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

Columbia University Data Science Institute

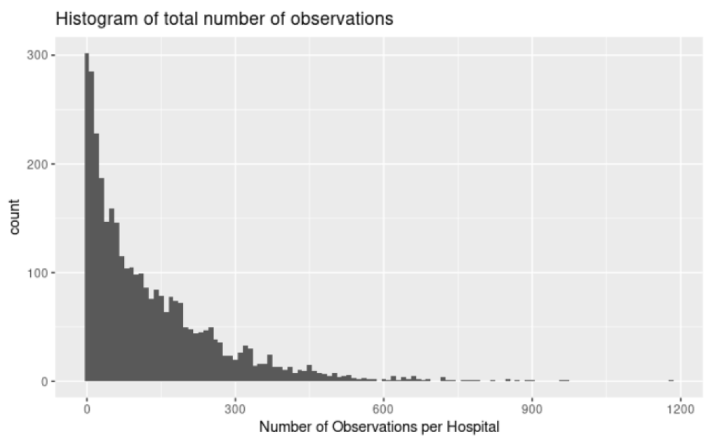
Johnson & Johnson

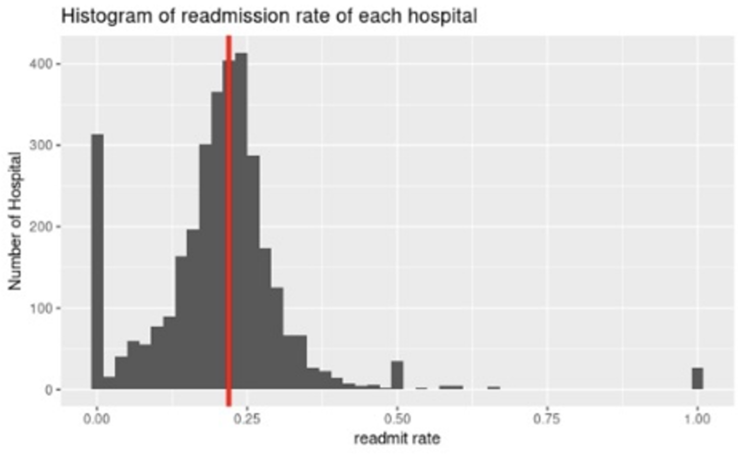
Data Exploratory Analysis

- 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

Columbia University
Data Science Institute

Johnson & Johnson

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 <=5, target encoder with classes>5

Columbia University
Data Science Institute

Johnson & Johnson

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

COLUMBIA UNIVERSITY
Data Science Institute

Johnson & Johnson

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

Columbia University Data Science Institute

Johnson & Johnson

Population Model

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

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

Columbia University Data Science Institute

Johnson & Johnson

Hospital Level Models

Goal: Build different models for each hospital.

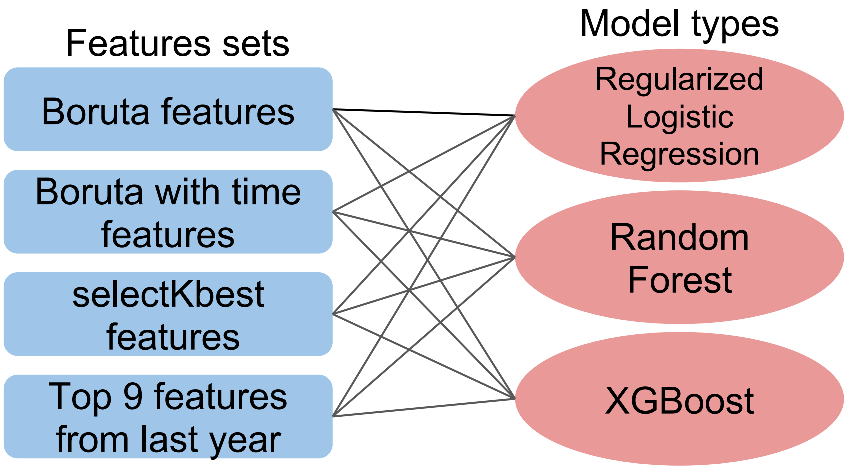
- Data Preprocessing
  - Grouped by hospitals, 3378 hospitals in total
  - Removed features with 0 variance within groups, 80 features left
- Sampling and Train Test Split by chronological order
  - Removed hospitals with less than 66 positive/negative cases, 367 hospitals left
  - 70% train, 15% validation and 15% test
  - Upsampled train dataset for each hospital

