J&J CAPSTONE PROJECT

AutoML Prediction Machine of Adverse Outcomes Following Hip Fracture Surgery

Mentors: Cindy Tong, Rhea Jang, Chin-Wen Chang, David Knowles
Students: Mike Wang, Zining Fan, Qiang Zhao, Siyuan Wang
Project Introduction

Objective:
Build a reusable, flexible platform for rapid prediction at the hospital level on hip fracture patients.

Why?
- Hip fractures are a major public health burden
- J&J works with hospitals to reduce complications after hip fracture surgeries

- Data source: Medicare + other public data (census) - 400K+ patient records.
- Outcome: 90-Day Risk of Readmission
Data Exploratory Analysis

- Over 430,000 observations, 247 features
- Feature types
  - hospital information
  - personal information
  - medical record
  - diagnosis and procedure of surgery
  - follow-up information
Data Exploratory Analysis

- Over 3,000 hospitals
- Imbalanced classes for target variable: 23% positive
Data Preprocessing

- Removed features with leaking info, with collinearity
- Added time related features: month, day of week…
- 137 features left
Categorical Features Processing

- Regrouping: group by type, keep the top classes and drop repeating information
- An encoding approach that combines both encoders: One-Hot encoder with classes $\leq 5$, target encoder with classes $> 5$
Approaches

- Develop model that can accurately predict individual patient’s risk of readmission.
  - Population Model
  - Hospital Model
  - Ensemble Method
- Then use the best approach from above to generate readmission rate for each hospital.
Population Model

- **Goals:** build one model for all hospitals
- **Sampling Method**
  - We only keep hospital which has >10 positive/negative observations to exclude extreme cases, which also meet the Medicare data policy
  - Remove hospital which has less than 100 observations
- **Train/Test Split by chronological order**
  - Train/validation/Test Ratio: 0.7: 0.15: 0.15
Population Model

- Three model types: regularized logistic regression, random forest, Xgboost
- Four feature sets: Base, Base with time features, Boruta features, Boruta with time features
- Metrics (F-1 score):
  - F-1 score can help balance metric when there is an imbalance dataset
  - F-1 score summarise Recall, Precision, True Positive, False Positive, False Negatives into one
Population Model

- Random Forest with base + time feature group as our final population model (train with 100% data)

<table>
<thead>
<tr>
<th>Trail</th>
<th>Model</th>
<th>Feature set</th>
<th>Cross Validation F1</th>
<th>Validation F1 Score</th>
<th>Log Loss</th>
<th>AUC</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Random Forest</td>
<td>Base + Time</td>
<td>0.6073</td>
<td>0.3670</td>
<td>0.8200</td>
<td>0.5300</td>
<td>0.3200</td>
</tr>
<tr>
<td>2</td>
<td>Random Forest</td>
<td>Base + Time</td>
<td>0.6085</td>
<td>0.4066</td>
<td>0.6850</td>
<td>0.6379</td>
<td>0.5544</td>
</tr>
<tr>
<td>3</td>
<td>Random Forest</td>
<td>Base + Time</td>
<td>0.6058</td>
<td>0.3714</td>
<td>0.7868</td>
<td>0.5620</td>
<td>0.3585</td>
</tr>
<tr>
<td>4</td>
<td>Random Forest</td>
<td>Base + Time</td>
<td>0.6062</td>
<td>0.4075</td>
<td>0.6798</td>
<td>0.6398</td>
<td>0.5568</td>
</tr>
<tr>
<td>5</td>
<td>Random Forest</td>
<td>Base + Time</td>
<td>0.6063</td>
<td>0.4068</td>
<td>0.6879</td>
<td>0.6377</td>
<td>0.5537</td>
</tr>
<tr>
<td>6</td>
<td>Random Forest</td>
<td>Base + Time</td>
<td>0.6058</td>
<td>0.4038</td>
<td>0.6907</td>
<td>0.6339</td>
<td>0.5459</td>
</tr>
</tbody>
</table>
Hospital Level Models

Goal: Build different models for each hospital.

