Final Project

STATS 202: Data Mining and Analysis

Stanford University

Outline:

1. Treatment Effect
2. Patient Segmentation
3. Forecasting
4. Binary Classification

By Kelvin Ortiz
Partner: Artem Grigor
Stanford ID: 06239139
Kaggle ID: “Unlabeled” Team: “Moonchild”
3rd Classification, 5th Forecasting

Treatment Effect

The first assignment asked me to provide some reasoning on the validity and significance of treatment group in the effectivity of treatment for schizophrenia.
I started concatenating all studies together to get an estimate that encompassed all data given.
I took three approaches to identify the significance of the treatment variable in predicting efficacy of treatment:

1. T-testing moving average in relation to time and scores
2. OLS regression with 2 and 5 variables testing PANSS Total and testing each of the 30 PANSS Scores
3. XGB Feature importance F score with 5 variables predicting PANSS Total

All approaches taken gave negative results and indicated that the anonymized treatment does not have an effect of schizophrenia. I used T-testing and Z-score to obtain the results found.

Features used

For a simple OLS regression I used all patient data from all studies and selected a few variables that I considered might have some impact in predicting PANSS Total or each individual Score. The features that I selected were:

- TxGroup: Binary variable to be tested on for significance.
- Visit Day: Useful because it is changeable over time
- Country Mean (mean of PANSS Total): Country might have an effect in predicting PANSS Total.
- Rater Mean (mean PANSS Total): Similarly, Rater mean might have some effect in predicting PANSS Total.

Approach 1:

The first approach that I did was to try to account for differences in rate of change on several variables such as PANSS Total, Positive total scores, Negative total scores, and General total scores relative to Visit Day. This approach was done so in order to see if the treatment variable has any effect in the rate of change of those total variables.

1. We grouped each patient and calculated the moving average of total scores relative to visit Day. I then obtained the mean of that rate of change for each patient, concatenated all means for each patient, and then calculated significance among two populations: Treatment and Control.
2. I also accounted for relevance in time for the treatment to take effect. I redid all measurements accounting for "Visit Days" that were after 90, given that in order for treatment to have any effect, the patient should be under treatment for at least 60-90 days as parts of the reading state.
3. I also accounted for days that were before 90 using the same approach as before, to see if there is any difference from before 90 and after 90 days under treatment.

In order to show significance, I considered 0.05 to be the threshold for the p-value. The results showed that there is no significance in treatment group relative to the rate of change in Negative, Positive, General, nor PANSS Total for all data, for data before and after 90 days.

- BEFORE 90 DAYS, MEAN TREATMENT:0.3236 CONTROL: 0.3401. P-VALUE: 0.4052
- AFTER 90 DAYS, MEAN TREATMENT: 0.0469 CONTROL: 0.0347. P-VALUE: 0.5784
- ALL DATA, MEAN TREATMENT: 0.1985 CONTROL: 0.2034. P-VALUE: 0.7165

Approach 2:

The second approach involved fitting regression lines and checking for significance in the treatment predictor relative to the PANSS Total, and each individual score.

1. I first fit an OLS regression with two predictors:

Y = PANSS Total X = TxGroup, Visit day

   This was a naïve approach, as I ran a t-test without considering other variables. The p-value was indeed very high at 0.85, showing that TxGroup is not significant when only accounting two variables.

2. I added some additional predictors

Y = PANSS Total X = "TxGroup", "Visit day", "Country mean", "Rater Mean", "Site Mean"

   After adding some additional variables, I was still left with a very high p-value for the treatment group of nearly 0.58.

3. Same variables as before, but now I test the hypothesis on each of the 30 PANSS scores, along with Positive, Negative and Bipolar index scores.

