**Using Natural Language Processing (NLP) to Predict Psychiatric Consultation in Inpatient Hospitals**

Capstone Project for Springboard Data Science Intensive Program

Robert F. Arao

**THE PROJECT:**

As the number of people with mental health problems increases, the need for time-sensitive psychiatric consultation is critical, especially where psychiatric health-care resources are low. Subsequently, my client wanted an instrument that could predict whether a patient would need psychiatric consultation. My client also wanted to know the probability of psychiatric consultation for six specific phrases chosen *a priori* for their clinical salience in the psychiatric consultation literature.

Although there are few studies in psychiatry that have developed tools to predict the need for psychiatric consultation1,2, the main aim of my project is to develop a model that predicts the need for psychiatric consultation soon after inpatient admission. This predictive model could affect the development of interventions, improve patient safety, and better manage psychiatric health-care resources in a timely manner. The second aim is to calculate the probability of psychiatric consultation based on the specific phrases chosen *a priori*.

**THE DATA:**

Data was extracted from electronic health records (EHR) from two medical centers- the University of Washington Medical Center (UWMC) and Harborview Medical Center (HMC). From their in-house SQL data warehouse (AMALGA), 108,530 patients were identified who were between the ages of 18 and 100 and admitted as inpatients between January 1, 2015 and September 20, 2017, and who were not deceased. A query of the database also indicated whether they received psychiatric consultation (cases) or not (controls).

I simulated a quasi-case-control design by using propensity score matching to identify matched controls. Propensity score matching is a matching technique used to estimate an outcome. To calculate the propensity score, logistic regression was used to estimate the probability of receiving a psychiatric consult vs. not receiving a psychiatric consult while controlling for gender, race, ethnicity, language spoken, veteran status, employment status, and admission year. I used a SAS macro3 to match 1:1 (case:control) using the logit of the propensity score with a caliper width of 0.2. Matching was performed within each medical center. While it is unknown what the actual distribution of cases and controls are in the two different hospitals, a 1:2 match proved to be too memory intensive on my computer resulting in my computer hanging up and eventually crashing. My client was satisfied with using a 1:1 match.

My dataset consisted of 6,775 medical records with medical narratives (blob content) and consisted of 2,675 cases and 2,675 controls. Most of these text-based medical notes were already collated into a single blob content. That is, on average, each patient was linked to at least one blob content comprising the complete history of medical notes. These blob contents correspond to the time of admission to either the time of discharge or time of psychiatric consultation. This were exported out from the SQL (AMALGA) database as two excel files - blob content for cases and blob content for controls.

**METHODOLOGY:**

I have treated this project as a supervised classification problem. The main outcome (a binary variable) is whether a patient received psychiatric consultation (cases) or not (control). The medical narrative was analyzed using Natural Language Processing (NLP) to identify words and phrases to predict whether a patient needs psychiatric consultation as well as compute the probability of the need for psychiatric consultation.

I used natural language processing (NLP) to analyze the blob content for each patient. This process first involves pre-text processing whereby familiar words that do not add any unique meaning were excluded (E.g. ‘ID/CC’, ‘a’,’is’,’between’). These words were added to an existing ‘stopword’ list found in the scikit-learn library. With the blob content represented as vectors and the removal of familiar words, I trained the model using the Multinomial Naïve Bayes classifier algorithm. The hyperparameters of the model was fine-tuned using K-fold cross-validation, using K=5. I then calculated the log-probability estimate for each word. This was done for two models: i) a single-word model, and ii) a tri-gram model.

After assessing the performance of the model, I identified some of the top 10 words and phrases with the highest probability of predicting the need for psychiatric consult. The probability for the *a priori* list was also calculated.