- **Data Preprocessing**
  - Grouped by hospitals, 3378 hospitals in total
  - Removed features with 0 variance within groups, 80 features left

- **Sampling and Train Test Split by chronological order**
  - Removed hospitals with less than 66 positive/negative cases, 367 hospitals left
  - 70% train, 15% validation and 15% test
  - Upsampled train dataset for each hospital
Hospital Level Models

- **Model Process**
  - 4 types of feature sets and 3 types of models, 12 combinations in total
  - Selected the best model from 12 models based on the F1 score on validation set

- **Features sets**
  - Boruta features
  - Boruta with time features
  - selectKbest features
  - Top 9 features from last year

- **Model types**
  - Regularized Logistic Regression
  - Random Forest
  - XGBoost
Hospital Level Models - Results

Validation F1-Score of 367 Hospitals for Hospital model

Best Type of Hospital Model

- feature_group
  - 2019_top9
  - Boruta
  - Boruta_time
  - Kbest

- Count
  - Random Forest
  - Reg. Logistic Regression Model
  - XGBoost
Ensemble Model Result

- Alpha * Population Model + (1 - Alpha) * Hospital Model
- Selected best alpha based on validation F1 Scores
Comparison - Validation Set

Validation F1-Score for each Hospital

Provider index (ordered by validation F1 score)
### Accumulative Validation and Test Score

- Gather all 367 hospitals’ validation sets into one big validation sets, same thing for the test set
- Random Guessing: 0 or 1 based on uniform distribution

<table>
<thead>
<tr>
<th>Model</th>
<th>Val F1</th>
<th>Log Loss</th>
<th>AUC</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Population Model</td>
<td>0.4196</td>
<td>0.6822</td>
<td>0.6404</td>
<td>0.5623</td>
</tr>
<tr>
<td>Hospital Model</td>
<td>0.4211</td>
<td>0.8201</td>
<td>0.6074</td>
<td>0.6442</td>
</tr>
<tr>
<td>Ensemble Model</td>
<td>0.5007</td>
<td>0.6497</td>
<td>0.6603</td>
<td>0.6900</td>
</tr>
<tr>
<td>Random Guessing</td>
<td>0.3210</td>
<td>0.9915</td>
<td>0.5504</td>
<td>0.5037</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Model</th>
<th>Test F1</th>
<th>Log Loss</th>
<th>AUC</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Population Model</td>
<td>0.4179</td>
<td>0.6891</td>
<td>0.6326</td>
<td>0.5543</td>
</tr>
<tr>
<td>Hospital Model</td>
<td>0.3331</td>
<td>0.8905</td>
<td>0.5479</td>
<td>0.5894</td>
</tr>
<tr>
<td>Ensemble Model</td>
<td>0.3705</td>
<td>0.6784</td>
<td>0.5945</td>
<td>0.6079</td>
</tr>
<tr>
<td>Random Guessing</td>
<td>0.3307</td>
<td>0.9892</td>
<td>0.5073</td>
<td>0.5760</td>
</tr>
</tbody>
</table>
Accumulative F1 Scores

- Hospital model and ensemble model results are biased and not generative
- Does population model work for predicting the hospital’s future readmission rate?
Population Model (Classification) Logic

Data:
- Last Year Patient Info
- Prediction Horizon Patient Info

Model:
- Training Input
- Validation Input
- Test Input

Current

Exact Copy

Prediction:
- Prediction Input

- Then use the prediction of next year patients’ Readm_flag to calculate readmission rate for corresponding hospital
- Exactly Same, thus even if our model can 100% predict Readm_flag correctly, we are only doing as good as just using last year’s readmission rate
Other Approach

