN.opt.threshold Logit.opt.threshold
## 1         0.1132479           0.1513443
```

For KNN, the optimal threshold is 0.11 and it is 0.15 for logit model. After the transformation, errors with threshold adjusted seemed normal, even though the error rate drastically increased.

```
##      adj.Train.Error adj.Test.Error
## KNN        0.4658470      0.5037543
## Logi       0.4489982      0.4648464
```

To make these two models even more robust in their predictions, 5-fold Cross validation was applied, respectively. With cross validation applied, the classification error for KNN improved by around 32.2%, and it was a 0.73% improvement in the Logit model.

```
## Warning in predict.lm(object, newdata, se.fit, scale = 1, type = if (type == :
## prediction from a rank-deficient fit may be misleading

## Warning in predict.lm(object, newdata, se.fit, scale = 1, type = if (type == :
## prediction from a rank-deficient fit may be misleading

## Warning in predict.lm(object, newdata, se.fit, scale = 1, type = if (type == :
## prediction from a rank-deficient fit may be misleading

## Warning in predict.lm(object, newdata, se.fit, scale = 1, type = if (type == :
## prediction from a rank-deficient fit may be misleading
## Warning in predict.lm(object, newdata, se.fit, scale = 1, type = if (type == :
## prediction from a rank-deficient fit may be misleading
```

```
##             Error Improvement
## cv-KNN  0.3416473   -1.540677
## cv-Logi 0.4614605   -2.431673
```

Random Forest

With Random Forest applied to model the data, 500 trees were built and the threshold for classification was adjusted according to the ROC curve. The resulted Classification errors are much lower than the previous models, which is more favorable.

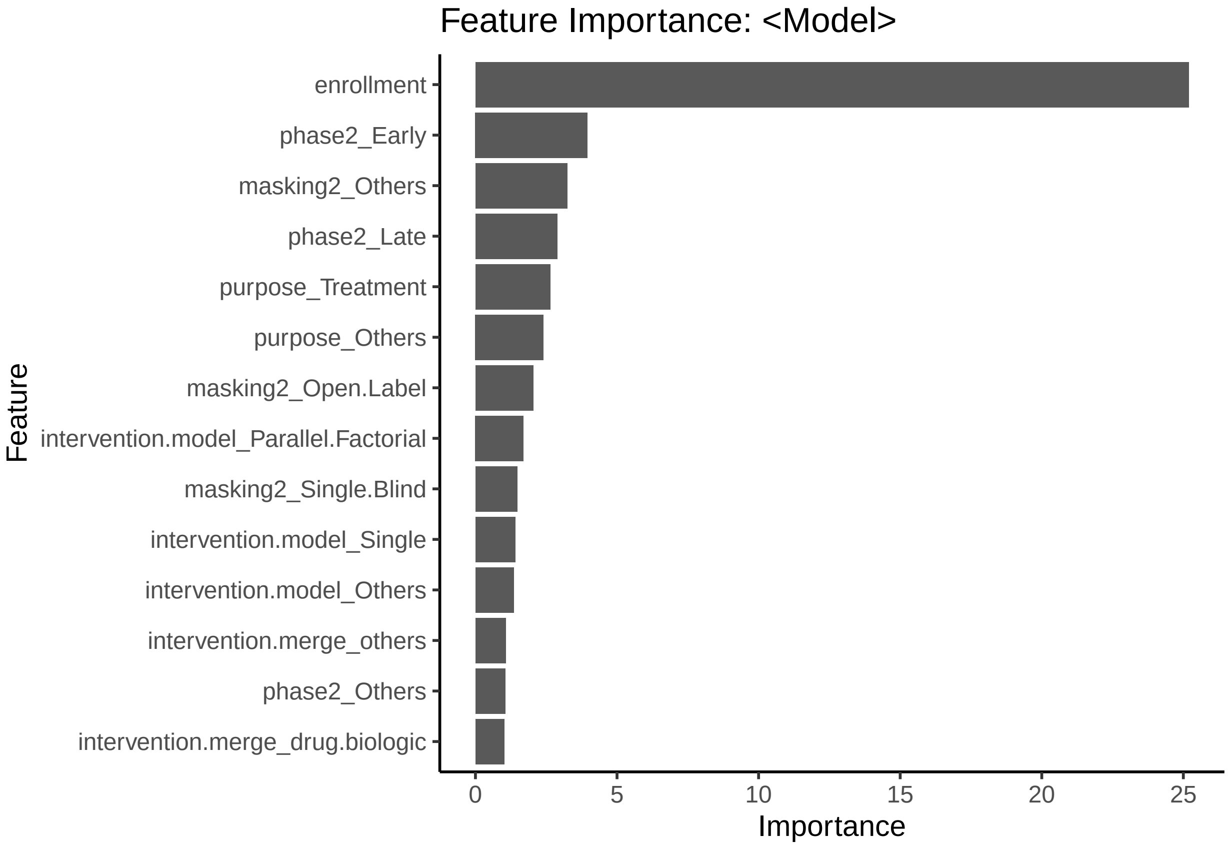
```
## Setting levels: control = 0, case = 1
```

