Dravet Prediction Model

Chong Kim

March 7, 2017

### Objective

Find algorithm/equation to best differentiate severe (dravet and infantile spasm) from mild state.

#### Steps

1. Data Management
2. Model Building
3. Evaluation of Model
4. Conslusion/Summary

#### Step 1: Data Management

***1a : Merge Chronic Comorbidity data with Patient Record Data***

***1b : Merge Medication data with Patient Record Data***

## 'data.frame': 191 obs. of 63 variables:  
## $ UNIQUE\_ID : int 1 2 3 4 5 6 7 8 9 10 ...  
## $ freq.DiazePAM : num 1 1 2 0 3 0 0 13 2 2 ...  
## $ freq.CloBAZam : num 0 5 16 19 7 9 3 7 5 12 ...  
## $ freq.Divalproex Sodium : num 0 5 0 0 4 0 1 0 0 0 ...  
## $ freq.LevETIRAcetam : num 0 0 16 0 0 0 0 4 0 0 ...  
## $ freq.Midazolam HCl : num 0 0 1 4 2 1 0 0 1 0 ...  
## $ freq.Sodium Chloride : num 0 0 4 12 0 0 0 1 0 0 ...  
## $ freq.ClonazePAM : num 0 0 0 8 0 0 3 0 2 0 ...  
## $ COHORT : int 0 0 0 0 0 0 0 0 0 0 ...  
## $ neuromusc\_ccc : int 0 1 1 1 1 1 1 1 1 1 ...  
## $ cvd\_ccc : int 0 1 0 0 0 0 1 1 0 0 ...  
## $ respiratory\_ccc : int 0 1 0 1 0 0 0 1 0 1 ...  
## $ renal\_ccc : int 0 0 0 0 0 0 0 0 0 0 ...  
## $ GI\_ccc : int 0 1 0 1 1 0 0 1 1 0 ...  
## $ hemato\_immu\_ccc : int 0 0 0 0 0 0 0 0 0 0 ...  
## $ metabolic\_ccc : int 0 0 0 1 0 0 0 1 0 0 ...  
## $ congeni\_genetic\_ccc : int 0 1 1 1 1 0 1 0 0 1 ...  
## $ malignancy\_ccc : int 0 0 0 0 0 0 0 0 1 0 ...  
## $ neonatal\_ccc : int 0 0 0 0 0 0 0 0 0 0 ...  
## $ tech\_dep\_ccc : int 0 1 0 1 1 0 1 1 0 0 ...  
## $ transplant\_ccc : int 0 0 0 0 0 0 0 0 0 0 ...  
## $ num\_ccc : int 0 5 2 5 3 1 3 5 3 3 ...  
## $ ccc\_flag : int 0 1 1 1 1 1 1 1 1 1 ...  
## $ COHORT\_ID : int 0 0 0 0 0 0 0 0 0 0 ...  
## $ MATCHED\_TRIAD : int 48 73 20 38 75 37 41 70 16 50 ...  
## $ SEX\_NUM : int 0 0 1 1 1 0 0 0 0 0 ...  
## $ AGE\_YEARS : int 7 7 3 3 11 3 3 5 4 2 ...  
## $ INDEX\_DATE\_SHIFT : Factor w/ 188 levels "1-Feb-15","1-Jul-15",..: 80 140 9 76 33 130 73 54 100 131 ...  
## $ ED\_COUNT : int 0 0 0 0 0 0 0 1 1 0 ...  
## $ ED\_MEDIAN\_MINS : num 0 0 0 0 0 0 0 135 193 0 ...  
## $ ED\_MEDIAN\_HOURS : num 0 0 0 0 0 0 0 2.25 3.22 0 ...  
## $ ED\_MEDIAN\_DAYS : num 0 0 0 0 0 0 0 0.09 0.13 0 ...  
## $ UC\_COUNT : int 0 0 0 0 0 0 0 0 0 0 ...  
## $ UC\_MEDIAN\_MINS : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ UC\_MEDIAN\_HOURS : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ UC\_MEDIAN\_DAYS : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ ICU\_COUNT : int 0 0 1 1 0 0 0 0 0 0 ...  
## $ ICU\_MEDIAN\_MINS : num 0 0 13753 21925 0 ...  
## $ ICU\_MEDIAN\_HOURS : num 0 0 229 365 0 ...  
## $ ICU\_MEDIAN\_DAYS : num 0 0 9.55 15.23 0 ...  
## $ IP\_COUNT : int 0 1 0 0 1 1 0 4 2 3 ...  
## $ IP\_MEDIAN\_MINS : num 0 882 0 0 2156 ...  
## $ IP\_MEDIAN\_HOURS : num 0 14.7 0 0 35.9 ...  
## $ IP\_MEDIAN\_DAYS : num 0 0.61 0 0 1.5 1.17 0 1.97 0.49 1.08 ...  
## $ ANNUAL\_AIR : int 0 0 0 0 0 0 0 0 0 0 ...  
## $ ANNUAL\_GROUND : int 0 0 0 0 0 0 0 0 0 0 ...  
## $ ANNUAL\_WALK : int 0 1 0 0 1 1 0 4 3 3 ...  
## $ PROCEDURE\_COUNT : int 0 5 7 14 1 2 0 9 0 3 ...  
## $ OTHER\_OUTPATIENT\_COUNT : int 1 11 12 14 6 6 4 22 10 9 ...  
## $ BEHAVIORAL\_COUNT : int 0 0 0 0 0 2 0 31 0 0 ...  
## $ TOTAL\_OUTPATIENT\_COUNT : int 1 11 12 14 6 8 4 53 10 9 ...  
## $ LAB\_COUNT : int 0 4 23 25 9 0 8 14 6 20 ...  
## $ MED\_GIVEN\_OR\_PRESCRIBED: int 1 1 1 1 1 1 1 1 1 1 ...  
## $ PRESCRIPTION\_COUNT : int 1 20 144 256 23 13 10 121 25 81 ...  
## $ DISTANCE : int 3 3 2 3 3 2 2 2 0 3 ...  
## $ INSURANCE : int 1 1 1 0 0 0 1 1 0 0 ...  
## $ TOTAL\_CHARGES : num 140 14317 426846 2361197 24002 ...  
## $ TOTAL\_COST : num 46.3 4724.7 140859.3 779195 7920.6 ...  
## $ TOTAL\_OUTPATIENT\_CHARGE: num 140 14213 321422 2164491 3797 ...  
## $ TOTAL\_OUTPATIENT\_COST : num 46.3 4690.3 106069.4 714282 1253.1 ...  
## $ TOTAL\_INPATIENT\_CHARGE : num 0 104 105424 196706 20205 ...  
## $ TOTAL\_INPATIENT\_COST : num 0 34.3 34789.9 64913 6667.6 ...  
## $ X : logi NA NA NA NA NA NA ...

