STATS 202 Final project

Yu-chi Tsao

Team member: Tzu-Ling Liu

Stanford University

Obejective1

Purpose

Using the data to find out does the treatment have any effect on schizophrenia?

Selection

The goal is to investigate whether the treatment has any effect on schizophrenia. If the treatment has effect on schizophrenia, the Positive and Negative Syndrome Scale (PANSS) score for patient with treatment should be lower than the patient with control when the visit day increases. To find out if this is the case, the linear regression model shown in Figure 1-1 can let us know if this statement is true or not. Visitday, treatment and the interaction term between Txgroup and Visitday are used as the predictors and PANSS score is used as the response. From this model, the slope for patient with treatment is β1 + β2 +β3 and the slope for control group is β1. Hence, if $\beta_2 + \beta_3 > 0$, we can confirm that the treatment has effect on schizophrenia. If $\beta_2 + \beta_3 < 0$, we can understand that the treatment has no effect on schizophrenia.

$$\text{PANSS\_total} = \beta_0 + \beta_1 * \text{VisitDay} + \beta_2 * \text{Treatment} + \beta_3 * \text{TxGroup} * \text{VisitDay}$$

$$\begin{cases} \text{TxGroup} = 1\ (\text{Treatment}) \rightarrow \text{PANSS\_total} = \beta_0 + \beta_1 * \text{VisitDay} + \beta_2 * \text{Treatment} + \beta_3 * \text{TxGroup} * \text{VisitDay} \\ \text{TxGroup} = 0\ (\text{Control}) \rightarrow \text{PANSS\_total} = \beta_0 + \beta_1 * \text{VisitDay} \end{cases}$$

Figure 1. The function of the linear model for objective 1.

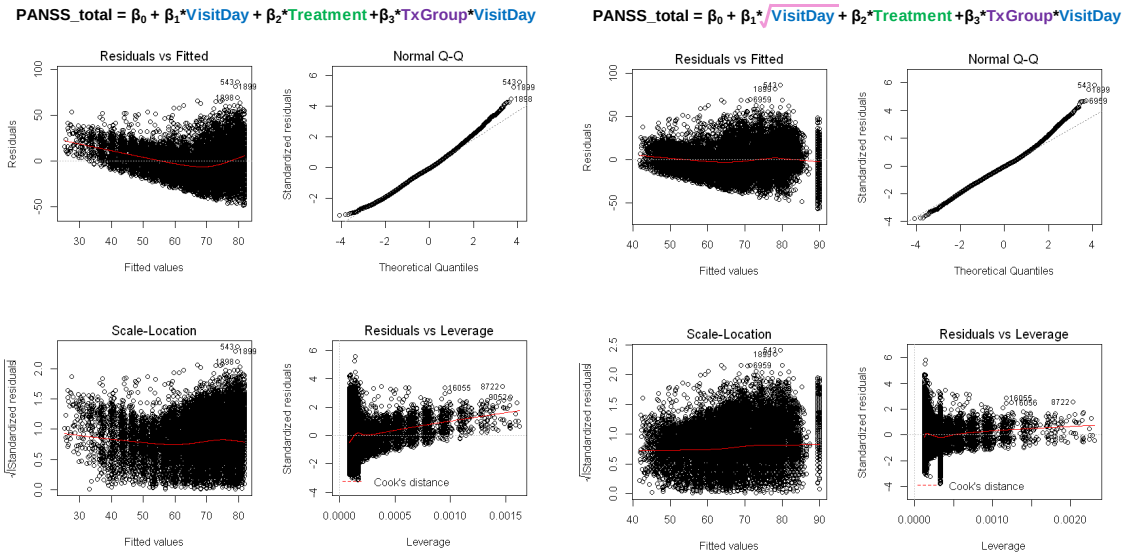
Preprocessing

To study the overall treatment effect on the schizophrenia, the study A to study E are combined together.