*LIBRARIES*

numpy:

array data structure, primary input for classifiers model, comparison matrix manipulation.

pandas:

data loading, data wrangling, data cleaning, dataframe manipulation, feature selection, and descriptive statistics

datetime

data wrangling and extracting month and year

matplotlib.pyplot

data visualization

seaborn

data visualization

nltk.corpus

stopwords list to remove familiar words

import string

converting blob content into string format to perform length computations.

sklearn.feature\_extraction.text

tokenization and vectorization (i.e. CountVectorization)

sklearn.naive\_bayes

multinomial Naïve Bayes classifier algorithm

sklearn.model\_selection

to split analytical dataset into train and test subset

sklearn.cross\_validation

for k-fold cross-validation

**DATA WRANGLING AND CLEANING:**

I used Python 3.4 to vertically merge the two excel files (blob content for cases and blob content controls). There were no duplicate records found. I assessed for any missingness and found two medical records with empty blob contents and were both controls. Since there were only two control medical records, these were excluded from further analysis. My analytical dataset consisted of 6,775 medical records corresponding to 2,767 patients who received psychiatric consultation and 2,767 patients who did not receive psychiatric consultation.

A few variables needed to be formatted for further analysis. The admission date was changed from an object-type to a useable datetime format, and a new variable (‘admit\_monyr’) was created to record the month and year of admission. The blob content was converted from an object-type into a string format.

After performing a preliminary review of the blob content, my client requested to exclude certain sections that had irrelevant material. This included specific drug names and their concentrations and evaluation times (reported in hours and minutes). This information was found between two key words: “Results Review” and “Problem List”. Subsequently, the use of a string wild card (\*) in between the first and last key words was used to exclude these sections. Reducing the word list improved the sparsity matrix thus allowing for faster computations. My client also identified certain phrases they were not interested in for this analysis but were interested in looking at them later. So, I flagged these phrases without excluding them by concatenating the first and last word of the phrases:

df\_combined\_working.replace('time \*evaluation','timeevaluation', regex=True)

df\_combined\_working.replace('consult \*history','consulhistory', regex=True)

df\_combined\_working.replace('time \*consult','timeconsult', regex=True)

df\_combined\_working.replace('hours \*yes','hoursyes', regex=True)

df\_combined\_working.replace('mins \*present','minspresent', regex=True)

df\_combined\_working.replace('consult \*illness','consultillness', regex=True)

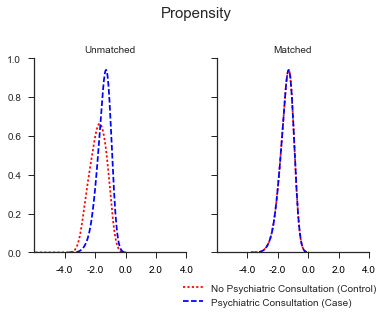
df\_combined\_working.replace('patient \*within','patientwithin', regex=True)

Unusual formatting characters was also found in the blob content (i.e. ‘\n’, ‘\t’, ‘\’). These were also removed from the dataset. The length of each blob content was then calculated for further analysis.

**EXPLORATORY ANALYSIS**

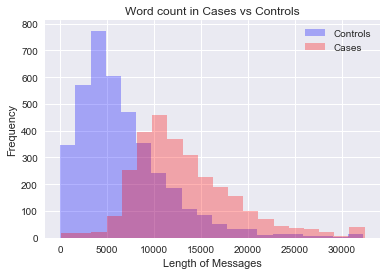
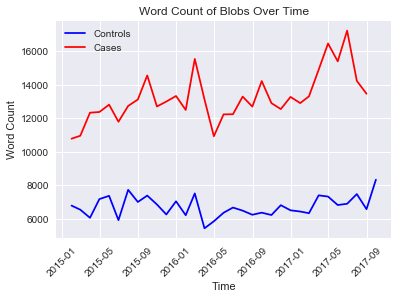
I plotted the propensity scores between the cases and controls for the matched and unmatched set to see how well the propensity score matching performed. The exactly lining of the two curves in the matched dataset reveals good matching performance (Figure 1).