* The table is in a 'one-patient-one-record' form. The top 6 observations above indicate the structure of the data that will be used for the prediction modelling.

***1c : Discard unnecessary variables in prediction and make dummy variables for those that are necessary and non numeric***

***1d : Missing data check***

|  |  |
| --- | --- |
| Data | Values |
| observations | 189 |
| variables | 58 |

* 1 person with missing insurance information excluded

***1e : Combine the severe epilepsy group***

***1f : Remove variables that produce Near Zero or Zero Variance***

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | freqRatio | percentUnique | zeroVar | nzv |
| UC\_COUNT | 25.71429 | 1.5873016 | FALSE | TRUE |
| UC\_MEDIAN\_MINS | 90.00000 | 4.7619048 | FALSE | TRUE |
| UC\_MEDIAN\_HOURS | 90.00000 | 4.7619048 | FALSE | TRUE |
| UC\_MEDIAN\_DAYS | 90.00000 | 3.7037037 | FALSE | TRUE |
| ANNUAL\_AIR | 46.25000 | 1.0582011 | FALSE | TRUE |
| ANNUAL\_GROUND | 187.00000 | 1.5873016 | FALSE | TRUE |
| MED\_GIVEN\_OR\_PRESCRIBED | 0.00000 | 0.5291005 | TRUE | TRUE |
| INSURANCE.2 | 188.00000 | 1.0582011 | FALSE | TRUE |
| INSURANCE.3 | 36.80000 | 1.0582011 | FALSE | TRUE |
| Data | Values |
| observations | 189 |
| variables | 51 |

* There are now 51 variables in the dataset.