```
## Setting direction: controls < cases
```

```
##   threshold
## 1     0.007
```

```
##               Train.Error Test.Error
## Random Forest   0.1316029  0.1535836
```

Also, information on feature importance in the training process could be obtain from the model. The features that are the most informative in this model are `enrollment`, `early-phase` (indicating trials in phase 1 or 2), `masking2-Others`, and `late-phase`(trials in phase 3 or 4).



Neural Network

Lastly, single-layer neural network was applied to model the output. 15 input layer neurons, 10 hidden layer neurons and 2 output neurons were used. The activation function used for Input and Hidden layers is ReLu, while Sigmoid was applied for the Output layer. Adam optimizer was chosen in refining the model. The resulting test error obtained was 0.13, which is also the lowest among all previously fitted models. However, the model could have been underfitted - the loss and accuracy learning plot drawn while fitting the model showed minimal changes and parallel lines for both training and validation sets, indicating a minimal optimization and learning during the modelling.

```
## Loaded Tensorflow version 2.8.0
```

```
## Model: "sequential"
## ________________________________________________________________________________
##  Layer (type)                       Output Shape                    Param #
## ================================================================================
##  dense_2 (Dense)                    (None, 200)                     3000
##
##  dropout_1 (Dropout)                (None, 200)                     0
##
##  dense_1 (Dense)                    (None, 140)                     28140
##
##  dropout (Dropout)                  (None, 140)                     0
##
##  dense (Dense)                      (None, 2)                       282
##
## ================================================================================
## Total params: 31,422
## Trainable params: 31,422
## Non-trainable params: 0
## ________________________________________________________________________________
```

```
## Test loss: 0.3921981
```

```
## Test accuracy: 0.8651617
```

```
##                Train.Error Test.Error
## Neural Network   0.1424332  0.1348383
```

Model Evaluation

In order to compare the models fitted, a table compiling the classification errors for each model is shown below. Comparing the testing error, it seems that Neural Network with the lowest error rate should be the best algorithm. However, since the model has a high probability of being undertrained, also the fact that it has a lower testing error than training error is highly suspicious, it is believed that the model fitted using Random Forest should be concluded as the best model instead. It has the lowest training error, showing that the data fitted the model well; also a low testing error, showing its high prediction accuracy. Also, as Random Forest is a simple and non-parametric algorithm - it makes no assumption of our dataset, this characteristic also increases the validity and credibility of this model.

```
##                Train.Error Test.Error
## Neural Network   0.1424332  0.1348383
## Random Forest    0.1316029  0.1535836
## cv-KNN                  NA  0.3416473
## cv-Logi                 NA  0.4614605
## adj.Logi         0.4489982  0.4648464
## adj.KNN          0.4658470  0.5037543
```

Discussion

With the final model obtained, NIH's funding behavior in the GI clinical trials could be predicted with an accuracy rate of 84.8%. The features that affect the NIH's funding the most are the size of enrollment, the phase, and the masking type of each trial. Moreover, there are some limitations and thoughts for next steps to be addressed.

Firstly, it is true that One Hot Encoding is an effective data transformation and preprocessing technique that helps our Models understand the data better. However, it also has its setback: the dataset is likely to face the problem of having highly correlated dummy-coded varaibels, which causes multicollinearity. One way to tackle the problem is to drop one of the dummy variable in each original variables. However, subject matter specialist should be consulted and further investigations needed to decide on the class to be dropped. Secondly, it is believed that the parameters, such as the ones in neural network, or the ones in cross validation and the prediction threshold should be tuned to find the optimal values so as to improve the model performances. Lastly, feature selection could be carried out according to the random forest's measure of feature importance, or other methods such as Lasso or stepwise selection.