Y = each of the 30 PANSS score, Pos/Neg Total, and Bipolar index X = "TxGroup", "Visit Day", "Country mean", "Rater Mean", "Site Mean"

   After testing the significance of treatment in predicting each of the scores, I clearly observed that all of the 30+ regression lines had the TxGroup showing high p-values. However, through looking at each of these regressions, I noticed that when the outcome label was "P5", "N6", "G3", and "G5", the p-value wasn't as high as with the other results. However, they were still high ranging from 0.15 to 0.3 p-values and did not present convincing evidence to reject the null hypothesis under the variables selected and for each score as the outcome label.

Approach 3:

After having done data analysis in question 3 and 4, and through tree-based feature selection along with ANOVA and CHI-Squared, I was able to observe that Treatment was neither important in predicting each patients Visit Day nor whether the patient Assessment would be flagged or not.

I fit the same predictors as in 2 and 3, having the y label being PANSS Total under XGBoost feature importance, which calculates the F score for each variable. XGBoost returned very low importance for the TxGroup predictor, overall asserting the previous findings that the anonymized treatment has little significance in measuring any effect on treatment of schizophrenia.

Conclusion:

After trying all three approaches mentioned I was greatly convinced, through rejecting the null hypothesis with high p-values, that the anonymized treatment has no effect on schizophrenia, based on the methods and variables selected.

Patient Segmentation

The variables that I used for clustering were the 30 PANSS Scores, along with the Positive Chronic, Negative Chronic, and Bipolar index variables created.
To further explain the features created:

- Positive Chronic: Patients that have 3 Positive scores with 4 or more points. Binary
- Negative Chronic: Patients that have 3 Negative scores with 4 or more points. Binary
- Bipolar index: Difference between Total Positive Score and Total Negative Score. This index is positive for patients that have larger Positive Score.

Just through looking at the data, and through some thought process, I had anticipated that K-Means would separate the patients into patients with higher positives, patients with higher negatives, and patients with even positives and negative scores.

K-Means Clustering, K=3:

After near 100 iterations of K-MEANS, I was able to better distinguish how K-Means allocated the patients into the clusters. I began with 3 clusters, as it seemed reasonable to me to separate the patients in three groups.
It wasn't uncommon to find iterations where low bipolar index was highly associated with low Positive Chronic, and high bipolar index associated with low Negative Chronic. However, I was unable even after a hundred iterations, to find one iteration that made a perfect separation of positive, negative, and average, and so I increased K.

K-Means Clustering, K=4:

With K=4, I convinced myself that I had found a better clustering of patients.
The clustering goes as follows:

Cluster 1: High Bipolar index / high Pos Chronic & low Neg Chronic (871 observations)
Cluster 2: Low Bipolar index / low Pos Chronic & high Neg Chronic (806 observations)
Cluster 3: Average Bipolar index / high Pos & high Neg Chronic (735 observations)
Cluster 4: Average Bipolar index / low Pos & high Neg Chronic (588 observations)

Further looking at the means of the variables when grouping by cluster, it was clear to me that K-Means had clustered the patients into how I was expecting to see:

- **Cluster 1:** High Bipolar index. This cluster accounts for an overview of patients that are very likely to be Positive Chronic patients (those with 3 or more positive scores with scores of 4 or more). This cluster has patients that very unlikely to be Negative Chronic. As a consequence, these patients have high positive scores and low negative scores. This cluster shows an average PANSS Total scores and also average general scores; however, G5 (Mannerisms and Posturing) and G7 (Motor Retardation) are particularly low.
- **Cluster 2:** The opposite side of cluster 1. These are patients with low Positive scores and high Negative scores. This cluster focuses in patients that are very high in negative scores, and as a result becoming Negative Chronic patients, while at the same time showing low positive scores, and highly unlikely to be a positive Chronic patient. This cluster shows lower PANSS Total compared to cluster 1, which might indicate that patients with higher negative scores are likely to have lower Total score compared to patients with higher positive scores.

This cluster shows average general scores; however, G9 (Unusual Thought Content) is particularly low.