***1g***: Split data into training and test.

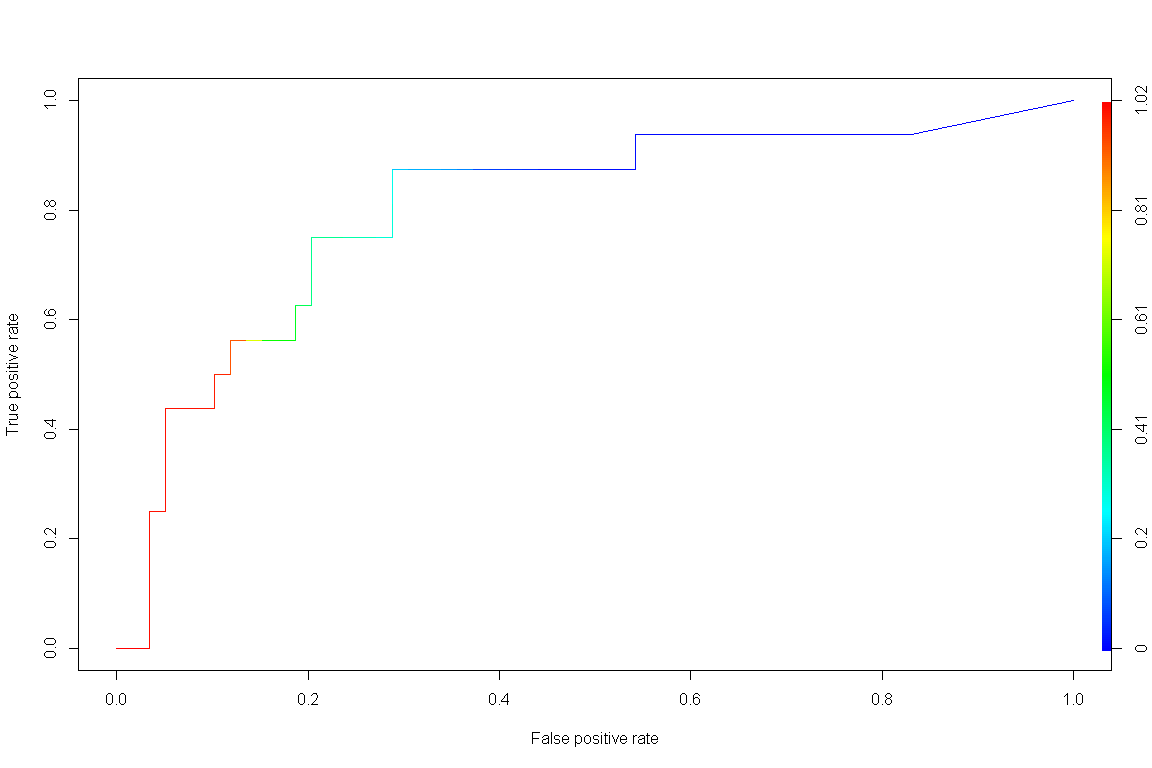
* Here we split the data into 60/40 (Training/Testing).

##### Logistic Regression Modedl 2

***2A: GLM model without model selection***

|  |  |  |
| --- | --- | --- |
|  | Severe | Non Severe |
| Severe | 49 | 7 |
| Non Severe | 10 | 9 |

* Given a 50% cutoff for predicting Severe vs. Non Severe cases, the ***sensitivity*** of the logistic regression model is 0.8305085.
* Given a 50% cutoff for predicting Severe vs. Non Severe cases, the ***specificity*** of the logistic regression model is 0.5625.
* Given a 50% cutoff for predicting Severe vs. Non Severe cases, the ***positive predictive value*** of the logistic regression model is 0.875.
* Given a 50% cutoff for predicting Severe vs. Non Severe cases, the ***negative predictive value*** of the logistic regression model is 0.4736842.
* The ***accuracy*** of the model in terms of correctly classifying the outcome is 0.7733333 and the 2x2 contingency table indicates that there are only out of the misclassified.