Transformation/Data Mining/Evaluation

Checking the non-linearity relationship between response and predictors: The linear regression model assumes that there is a straight-line relationship between the predictors and the response. The residual plot showed a strong U-shape, providing strong indication of non-linearity in the data. Hence, non-linear transformation of the predictor is tested. By using the square-root of variables for visit days, the residual plot did not show strong U-shape. The residual standard error decreased from 15.46 to 14.87 and the R^2 is improved from 0.3318 to 0.3815.

From the results, the regression model with transformation on the predictors -visit days is used to investigate whether the treatment has effect on the schizophrenia. Figure 2 shows the comparison between the linear regression model and the model with transformation on the predictor -visit days. Figure 3 shows the model results using regression with transformation on visit days.



	Linear regression	Regression with transformation on visit days
Residual standard error	15.46	14.87
R^2	0.3318	0.3815

Figure 2. The comparison between the linear regression model and the model with transformation on the predictors - visit days.

```
Residuals:
    Min      1Q  Median      3Q     Max
-56.513  -9.513  -0.513   8.600  86.506

Coefficients:
                             Estimate Std. Error t value Pr(>|t|)
(Intercept)                 89.512956   0.268467 333.422  < 2e-16 ***
sqrt(VisitDay)              -2.518218   0.058689 -42.908  < 2e-16 ***
TxGroupTreatment             0.405221   0.273492   1.482    0.138
VisitDay                     0.016806   0.003438   4.889 1.02e-06 ***
TxGroupTreatment:VisitDay -0.002360   0.002114  -1.116    0.264
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 14.87 on 22904 degrees of freedom
Multiple R-squared:  0.3815,    Adjusted R-squared:  0.3814
F-statistic:  3532 on 4 and 22904 DF,  p-value: < 2.2e-16
```

	β_2	β_3	Overal p value	p value for visit days
Model result	0.405	-0.002360	<2.2 e^{-16}	1.02 e^{-6}

Figure 3. The model results

Interpretation/Conclusion:

From the model results (Figure 3), we can see the $\beta_2 + \beta_3 > 0$, indicating that the treatment has effect on the PANSS score. This means that the treatment has effect on schizophrenia.

Moreover, the overall p value is less than 0.05 and the p value for the visit day predictor is less than 0.05, meaning that this predictor is statistically significant. This indicates that the PANSS score is related to the visit days. And since the $\beta_2 > \beta_3$, the treatment may have stronger effect in the earlier visit days compared to later visit days.

Obejective2

Purpose:

Creating clustering group to investigate which clustering group has stronger treatment effect and investigate the percentage population for that clustering group. By understanding which group of patients can have stronger treatment effect, the stakeholders like sellers can strongly recommend the treatment for those patients because they have stronger treatment effect. Moreover, clinicians can understand and research more on why those group can have stronger treatment effect compared to other clustering group.

Selection:

◆Variable selection: The mean of positive syndrome (Pmean), negative syndrome (Nmean), general syndrome (Gmean) and PANSS score are selected to be the clustering variables. The reason for selecting those 4 variables is that first, the PANSS score were summed up by those 3 main category syndrome measurements. Second, this allows to understand how does the mean of those major syndrome score will be clustered together. Lastly, by clustering those variables, this may allow to distinguish which group of patients can have stronger treatment effect.
◆Model selection: K nearest neighbors (KNN) model and Hierarchical clustering method (complete linkage and single linkage) were implemented and compared. Within-cluster sum of variance was used to examine the error for KNN.

Preprocessing/Transformation

The data is preprocessed into Pmean, N mean, Gmean and PANSS score to be the variables. All the variables are rescaled before creating the clustering model.

Data Mining/Evaluation

◆Comparison between different models: Hierarchical clustering method (complete linkage and single linkage) and K nearest neighbors (KNN) model is implemented and compared. Compared between complete linkage and single linkage hierarchical clustering algorithm, strong chaining effect is shown for single linkage model (Figure 4), indicating the single linkage method is not suitable to be used as the clustering model for this objective. Compared between complete linkage hierarchical clustering and KNN by plotting the PANSS score v.s Nmean/Pmean/Gmean (Figure 5 and 6), the KNN separated the group better than the complete linkage hierarchical clustering model. Hence, the KNN model is selected to interpret the results. The reason to choose k = 4 is first, from the complete linkage model, it seems having k=4 is reasonable number to cluster into 4 major group. Second, from KNN model, the within-cluster sum of variance is dramatically decreased when k is changed from 1 to 4. Lastly, having lower k is easily interpretable.