Columbia University Data Science Institute

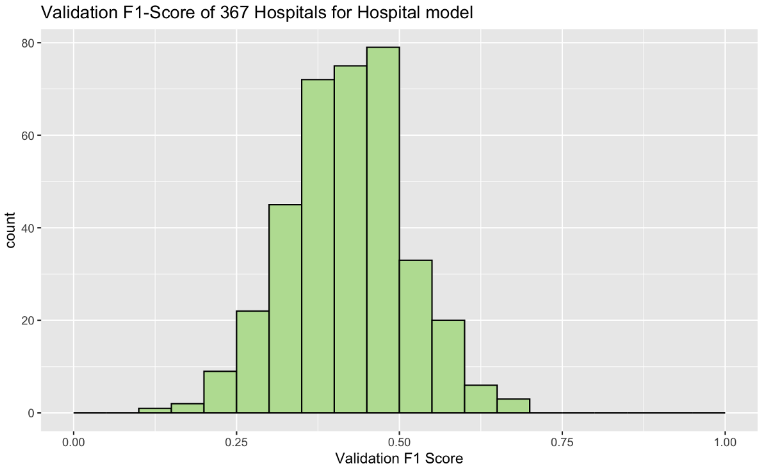
Johnson & Johnson

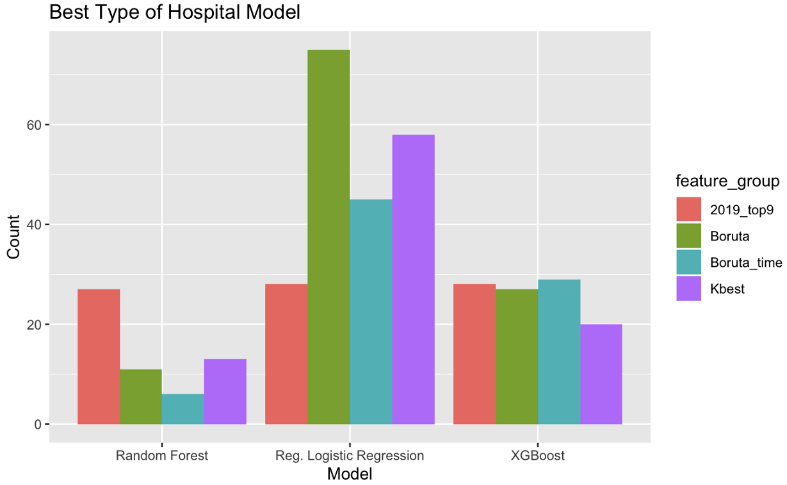
Hospital Level Models

- 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



Hospital Level Models - Results





Columbia University
Data Science Institute

Johnson & Johnson

Ensemble Model Result

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

Columbia University
Data Science Institute

Johnson & Johnson

Comparison - Validation Set

Columbia University Data Science Institute

Johnson & Johnson

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

Validation Scores				
Model	Val F1	Log Loss	AUC	Accuracy
Population Model	0.4196	0.6822	0.6404	0.5623
Hospital Model	0.4211	0.8201	0.6074	0.6442
Ensemble Model	0.5007	0.6497	0.6603	0.6900
Random Guessing	0.3210	0.9915	0.5504	0.5037

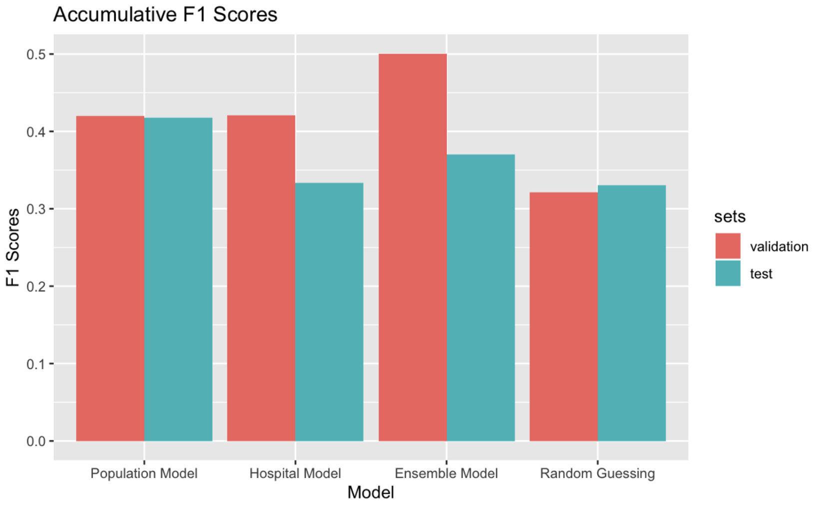
Test Scores				
Model	Test F1	Log Loss	AUC	Accuracy
Population Model	0.4179	0.6891	0.6326	0.5543
Hospital Model	0.3331	0.8905	0.5479	0.5894
Ensemble Model	0.3705	0.6784	0.5945	0.6079
Random Guessing	0.3307	0.9892	0.5073	0.5760

Columbia University
Data Science Institute

Johnson & Johnson

Accumulative F1 Scores

- Hospital model and ensemble model results are biased and not generative
- Does population model works for predicting the hospital's future readmission rate?