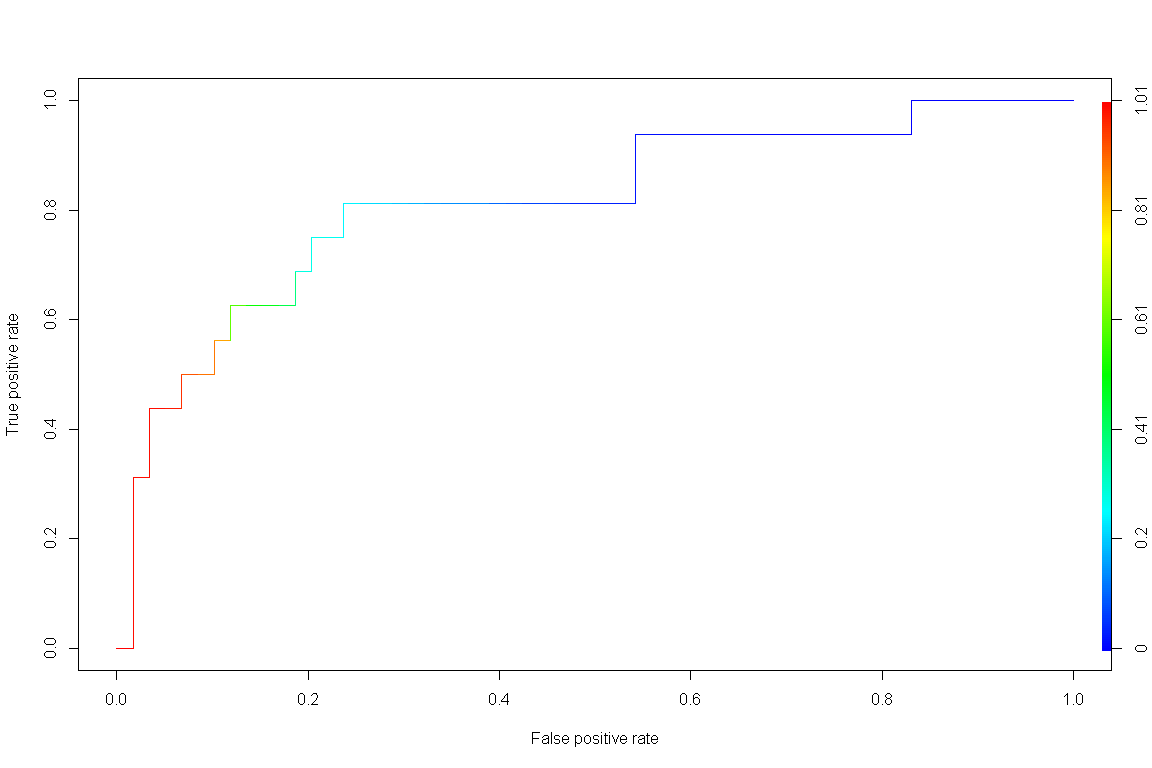


## [1] "AUC :" "0.804025423728814"

* The Area Under the Curve (AUC) of the ROC curve is 0.8040254, which indicates that the probability of the model correctly identifying Severe vs. Non severe is 0.8040254.