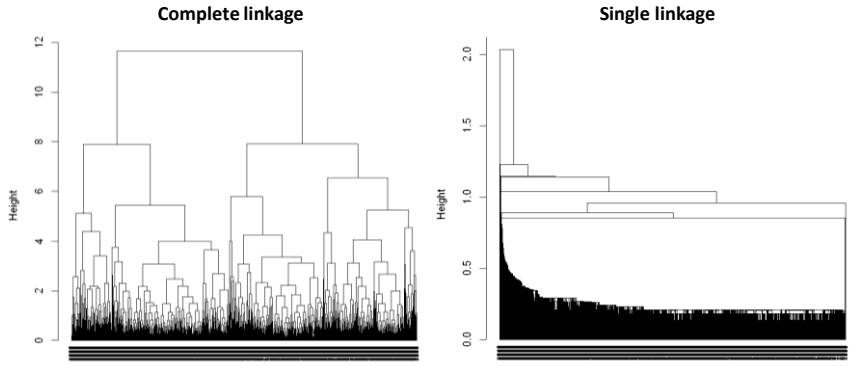


Figure 4. The dendrogram of complete linkage and single linkage hierarchical clustering.

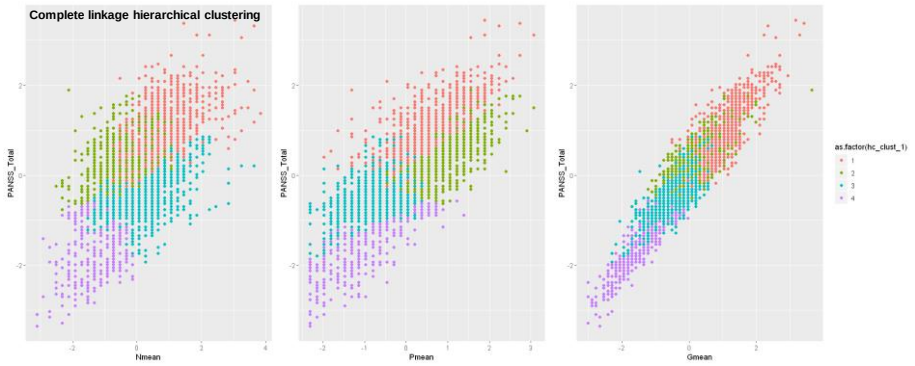


Figure 5. Plot of PANSS score v.s Nmean/Pmean/Gmean with complete linkage hierarchical clustering model as k=4.