COLUMBIA UNIVERSITY
Data Science Institute

Johnson & Johnson

Population Model (Classification) Logic

Current

Data:

Last Year Patient Info	Last Year Readm_flag

Prediction Horizon Patient Info	Next Year Readm_flag

Exact Copy

Model:

Training Input	Output Outcome
Validation Input	Output Outcome
Test Input	Output Outcome

Prediction:

Prediction Input	Prediction Next Year Readm_flag

- 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

Columbia University
Data Science Institute

Johnson & Johnson

Other Approach

Regression on Hospital Level Data

Current

Data: Patient Level to Hospital Level by year

Hospital Last Year Avg Info | Last Year Readm Rate

Next Year Readm Rate

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

Train Input
Validation Input
Test Input

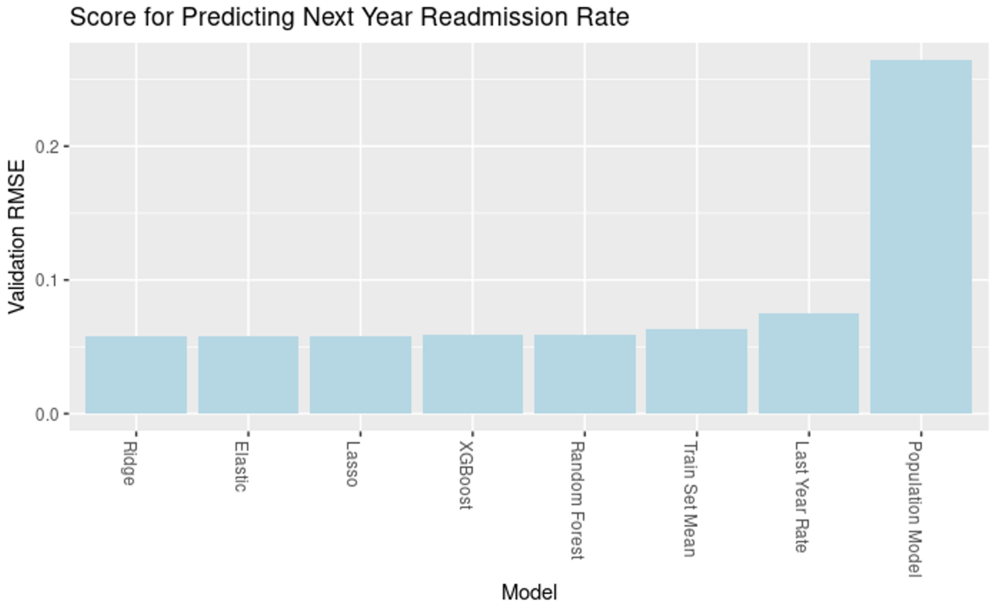
Output Outcome
Output Outcome
Output Outcome

Prediction:

New Hospital with Historical Record

Output Outcome

Result - RMSE



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

COLUMBIA UNIVERSITY
Data Science Institute

Johnson & Johnson

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

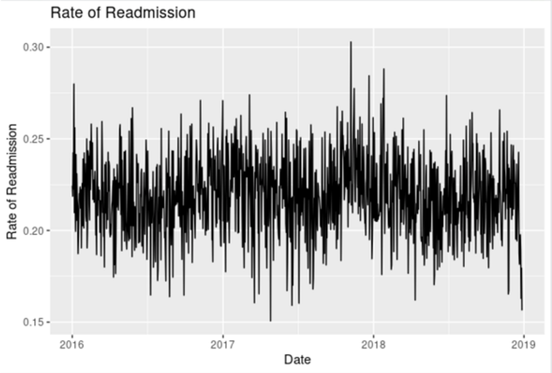
Thank You

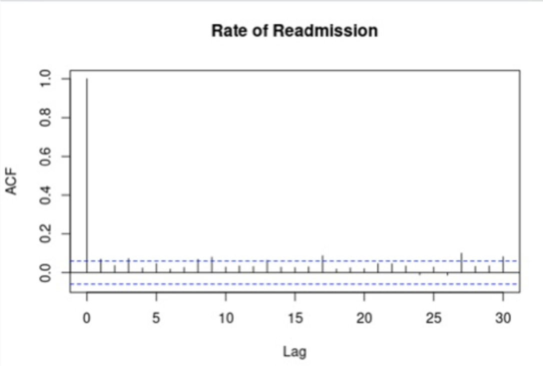
Presenters: MIKE WANG, ZINING FAN, QIANG ZHAO, SIYUAN WANG