- **Cluster 3:** This group accounts for patients that have both high positive and negative scores (very likely to be positive or negative Chronic patients), and as a result they have high positive, negative, and general scores. This cluster has an overall high PANSS total score.
- **Cluster 4:** This cluster contrasts cluster 3, because it accounts for patients that have both low positive and negative scores) (very unlikely to be positive or negative Chronic patients). They have low positive, negative, and general scores. Overall having a low PANSS total.

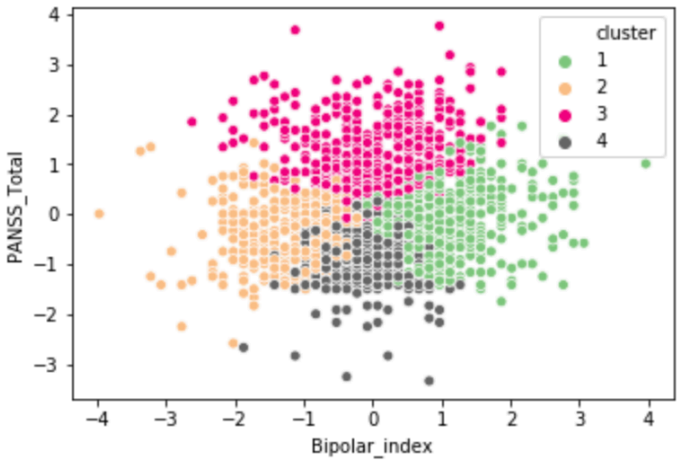
Additional score findings:

G1(Somatic Concern), G3 (Guilt Feeling), and G6 (Depression) showed almost no change among clusters. These three scores seemed not to change much among patients throughout the dataset, and within each of the 4 clusters.

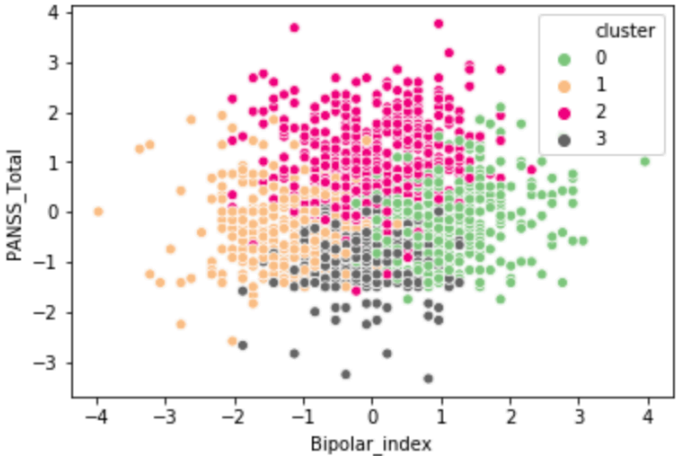
In cluster 1, Patients with high positive scores did not necessarily have to have high P2 (Conceptual Disorganization).

In cluster 2, Patients with high negative scores did not necessarily have to have high N7 (Stereotyped thinking).

K-Means with K=4



Gaussian Mixture Model, 4



components

Gaussian Mixture Model – 4 Components

As the scatterplot shows, GMM showed very similar results, although it isn't as clean as the separation seen in K-means.

The last notes seen in K-means were also visible using GMM, showing that P2 and N7 are not requirement scores in order to be highly positive/chronic or being highly negative/chronic among patients in the first two main groups.

Likewise, G1 and G3 are scores that are the most unchangeable among all patients within all four clusters.

Additional notes:

I had initially only included datasets A-D, but then I included E, and I got similar results. When I used 4 clusters on K-means and GMM, I would be finding similar separation among group of patients as well as findings with G1 and G3 having low variation among patients.

Forecasting

Forecasting was the question that I tackled last, and so a large amount of the dataset that I arranged for Classification, was used for this prediction as well.
The simple approach that I used was to use all patient data accounting for all studies and predicting the following PANSS Total Score from each training example. I was surprised to find that once I had arranged the model with multiple features, I placed 7th on the leaderboard. The following day I added three additional features, and I went up to place 5th.