Figure 1. Propensity score matching between cases and controls in unmatched and matched cohort



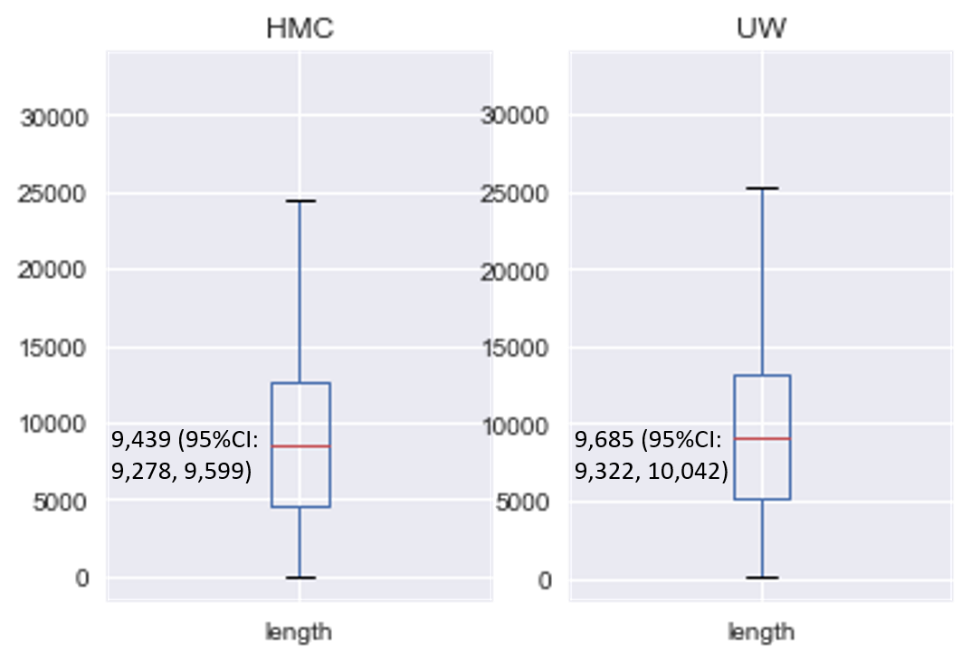
A histogram of the word count between cases and controls showed a unique area where cases uses more words in their blob contents (Figure 2). This supports the idea that there are certain words that may be better predictors of psychiatric consultation. The mean word count per message among cases (13,317 (95%CI: 13,111, 13,521)) was significantly higher than that among controls (6,679 (95%CI: 6,525, 6,832)). I also plotted a line chart to see if there were any observable time trends in word count between cases and controls. There were no observable time trends among cases nor controls (Figure 3).

Figure 2. Distribution of word count in cases and controls Figure 3. Time trend of word count among cases and controls.

Further, there did not appear any significant difference in mean word count between the two medical centers (UW: 9,682 (95%CI: 9,322,10,042) vs. HMC: 9,439 (95%CI: 9,278, 9,599) (Figure 4).

Figure 4. Word count between HMC and UW



**MODEL FITTING**

The matched dataset with blob content was randomly split such that 70% of the data was used to train the model (N=4,743), and 30% was used to test the model. (N=2,032). By splitting the matched dataset into a training/test set, the model only ever sees the training data during its model fitting and parameter tuning. That is, the final evaluation of the test data is representative of true predictive performance.

I used the bag-of-words approach to identify single words that were most predictive of receiving a psychiatric consult. The bag-of-words approach (vector space model) is a method to convert a collection of texts (corpus) into a vector format. In this approach, I disregard grammar and word order and focus mainly on multiplicity of each included word. I also developed a program to identify a contiguous sequence of words (n-grams) and to predict the probability of a user input word(s). With the blob content represented as vectors, I trained the model using the Multinomial Naïve Bayes classifier algorithm. A 5-fold cross-validation using revealed an alpha of 50 and min\_df of .1 would be suitable for my models. K=5. I then calculated the log-probability estimate for each word. This was used to develop a single-word model and a tri-gram model.

**MODEL EVALUATION:**

I used chi-square test to statistically assess the performance of propensity score matching. As table 1 indicates, there was no statistically significant differences in patient characters after performing propensity score matching.