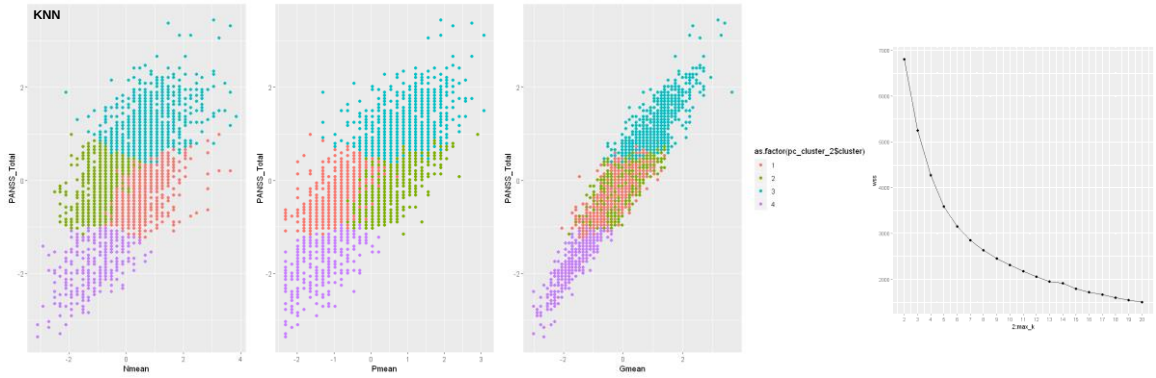


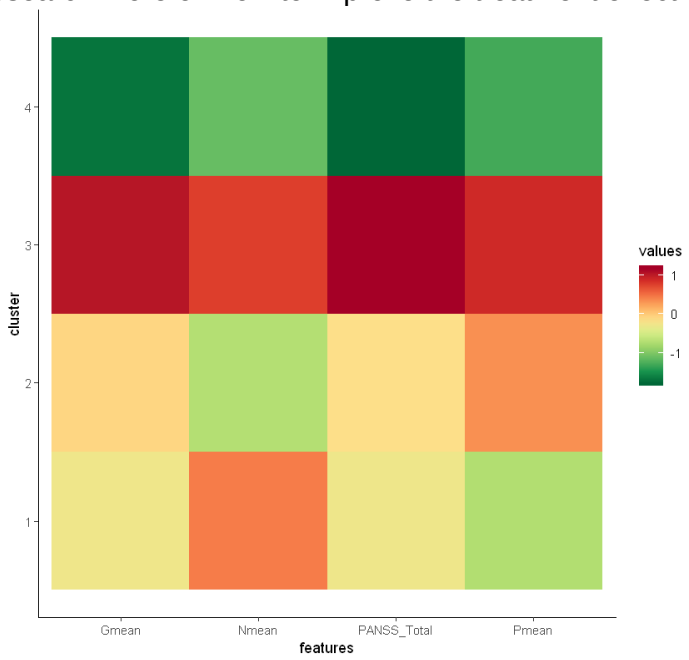
Figure 6. Plot of PANSS score v.s Nmean/Pmean/Gmean with KNN model as k=4.

Interpretation/Evaluation

◆Cluster group - Pmean, Gmean, Nmean and PANSS score value comparison:
From the heat map and the centroid value for each cluster group (Figure 7), the cluster group 4 has the lowest Pmean, Nmean, Gmean and PANSS score compared with other groups. Group 3 has the highest Pmean, Nmean, Gmean and PANSS score compared

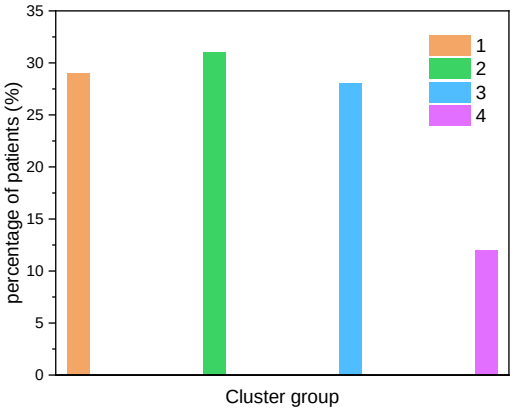
with other groups. Group 1 and 2 have intermediate values of Pmean, Nmean, Gmean and PANSS score. Compared between group 1 and 2, group 1 has lower Pmean, higher Nmean, and lower Gmean compared with group 2. For the percentage of patients in each group (the number of patients in the cluster group divided by the total number of patients), group 1, 2 and 3 have similar percentage of patients (28%-31%), but the group 4 has less percentage of patients (12%) (Figure 7).

◆Compared the treatment effect between the cluster group:
We used the same model from the objective 1 and investigated the treatment effect between the cluster groups by comparing the total value of $\beta_2 + \beta_3$. Higher total value of $\beta_2 + \beta_3$ means that the treatment effect is stronger. From the results (Figure 8 table), the strength order of the treatment effect is group 4> group 1>group 3> group2. Hence, the stockholders like sellers could strongly recommend group 4 and 1 to have the treatment because they tend to have stronger treatment effect on schizophrenia. Moreover, clinicians could research more on how to improve the treatment effect on group 2 and 3.