***2B: Evaluation of GLM model with Stepwise Variable Selection***:

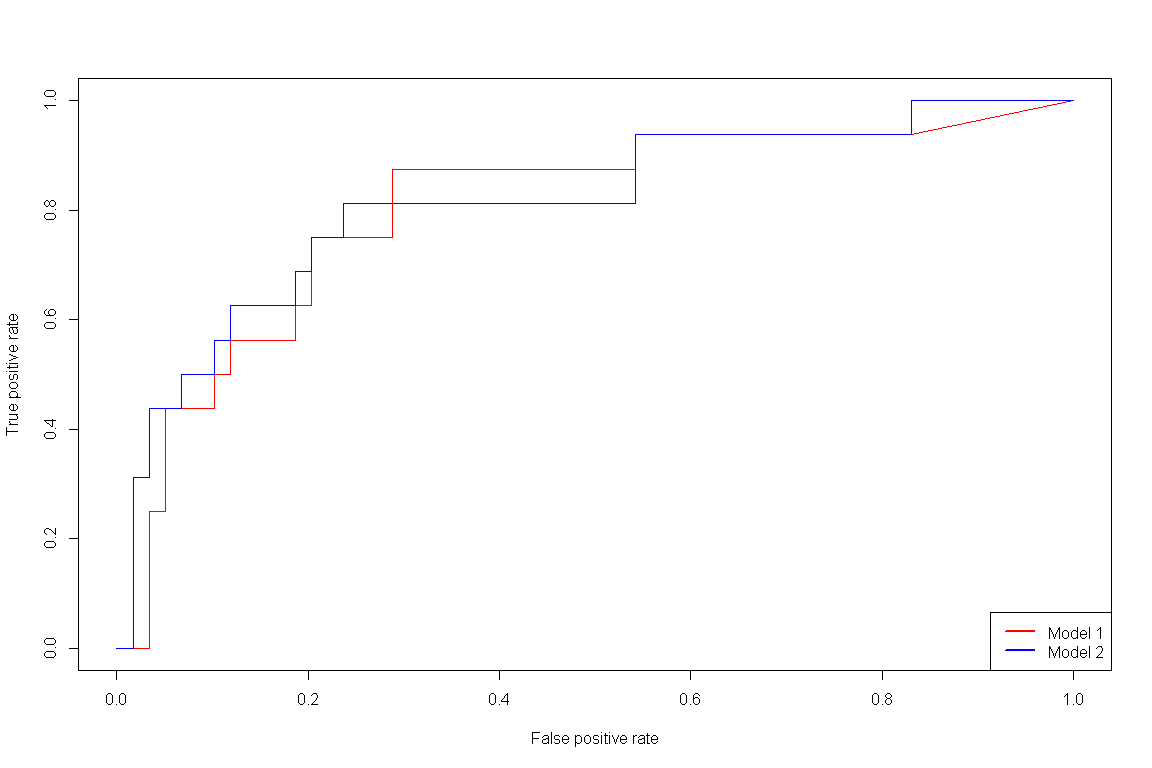
|  |  |  |
| --- | --- | --- |
|  | Severe | Non Severe |
| Severe | 51 | 6 |
| Non Severe | 8 | 10 |



## [1] "AUC :" "0.804025423728814"

* Using the new model created with stepwiseAIC variable selection, given a 50% cutoff for predicting Severe vs. Non Severe cases, the ***sensitivity*** of the logistic regression model is 0.8644068.
* Given a 50% cutoff for predicting Severe vs. Non Severe cases, the ***specificity*** of the logistic regression model is 0.625.
* Given a 50% cutoff for predicting Severe vs. Non Severe cases, the ***positive predictive value*** of the logistic regression model is 0.8947368.
* Given a 50% cutoff for predicting Severe vs. Non Severe cases, the ***negative predictive value*** of the logistic regression model is 0.5555556.
* The ***accuracy*** of the model in terms of correctly classifying the outcome is 0.8133333 and the 2x2 contingency table indicates that there are only out of the misclassified.
* The new model includes variables COHORT3, freq.LevETIRAcetam, cvd\_ccc.0, neuromusc\_ccc.0, TOTAL\_OUTPATIENT\_COUNT, congeni\_genetic\_ccc.0, num\_ccc, freq.CloBAZam, ICU\_COUNT, PROCEDURE\_COUNT which are frequency of Levetiracetam prescribed, Count of # of cardio-vascular chronic comorbid conditions, Count of # of neuro-muscuclar chronic comorbid conditions, Count of total oupatient claims, Count of # of congenial genetic chronic comorbid conditions, Total # of chronic comorbid conditions, frequency of Clobazam prescribed, Count of # of ICU claims, and Count of # of procedure claims.

***2C: Visual Comparison of Model***



* Based on the AUC statistics, the new model yields a slightly higher ***AUC*** of 0.8135593 vs. 0.8040254.