Table 1. Comparison of patient characteristics before and after matching

|  | *Unmatched* | | |  | *Matched (1:1)* | | |
| --- | --- | --- | --- | --- | --- | --- | --- |
| *Characteristics* | *controls* | *cases* | *p-value* |  | *controls* | *cases* | *p-value* |
| Total | 105,763 | 2,767 | < 0.0001 |  | 2,767 | 2,767 | 1.0000 |
| UWMC | 46,451 | 451 |  |  | 451 | 451 |  |
| HMC | 59,312 | 2,316 |  |  | 2,316 | 2,316 |  |
|  |  |  |  |  |  |  |  |
| Gender |  |  |  |  |  |  |  |
| Female | 50,303 (47.6) | 1,090 (39.4) | < 0.0001 |  | 1,097 (39.6) | 1,090 (39.4) | 0.8474 |
| Male | 55,448 (52.4) | 1,677 (60.6) |  |  | 1,670 (60.4) | 1,677 (60.6) |  |
| Unknown | 12 (0.0) | 0 (0.0) |  |  | 0 (0.0) | 0 (0.0) |  |
|  |  |  |  |  |  |  |  |
| Race |  |  |  |  |  |  |  |
| White | 74,064 (70.0) | 2,006 (72.5) | < 0.0001 |  | 2,037 (73.6) | 2,006 (72.5) | 0.5360 |
| Black or African American | 13,098 (12.4) | 367 (13.3) |  |  | 372 (13.4) | 367 (13.3) |  |
| Asian | 8,441 (8.0) | 155 (5.6) |  |  | 158 (5.7) | 155 (5.6) |  |
| Native Hawaiian or Other | 1,121 (1.1) | 24 (0.9) |  |  | 14 (0.5) | 24 (0.9) |  |
| America Indian or Alaskan | 2,698 (2.6) | 87 (3.1) |  |  | 76 (2.7) | 87 (3.1) |  |
| Multiple races | 1,124 (1.1) | 63 (2.3) |  |  | 53 (1.9) | 63 (2.3) |  |
| Unknown | 5,217 (4.9) | 65 (2.3) |  |  | 57 (2.1) | 65 (2.3) |  |
|  |  |  |  |  |  |  |  |
| Primary Language Spoken |  |  |  |  |  |  |  |
| English | 94,469 (89.3) | 2,605 (94.1) | < 0.0001 |  | 2,609 (94.3) | 2,605 (94.1) | 0.6074 |
| Spanish | 4,306 (4.1) | 50 (1.8) |  |  | 54 (2.0) | 50 (1.8) |  |
| Other | 6,639 (6.3) | 105 (3.8) |  |  | 101 (3.7) | 105 (3.8) |  |
| Unknown | 349 (0.3) | 7 (0.3) |  |  | 3 (0.1) | 7 (0.3) |  |
|  |  |  |  |  |  |  |  |
| Veteran Status |  |  |  |  |  |  |  |
| Yes | 8,142 (7.7) | 230 (8.3) | < 0.0001 |  | 216 (7.8) | 230 (8.3) | 0.5358 |
| No | 90,263 (85.3) | 2,402 (86.8) |  |  | 2,429 (87.8) | 2,402 (86.8) |  |
| Unknown | 7,358 (7.0) | 135 (4.9) |  |  | 122 (4.4) | 135 (4.9) |  |
|  |  |  |  |  |  |  |  |
| Employment status |  |  |  |  |  |  |  |
| Full-time | 30,508 (28.8) | 292 (10.6) | < 0.0001 |  | 297 (10.7) | 292 (10.6) | 0.9854 |
| Part-time | 5,026 (4.8) | 78 (2.8) |  |  | 78 (2.8) | 78 (2.8) |  |
| Self-Employee | 3,720 (3.5) | 50 (1.8) |  |  | 49 (1.8) | 50 (1.8) |  |
| Not Employed | 41,143 (38.9) | 1,862 (67.3) |  |  | 1,854 (67.0) | 1,862 (67.3) |  |
| Student | 3,120 (2.9) | 36 (1.3) |  |  | 39 (1.4) | 36 (1.3) |  |
| Active Duty | 79 (0.1) | 1 (0.0) |  |  | 0 (0.0) | 1 (0.0) |  |
| Retired | 18,708 (17.7) | 362 (13.1) |  |  | 369 (13.3) | 362 (13.1) |  |
| Unknown | 3,459 (3.3) | 86 (3.1) |  |  | 81 (2.9) | 86 (3.1) |  |
|  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |
| Year of Admission |  |  |  |  |  |  |  |
| 2015 | 38,513 (36.4) | 1,028 (37.2) | 0.0020 |  | 1,045 (37.8) | 1,028 (37.2) | 0.8492 |
| 2016 | 39,898 (37.7) | 960 (34.7) |  |  | 960 (34.7) | 960 (34.7) |  |
| 2017 | 27,352 (25.9) | 779 (28.2) |  |  | 762 (27.5) | 779 (28.2) |  |
|  |  |  |  |  |  |  |  |
| Age, mean (std) | 49.2 (18.0) | 47.4 (16.4) | < 0.0001 |  | 47.6 (16.5) | 47.4 (16.4) | 0.7095 |