Data Processing:

To begin with, I added all studies together, including study E, and created a separate test set E maintaining an equal number of predictors on both datasets. I then turned categorical variables into numerical, added new time predictors, and obtained each patient's following visit as a new label to be predicted, deleting their last visit, as their last visit did not have a next visit to be added as an output label.
I used different regression algorithms, but the most compact model was XGB, as it was able to take in a large number of predictors without much difficulty in trying to find regularization due to overfitting.

Categorical variables

The most useful categorical variables to be turned into new features were Rater ID and Site ID

1. **Rater ID:** From this feature I was able to obtain the mean PANSS Total assigned by each rater. I considered It important to keep this variable, as it accounts for bias and relevant information that could impact each patient's final score. As it turned out, the mean Rater ID showed high F score under feature importance and raised my score overall.
2. **Site ID:** Similar to Rater ID, I obtained the mean PANSS Total assigned by each site. From the start, I was suspicious of adding an additional variable that accounted for the mean Total, having observed that this variable contained very similar information to Site Mean. I was skeptical of including it along with Rater mean as it might amount for variables that are highly correlated.
3. **Country:** I initially turned this categorical variable into the main two countries observed in the Test set (USA and Russia), along with two other major continents (Europe and Asia). However, after observing the feature importance, F score, I figured that these binary predictors had very small value in predicting the label Y.
   I then tried accounting for each grouped continent mean of PANSS Scores, but it also had a low importance. Lastly, I went back and obtained the mean of each country for all the dataset. The last feature: "Country Mean", showed high importance under higher F score, and therefore I am keeping it for now.
4. **Patient ID:** This variable was useful for grouping the patients and creating new variables relevant to each patient. One of those variables that I created was the rate of change in total score. This variable accounted to the percentage change in Total Score from their last visit. (Since the first percentage change was *NaN*, I turned it into zeros. I also did something similar to "Visit Day". I created a time of visit rate of change, which also accounted for percentage change from the previous visit. Lastly, I ranked each patient's Visit Day ranging

from 1 to 20. This last variable was useful in ordering each patient's visit (although it did not account for missing values and long absences in the treatment).
On the day of submission, I introduced one last variable: "Patient Score Mean", which accounted for the mean score for each patient.

5. **TxGroup:** This variable was turned into a binary predictor, 1: Treatment, 0: Control, although this binary variable had little importance in predicting the following PANSS Score. F-Score under XGB showed very little importance every single time I tried to do either regression or classification tasks. I then figured that TxGroup might indeed have no relevance for treatment effects. I took this finding to further complement my previous finding in question 1.
6. **Study:** The variables were turned into one-hot vectors, taking aside a base E variable. Similarly, this predictor was not useful at predicting each patient's last visit Score.
7. **Assessment ID:** Taken out for being unique values.

Numerical Variables

I initially used all numerical variables and then tried to create more variables related to Visit Day and PANNS Total, trying to find interaction predictors with high importance in predicting the last score.

1. **Visit Day:** This variable was difficult to interpret at first, as I was trying to organize each patient's visit in a chronological way, which efficiently dealt with discrepancies in attendance in treatment as well as figuring which Visit day would be the first for each patient, as some patients had long gaps before they actually began their treatment.
   Visit Day was used primarily to create the rank column, which ranked the Visit Day for each patient. Some transformations that I made were: Time percent change, which showed the percent change of visits from the prior one. The base time was turned into zero. **I** tried creating additional variables by transforming this variable in different ways. It actually worked out, as I was able to observe improvements in the Kaggle Leaderboard.
2. **PANSS Total:** Similar to Visit Day, I created another variable that accounted for percentage change of PANSS Total for each patient, with base percentage change switched to zero as there would be null values.
   In order to make the model work, I had to add a new variable that accounted for each patient following Visit Day. Since the last visit day had no following day, it was taken out.