Regression on Hospital Level Data

Data: Patient Level to Hospital Level by year

Hospital Last Year Avg Info | Last Year Readm Rate

Model: Linear Regression Ridge, Lasso, Elastic, Random Forest, XGBoost

Train Input | Validation Input | Test Input

Prediction:
New Hospital with Historical Record

Current

Next Year Readm Rate

Output Outcome

Output Outcome

Output Outcome

Output Outcome
Result - RMSE

- Population Model (Classification): 0.27
- Regression Model: 0.06
- Last Year Rate: 0.07
- Train Mean: 0.65
Conclusion

- The classification on patient level data does not help to predict future hospital readmission rate if using exactly same past data as prediction horizon data.
- Regression to directly predict hospital’s future readmission rate might be a better approach.
- Future step: better model and feature set for the regression method.
Thank You

Presenters: MIKE WANG, ZINING FAN, QIANG ZHAO, SIYUAN WANG
Appendix
Appendix- Temporal Dependency

- No significant temporal dependency for target variable
- Generated feature set combinations for further testing
Appendix - Population Model - Without Categorical Regrouping

- Downsampling perform better in random forest and XGboost models
- Upsampling perform better in regularized logistic regression
- Considering about the large size of data, we decide to use downsampling approach to fasten our running time
Appendix - Population Model - Compare Encoding method
Appendix - Comparison - Hospital Level (order by size)
Appendix - AUC Last Year

![AUC Score of Readmission Models](image-url)
Appendix - F1 All Three
Appendix - AUC All Three
Appendix - Log Loss All three
Appendix - Accuracy

Accuracy of 367 Hospitals for 3 Models

- Model: Ensemble, Hospital, Population

Accuracy vs. density for different models.
## Appendix - Validation Results

<table>
<thead>
<tr>
<th></th>
<th>validation F1 Score</th>
<th>Log Loss</th>
<th>AUC</th>
<th>Accuracy</th>
<th>PRAUC</th>
<th>Precision</th>
<th>Recall</th>
</tr>
</thead>
<tbody>
<tr>
<td>Population Model</td>
<td>0.4196</td>
<td>0.6822</td>
<td>0.6404</td>
<td>0.5623</td>
<td>0.3503</td>
<td>0.3054</td>
<td>0.6704</td>
</tr>
<tr>
<td>Hospital Model</td>
<td>0.4211</td>
<td>0.8201</td>
<td>0.6074</td>
<td>0.6442</td>
<td>0.3100</td>
<td>0.3418</td>
<td>0.5483</td>
</tr>
<tr>
<td>Ensemble Model</td>
<td>0.5007</td>
<td>0.6497</td>
<td>0.6603</td>
<td>0.6900</td>
<td>0.3499</td>
<td>0.4039</td>
<td>0.6585</td>
</tr>
<tr>
<td>Random Guessing</td>
<td>0.3149</td>
<td>1.0044</td>
<td>0.4929</td>
<td>0.4975</td>
<td>0.2311</td>
<td>0.2321</td>
<td>0.4893</td>
</tr>
</tbody>
</table>
## Appendix - Test Results

<table>
<thead>
<tr>
<th></th>
<th>test F1 Score</th>
<th>Log Loss</th>
<th>AUC</th>
<th>Accuracy</th>
<th>PRAUC</th>
<th>Precision</th>
<th>Recall</th>
</tr>
</thead>
<tbody>
<tr>
<td>Population Model</td>
<td>0.4179</td>
<td>0.6891</td>
<td>0.6326</td>
<td>0.5543</td>
<td>0.3389</td>
<td>0.3029</td>
<td>0.6737</td>
</tr>
<tr>
<td>Hospital Model</td>
<td>0.3331</td>
<td>0.8905</td>
<td>0.5479</td>
<td>0.5894</td>
<td>0.2685</td>
<td>0.2711</td>
<td>0.4319</td>
</tr>
<tr>
<td>Ensemble Model</td>
<td>0.3705</td>
<td>0.6784</td>
<td>0.5945</td>
<td>0.6079</td>
<td>0.2996</td>
<td>0.2994</td>
<td>0.4859</td>
</tr>
<tr>
<td>Random Guessing</td>
<td>0.3307</td>
<td>0.9892</td>
<td>0.5073</td>
<td>0.5076</td>
<td>0.2398</td>
<td>0.2441</td>
<td>0.5122</td>
</tr>
</tbody>
</table>
Appendix- Ensemble Model Result