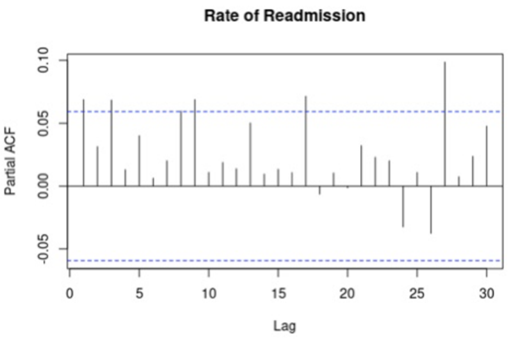
Appendix

Appendix- Temporal Dependency

- No significant temporal dependency for target variable
- Generated feature set combinations for further testing



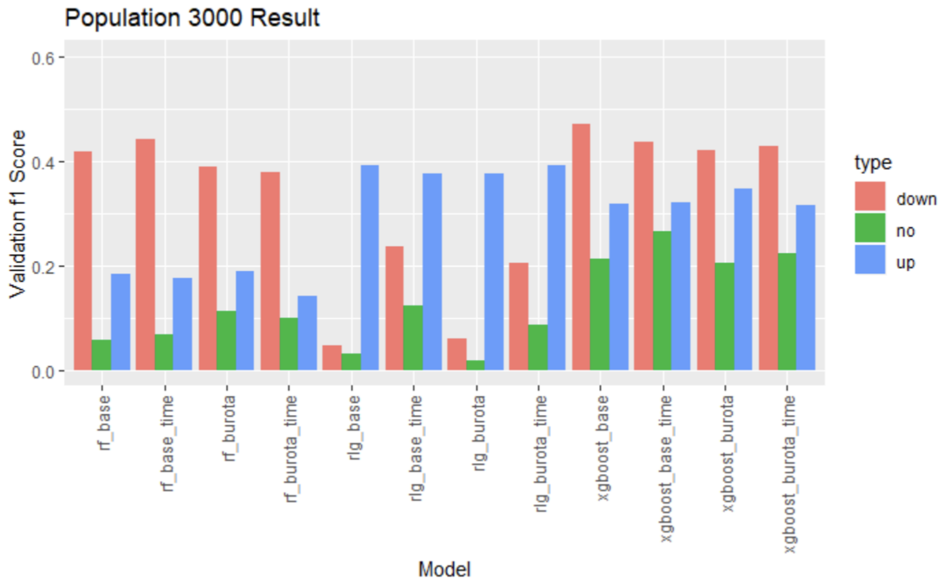




COLUMBIA UNIVERSITY
Data Science Institute

Johnson & Johnson

Appendix- Population Model - Without Categorical Regrouping

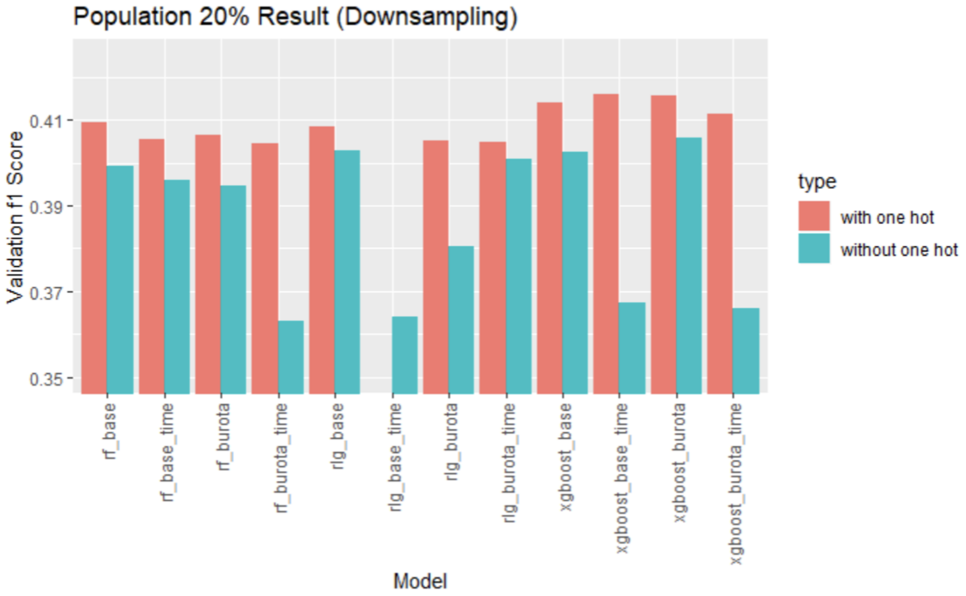