***2D: Regression parameters***

|  |  |
| --- | --- |
|  | **Model 2** |
| (Intercept) | 1.03 (2.12) |
| freq.LevETIRAcetam | 0.37 (0.20) |
| cvd\_ccc.0 | -2.69 (1.72) |
| neuromusc\_ccc.0 | 20.96 (6117.56) |
| TOTAL\_OUTPATIENT\_COUNT | -0.43 (0.15)\*\* |
| congeni\_genetic\_ccc.0 | 1.84 (1.32) |
| num\_ccc | -0.67 (0.45) |
| freq.CloBAZam | -0.22 (0.15) |
| ICU\_COUNT | -28.88 (2084.53) |
| PROCEDURE\_COUNT | 1.46 (0.43)\*\*\* |
| AIC | 75.40 |
| BIC | 102.76 |
| Log Likelihood | -27.70 |
| Deviance | 55.40 |
| Num. obs. | 114 |
| ***\*\*\*p < 0.001, \*\*****p < 0.01, \**p < 0.05 | |

* Based on the regression parameters estimated, the count of # of procedure claims and counts # of outpatient claims had a statistically significant association with the severity of Dravet (p<0.01 and p<0.001, respectively).
* The odds of being classified as Non severer epileptic (i.e. Childhood Absence Epilepsy) decreases by a factor of 0.6503081 for a one unit increase in the # of outpatient claims.
* The odds of being classified as Non severer epileptic (i.e. Childhood Absence Epilepsy) increases by a factor of 4.32 for a one unit increase in the # of procedure claims.

***Probability of classified as Non Severe Epilepsy: adjusting Procedure claims***

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **LevETIRAcetam** | **CVD CCC** | **Neuro musc CCC** | **TOTAL**  **OUTPAT**  **CNT** | **Congeni genetic CCC** | **Total CCC** | **CloBAZam** | **ICU CNT** | **PROC CNT** | **Prb. NS** |
| 2.28 | 0.66 | 0.10 | 12.69 | 0.65 | 2.73 | 2.77 | 0.186 | 2 | 0.001 |
| 2.28 | 0.66 | 0.10 | 12.69 | 0.65 | 2.73 | 2.77 | 0.186 | 4 | 0.019 |
| 2.28 | 0.66 | 0.10 | 12.69 | 0.65 | 2.73 | 2.77 | 0.186 | 6 | 0.267 |
| 2.28 | 0.66 | 0.10 | 12.69 | 0.65 | 2.73 | 2.77 | 0.186 | 8 | 0.872 |

* After setting all variables to their mean values except for # of procedure claims, adjusting the values of the # of procedure claims indicate that the probability of being classified as non severe epilepsy is 0.001, 0.019, 0.267, 0.872 as we adjust the # of procedure claims to 2,4,6,8, respectively.

***Probability of classified as Non Severe Epilepsy: Adjusting TOTAL\_OUTPATIENT\_COUNT***

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **LevETIRAcetam** | **CVD** | **Neuro musc** | **TOTAL**  **OUTPAT**  **CNT** | **Congeni genetic** | **Total CCC** | **CloBAZam** | **ICU CNT** | **PROC CNT** | **Prb. NS** |
| 2.28 | 0.66 | 0.10 | 2 | 0.65 | 2.73 | 2.77 | 0.186 | 3.82 | 0.601 |
| 2.28 | 0.66 | 0.10 | 4 | 0.65 | 2.73 | 2.77 | 0.186 | 3.82 | 0.389 |
| 2.28 | 0.66 | 0.10 | 6 | 0.65 | 2.73 | 2.77 | 0.186 | 3.82 | 0.212 |
| 2.28 | 0.66 | 0.10 | 8 | 0.65 | 2.73 | 2.77 | 0.186 | 3.82 | 0.102 |
| 2.28 | 0.66 | 0.10 | 10 | 0.65 | 2.73 | 2.77 | 0.186 | 3.82 | 0.0461 |

* After setting all variables to their mean values except for # of total outpatient claims, adjusting the values of the # of total outpatient claims indicate that the probability of being classified as non severe epilepsy is 0.602, 0.39, 0.213, 0.103,0.046 as we adjust the # of procedure claims to 2, 4, 6, 8 and 10, respectively.