- Alpha * Population Model + (1 - Alpha) * Hospital Model
- Selected best based on validation F1 Scores
Appendix - Patient Level Data to Hospital Level Data

### Patient Level Data

<table>
<thead>
<tr>
<th>Patient ID</th>
<th>prvdr_num</th>
<th>age</th>
<th>sex</th>
<th>M_onehot</th>
<th>state</th>
<th>State target encoder</th>
<th>dgnrs_cd</th>
<th>dgnrs_cd.S</th>
<th>dgnrs_cd.M</th>
<th>dgnrs_cd.T</th>
<th>Readmision</th>
<th>Year</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>A</td>
<td>50</td>
<td>M</td>
<td>1</td>
<td>NY</td>
<td>0.7</td>
<td>S</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>2016</td>
</tr>
<tr>
<td>2</td>
<td>A</td>
<td>60</td>
<td>M</td>
<td>1</td>
<td>NY</td>
<td>0.7</td>
<td>T</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>2016</td>
</tr>
<tr>
<td>3</td>
<td>A</td>
<td>70</td>
<td>F</td>
<td>0</td>
<td>NY</td>
<td>0.7</td>
<td>M</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>2016</td>
</tr>
<tr>
<td>4</td>
<td>A</td>
<td>80</td>
<td>F</td>
<td>0</td>
<td>NY</td>
<td>0.7</td>
<td>S</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>2016</td>
</tr>
<tr>
<td>5</td>
<td>A</td>
<td>55</td>
<td>M</td>
<td>1</td>
<td>NY</td>
<td>0.7</td>
<td>S</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>2016</td>
</tr>
<tr>
<td>6</td>
<td>A</td>
<td>60</td>
<td>F</td>
<td>0</td>
<td>NY</td>
<td>0.7</td>
<td>T</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>2016</td>
</tr>
<tr>
<td>7</td>
<td>A</td>
<td>65</td>
<td>F</td>
<td>0</td>
<td>NY</td>
<td>0.7</td>
<td>S</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>2016</td>
</tr>
<tr>
<td>8</td>
<td>A</td>
<td>50</td>
<td>M</td>
<td>1</td>
<td>NY</td>
<td>0.7</td>
<td>S</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>2017</td>
</tr>
<tr>
<td>9</td>
<td>A</td>
<td>60</td>
<td>M</td>
<td>1</td>
<td>NY</td>
<td>0.7</td>
<td>M</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>2017</td>
</tr>
<tr>
<td>10</td>
<td>A</td>
<td>70</td>
<td>F</td>
<td>0</td>
<td>NY</td>
<td>0.7</td>
<td>M</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>2017</td>
</tr>
</tbody>
</table>

### Hospital Level Data

<table>
<thead>
<tr>
<th>prvdr_num</th>
<th>age</th>
<th>Male Ratio</th>
<th>State</th>
<th>dgnrs_cd.S ratio</th>
<th>dgnrs_cd.M ratio</th>
<th>dgnrs_cd.T ratio</th>
<th>Year</th>
<th>Current Readmision Rate</th>
<th>Next Year Readmision rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>62.86</td>
<td>0.43</td>
<td>NY</td>
<td>0.57</td>
<td>0.14</td>
<td>0.29</td>
<td>2016</td>
<td>0.43</td>
<td>0.33</td>
</tr>
<tr>
<td>A</td>
<td>60.00</td>
<td>0.67</td>
<td>NY</td>
<td>0.33</td>
<td>0.67</td>
<td>0.00</td>
<td>2017</td>
<td>0.33</td>
<td>0.33</td>
</tr>
</tbody>
</table>