After fitting the model to the training (Table 2a) and test dataset (Table 2b) in the single-word model, the overall accuracy of the training-fitted model was slightly higher than the test-fitted model (88.5% vs.87.6%). The suggests the model was not overfitted. In identifying single words, the overall precision was found to be 90% with a recall rate of 88%. The F1-score was also fairly high (88%) and approximates the recall rate.

Table 2a. Classification report of single-word model prediction on training data

precision recall f1-score support

0 1.00 0.77 0.87 2754

1 0.76 1.00 0.86 1988

avg / total 0.90 0.87 0.87 4742

Accuracy=0.885280

Table 2b. Classification report of single-word model prediction on test data

precision recall f1-score support

0 0.99 0.80 0.88 1207

1 0.77 0.99 0.87 826

avg / total 0.90 0.88 0.88 2033

Accuracy=0.876045

In the tri-gram model, the overall accuracy of the training-fitted model (Table 3a) was .4% points lower when fitted to the test dataset (Table 3b) (90.2% vs. 90.6%). This difference of is not substantial enough to suggest any overfitting. In identifying tri-grams in the test dataset, the overall precision, recall rate, and F1-score were each found to be 91%. Moreover, the tri-gram model seemed to better perform than the single-word model.

Table 3a. Classification report of tri-gram model prediction on training data

precision recall f1-score support

0 0.99 0.84 0.91 2740

1 0.82 0.99 0.90 2002

avg / total 0.92 0.90 0.90 4742

Accuracy=0.902362

Table 3b. Classification report of tri-gram model prediction on test data

precision recall f1-score support

0 0.99 0.84 0.91 1173

1 0.82 0.99 0.90 860

avg / total 0.91 0.91 0.91 2033

Accuracy=0.90558

**RESULTS:**

Based on the single-word model, some of the top words that predicted psychiatric consultation include ’delirium’ (P=0.97), ‘restraints’ (P=0.95), ‘agitation’ (P=0.88), and ‘ama’ (P=0.86) (Table 4).

Table 4. Top 10 words that predict psychiatric consultations

UNIQUE WORDS P(PsychConsult | word)

delirium 0.97

restraints 0.95

judgement 0.89

agitation 0.88

collateral 0.87

ama 0.86

stressors 0.86

maintain 0.85

involuntary 0.83

attempt 0.83

”ama”: against medical advice

Based on the tri-gram model, some of the top phrases that predicted psychiatric consultation include ‘prn past psychiatric’ (Prob=1.00), ‘difficulty obtaining history’ (P=0.99), ’emergency department ita’ (P=0.99), and ‘nka medications scheduled’ (P=) (Table 5).