The near-final dataset contained the following:

Model: 1

- Y vector, 19962 labels with following scores.
- X matrix: 19962 examples, 54 variables.
- Real – Test matrix: 1962 examples, 54 variables

Although the dataset seemed to have too many predictors, and I was indeed worried that many predictors could potentially diminish the effectivity of other significant predictors, I started implementing the model with all variables, and gradually taking out some predictors through using different feature selection features such as Backward Elimination, F score feature importance under XGB, and I also looked at results under ANOVA and Chi-Squared.

Contrary to what I was expecting, XGB did not show any relevant improvement through reducing the number of features. I realized that the number of predictors was perhaps not too large compared to the number of examples, and so I decided to keep it, and perhaps do some feature reduction with PCA (specially for features below the top 4 predictors).

Cross-Validation Score with 10 folds

Linear Regression with multiple variables: 85% accuracy
LASSO Regression (alpha=0.1): around 86% accuracy
AdaBoost Regression: 77% accuracy
Random Forest Regression: ~ 85%
XGB Regression: ~ 88%

XGB was clearly showing higher results with the cross-validation score, and so most of my submissions were done on XGB.

Model 2:

I tried a second approach, as I overheard prof. Linh talking about using all observations per patient to predict the last one, after getting the mean of each patients predicted observation.
I implemented it and trained the model on all studies predicting each patient last visit for each training example. I then tested the model on study E, having all patients visits and predicting the last. I obtained a prediction for each visit day, and so I calculated the mean of each patient as a result for the final score.
The result wasn't what I was expecting. I got a score on the leaderboard through this approach. Perhaps I had to make some changes in the model.

Final comments:

As the Kaggle competition was about to conclude, I made additional changes to the dataset, as I tried to improve my positioning in the board. I tried finding some irregular data that I could take out in order to have a better approximation to the test set. I observed that STUDY A was showing a higher average of "mean average PANNS Total" per patient, and so I took "study A" out and ran the regression. The results were indeed better. I ended up placing 5th in forecasting.

Binary Classification

Data Processing:

Classification was the first problem that I tackled, and I truly found it the most fun assignment to do, plus I gained the most information about the entire dataset and the patients with schizophrenia. I focused a large amount of time in this task, and always discovered something new that could help me improve my score. I did many changes not only to the selection of features, but also to the shape of the data, attempting to select the most valuable information from the dataset given. I tested several hypotheses in attempting to approximate my training set's population distribution to the test set distribution in order to obtain better prediction results. In order to do this, I had to have a strong understanding of the data, preprocess the data, and apply a classifier algorithm. On the last step, I realized that doing hyperparameter tuning did not make much improvements, and it only largely caused my model to lose robustness, as I had not achieved a good population approximation prior to tuning the parameter.

I started by concatenating all studies (a, b, c, and d) that have an identified label (Pass/ Flagged-Assigned to CS), resulting in a matrix (20947 x 40).

- Integer values: 32 (30 PANSS scores, PANSS Total, and Visit Day)
- Categorical values: 7 (Study, Country, Patient ID, Site ID, Rater ID, Assessment ID, and TxGroup)
- **Output label**: Categorical with 3 values (Passed, Flagged, Assigned to CS)

The 32 integer values could be easily fed into an algorithm to predict output label after turning it into a binary vector, however the prediction accuracy was very poor and much valuable information was being left out.

I decided to dig a little bit deeper by using as much data as I could, while also transforming the data and implementing new predictors by using interaction variables. The number of features were not my concern, as the number of examples was large, and I wasn't intending to employ polynomial feature generation, besides, the models that were producing the best results such as XGBoost did not require much feature selection. I made several changes throughout the competition by adding combinations and transformations of variables, attempting to find the most optimal predictor for the binary classification task.