Group	Pmean	Nmean	Gmean	PANSS_Total
1	-0.7242403	0.4381298	-0.27458985	-0.2843609
2	0.3414757	-0.7077630	-0.08163574	-0.1440839
3	0.8979758	0.7733418	1.05237118	1.1687888
4	-1.3042634	-1.1190913	-1.68999628	-1.7840076

Figure 7. Heat map and the centroid value for each cluster group



Group	β_2	β_3	$\beta_2 + \beta_3$	RSS
1	0.495268	-0.002611	0.492657	13.26
2	0.259754	-0.001085	0.258669	14.46
3	0.330530	-0.001635	0.328895	14.56
4	1.304367	--0.008452	1.312819	12.23

Figure 8. Percentage of patients in each group (the number of patients in the cluster group divided by the total number of patients). Table: The regression result for different clustering group.

Obejective3

Purpose

Predicting the 18th PANSS score.

Selection

◆Input feature selection: Different input features size were compared based on their optimized learning rate (LR).
1. Sum of positive syndrome, sum of negative syndrome and sum of general syndrome, visit day and PANSS score are used as the input features.
2. Positive syndrome (P1-P7), negative syndrome (N1-7), general syndrome (G1-G17), visit days and PANSS score.
◆Model selection: Considering the data is sequential, recurrent neural network is selected as the model prediction.
◆Root mean square error was used to compared the performance on training and development set. Figure 9 shows the design for RNN model. Adam was determined as the optimizer in the training process.
◆Hyperparameter tuning: training set and development set were used to evaluated the optimized learning rate.

Preprocessing/Transformation

The study A to D were combined to be used as training set and development set (cross-validation set). Before splitting the dataset, the dataset was preprocessed by normalizing

and shuffling. Then the data was split and allocated with 70% of total to the training set and 30% to the development set.

To structure the input data, for example, for feature 2 engineering, a two-dimensional array was created with rows that contains 33 features, including visit day (VD), positive syndrome (P1-P7), negative syndrome (N1-7), general syndrome (G1-G17) and PANSS score (P). To predict 18th PANSS score, the output (label) is the actual PANSS score during the last visit day's score. The arrangement of the feature was shown in figure 9.

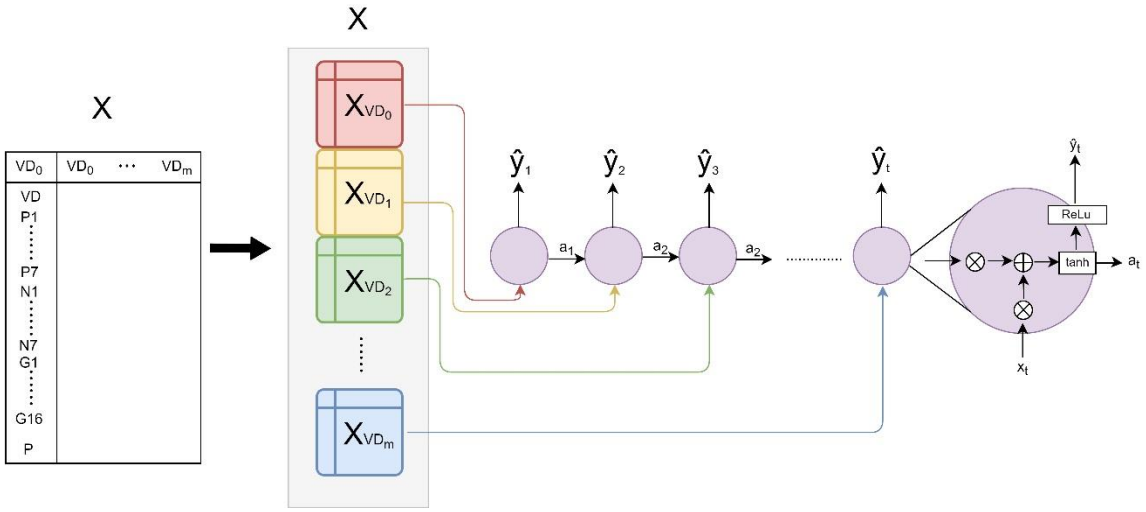


Figure 9. Arrangement of features and scheme of recurrent neural network (RNN model)