Table 5. Top 10 bi-grams that predict psychiatric consultations

UNIQUE PHRASES P(PsychConsult | word)

consult request evaluation 1.00

prn past psychiatric 1.00

ita involuntary treatment 1.00

examination appearance 1.00

difficulty obtaining history 0.99

emergency department ita 0.99

nka medications scheduled 0.99

evaluation consult requested 0.99

department ita involuntary 0.99

physical examination appearance 0.99

”prn*”: pro re nata* (when necessary)

”ita”: involuntary treatment act

”nka”: no known allergies

Table 6 presents specific phrases that my client chose *a priori* for their clinical salience in the psychiatric consultation literature. The phrases with the highest probability of predicting a psychiatric consultation include ‘suicide attempt’(P=0.99), ‘Substance use disorder’ (P=0.99), and ‘Parents separated’ (0.64), and ‘Family mental illness’(P=0.63).

Table 6. Probability of psychiatric consult using user-input phrases

|  |  |
| --- | --- |
| Phrases | Probability |
| Suicide attempt | 0.99 |
| Substance use disorder | 0.99 |
| Parents separated | 0.64 |
| Family mental illness | 0.63 |
| Sexual abuse | 0.56 |
| Verbal abuse | 0.56 |
| Firearm related | 0.55 |
| Overdose related | 0.55 |
| Out of control behavior | 0.50 |
| Physical abuse | 0.45 |
| Witness to | 0.41 |
| Loss of consciousness | 0.33 |

The context of these phrases is still being investigated. I’m currently working on scheduling a meeting with several psychiatrists and psychologists to go over these phrases and to better understand the context of these phrases.

**LIMITATIONS:**

Although, this study has successfully developed a model that identifies unique words and phrases that predict psychiatric consultation, this study is not without limitations. First, I did a 1:1 case-control match. This was done primarily due to the heavy burden in computer processing inherent in NLP computations. The true distribution of patients that see psychiatric consultation at inpatient hospitals is unknown. Moreover, this distribution may hard to approximate since each hospital varies in patient culture and in the number of available psychiatric wards. Second, the medical narrative is subjective and based on the user entry. That is, the medical staff that enters their text-based notes may have their own abbreviations and linguistic nuances, which may subsequently vary between medical staff. Third, EHR's from two medical hospitals was analyzed. Even though I matched within each of the hospital medical centers, the variance in note taking may vary from hospital to hospital.

**FUTURE RESEARCH:**

Future research includes improving the model even more. This would include implementing another classifier algorithms (e.g Random Forrest Classifier) to see if the same words are predicted. If a different algorithm produced similar words, this would lend to the generalizability of this study’s results. Another area of improvement would be investigating other variables that may be important in calculating the propensity score. For example, we did not control for the different medical wards.

After assessing the generalizability of these results, it would be interesting to create a stand-alone program that could be run overnight to produce a list of patients which contains unique phrases that predict the need for psychiatric consultation. The tending psychiatrist can then use this list to assign med students during their medical rounds.

**CLIENT RECOMMENDATIONS:**

Based on the above results, the tri-gram model seemed to better perform than the single-word model. Therefore, it is recommended that the tri-gram model be used to identify unique phrases.

However, it is further recommended that the context of which these phrases appear be explored to fully understand the relevance of these phrases. This is especially important where nuances and abbreviations are used.

**REFERENCES:**

1. Perlis RH. A clinical risk stratification tool for predicting treatment resistance in major depressive disorder. Biol Psychiatry 2013; 74: 7–14.

2. Cook BL, Progovac AM, Chen P et al. Novel use of natural language processing (NLP) to predict suicidal ideation and psychiatric symptoms in a text-based mental health intervention in Madrid. Comput Math Methods Med 2016; 2016: 8708434.

3. Coca-Perraillon M. Local and global optimal propensity score matching (Paper 185-2007). *SAS Global Forum 2007*. Cary, NC.: SAS Institute Inc.; 2007.