Categorical variables

I started by turning the categorical values + output into numerical, or binary values:

1. **Study**: (4 new predictors) There were 5 different studies (A, B, C, D, E), so I created 4 binary columns as predictors "A, B, C, D". I took out the E from the test set, and concatenated "A, B, C, D" with zero values into the test set. The E predictor was left out to prevent multicollinearity. This step helped me realize that study D was the most influential in predicting the output, although the predictions were misleading, as D is a study that showed large irregularities in the data.
2. **Country:** (4 new predictors) After realizing that the test set only had about 43% Russia, 55%, 2% UK, I separated the countries into: USA, Russia, Europe, Asia, and Americas. I then eliminated the dummy variable for Americas. This variable also helped me realize that I need to put extra weight on the two countries being evaluated.
3. **Patient ID:** This predictor was eliminated, although it helped me to find additional predictors that account for differences within each patient. I was able to extract rate of change, mean values, and totals.
4. **Site ID:** (1 new predictor) I turned this variable into mean PANSS Total belonging to each Site ID. ("Site mean")
5. **Rater ID:** (5 new predictors) I also turned each rater into its mean PANNS Total score that it assigned. ("Rater mean"), along with 4 variables for Rater USA, Rater Russia, Rater Europe, and Rater Asia.
6. **Assessment ID:** This variable was unique for each patient, so it was taken out after preprocessing.
7. **TxGroup:** (1 predictor) Binary value with 1: Treatment, 0: Control

Overall, I turned 7 categorical variables into 15 new ordinal, and binary predictors.

Numerical variables

I created a large number of features by mixing variables together, using squares and square roots. I did this to account for interactions within variables, as an example in the reading it mentions

that some scores assigned must go hand in hand with other scores assigned, else a flag would be raised**.**

1. **Scores:** An average of 20 new variables containing differences and first score squared then the difference with another score, etc. It took me some extra reading to find interaction scores, but I also mixed scores that were most important according to GXB feature_importances feature (based on F-Score).
2. **Totals:** Created 7 new features: Total Positive, Total Negative, Bipolar index (TP – TN), TP squared – TN, TN squared – TP, Positive Chronic patients and Negative Chronic patients (those that have 3 or more P/N with 4 or more points).

Overall, I created about 30 new features accounting for interaction, along with the previous scores plus the output label which was turned into a binary vector.

The preprocessed dataset contained the following:

- Y vector, 1: Passed and 0: Flagged/Assigned to CS.
- X matrix: 17999 examples, 74 variables.
- Real – Test matrix: 1962 examples, 74 variables

Data analysis and exploration

Further data exploration showed me that study D only contained China, while the test set contained USA and Russia. Additionally, Study D contained an unexplainably large number of Flagged examples.
I took Study D out and improved my score in the leaderboard largely.
I also tried taking out the flagged examples from study A, as they were significantly large. I didn't get positive results, my False Positives increased.

Modeling Approaches

I tried several classification models, and I also did some preprocessing for those that needed it. I figured that some tree models did not need scaling, so scaling was done for a few models.

I always plugged train test split into any model prediction, and also used CV score of 10, and MSE.

- Logistic Regression with multiple variables: 86% accuracy
- LASSO Regression: around 87% accuracy
- Random Forest: ~ 88%
- XGB Classifier: ~ 88-90%

I decided to further continue with XGB because it always had the highest score in the Kaggle Leaderboard.
I did several hyperparameter tuning on XGB. The most important were tree depth, learning rate, gamma, and alpha. These hyperparameters were very sensitive to little changes in the dataset, and so I always decided to leave the parameters unmodified whenever I was looking to make modifications to the dataset.
Overall, I'm currently placing 3rd in the Kaggle Leaderboard, and my highest score is 0.61795.

Conclusions:
I truthfully enjoyed the project from beginning to end. The challenge and competition pushed me harder to try to come up with a creative idea that would give me an advantage. I spent more than 80 hours throughout this project, working mainly in implementing models, analyzing the data, and preprocessing the dataset.
The amount of information that I obtained has been incomparable, as I have been able to build algorithms by myself, and to improve the model and dataset in such amount of time. I would not change this experience, nor my time with my partner Artem throughout our process of finding solutions for this challenge.