Data mining/ Interpretation/ Evaluation

Table 1 summarized the RMSE of training and development set with optimized learning rate for two different feature engineering. From the results, having larger feature size (Input feature 2) reduced the training error from 4.74 to 1.48 and dev. error and from 4.4 to 1.44. Moreover, for the study E (test set) in the Kaggle leaderboard, the MSE improved from 32.82946 to 30.57691 as changing from feature 1 to feature 2 input.

RMSE	Input feature 1	Input feature 2
Optimized LR	0.005	0.05
Training Set	4.74	1.48
Dev. Set	4.4	1.14

Table 1. training and dev. set RMSE with optimized learning rate for two different feature engineering. Input feature 1: sum of positive syndrome, sum of negative syndrome and sum of general syndrome, visit day and PANSS score are used as the input features.
Input feature 2: Positive syndrome (P1-P7), negative syndrome (N1-7), general syndrome (G1-G17), visit days and PANSS score.

Kaggle leader board results:

MSE score: 30.5769, Username: yuchi tsao/Team name: Mater.

Obejective4

Purpose:
Binary classification for predicting the probability of the assessment being either flagged or assigned to CS.

Selection:

◆Input data: considering the purpose of the prediction is to predict the assessment being either flagged or assigned to CS, Positive syndrome (P1-P7), negative syndrome (N1-7), general syndrome (G1-G17), and PANSS score are selected to be the input features because the relationship between those variables may allow to get some relationship.
◆Selecting model: Fully connected neural network (FCNN) was used as the model.
◆Four hidden layers were created and relu was chosen to be the activation function for them. For the output layer, sigmoid was selected to be the activation function. Binary cross entropy loss was used as the loss function for binary classification. Adam was determined as the optimized in the training process. Figure 10 shows the design for FCNN model.
◆Accuracy was used to compared the performance on training and development set.
◆Hyperparameter tuning: training set and development set were used to evaluated the optimized learning rate.

Preprocessing/Transformation

The study A to D were combined to be used as training set and development set (cross-validation set). Before splitting the dataset, the dataset was preprocessed by normalizing and shuffling. Then the data was split and allocated with 75% of total to the training set and 25% to the development set.

To structure the input data, for example, a two-dimensional array was created with rows that contains 33 features, including visit day (VD), positive syndrome (P1-P7), negative syndrome (N1-7), general syndrome (G1-G17) and PANSS score (P). Because there are total 16757 data in the training data, so a size of (33, 16757) is created as the input features. The arrangement of the feature was shown in figure 10. The output label, Y=0 is the passed in Lead status and Y=1 is the flagged or assigned to CS in the lead status (LD). Figure 11 shows the detail for the data set.

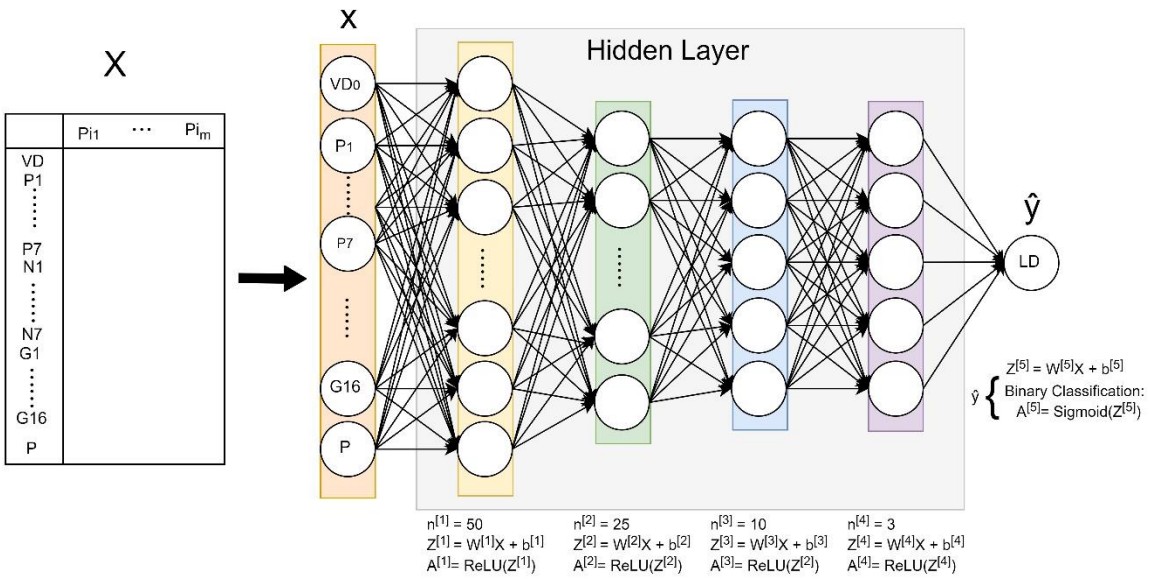


Figure 10. the design for FCNN model.

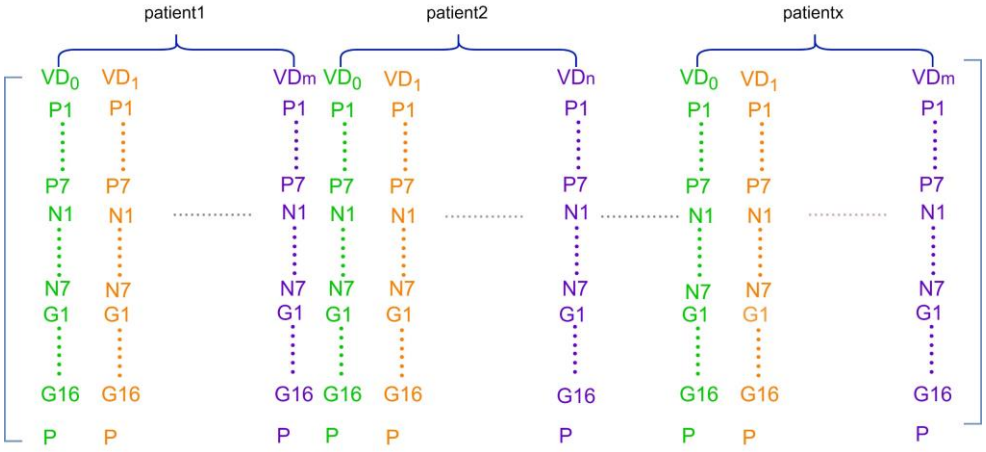


Figure 11. Details for the input data set.

Data mining/ Interpretation/ Evaluation

Table 2 summarized the training and development set accuracy with different learning rate (LR). We found the amount of data with passed in the lead status is greatly larger than the data with either flagged or assigned to CS. Hence, under-sampling is further implemented to solve the imbalanced dataset. Under-sampling means that we will select only some of the data from the majority class. Interestingly, training and development accuracy can be further improved to 94% when LR=0.01 and 91% when LR=0.001 (Table 3).

Learning rate	0.1	0.01	0.001	0.0001	0.00001
Training Set Accuracy	84	90	75	24	75
Dev. Set Accuracy	76	75	74	23	75

Table 2. The training and development set accuracy for model without under-sampling.

Learning rate	0.1	0.01	0.001	0.0001	0.00001
Training Set Accuracy	50	94	91	88	64
Dev. Set Accuracy	49	94	91	88	63

Table 3. The training and development set accuracy for model with under-sampling.

Kaggle leader board results:

Log loss: 0.76297, Username: yuchi tsao/Team name: Mater.

Contribution:

For objective1 and 2, we created the code individually but shared the results and discussed the results with each other. For objective 3 and 4, we developed the code together, tuned the model parameter together and discussed the result together.