- Downsampling perform better in random forest and XGboost models
- Upsampling perform better in regularized logistic regression
- Considering about the large size of data, we decide to use downsampling approach to fasten our running time

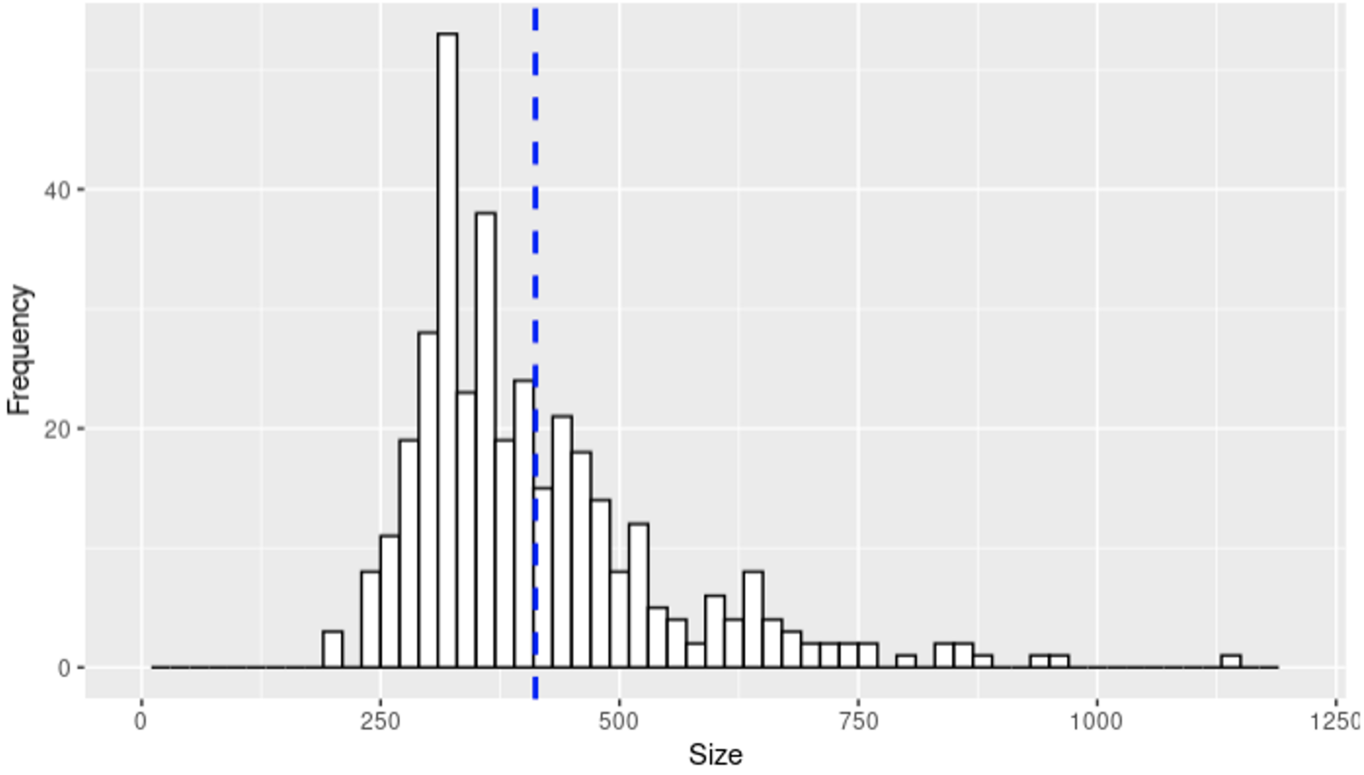
Columbia University Data Science Institute

Johnson & Johnson

Appendix- Population Model - Compare Encoding method



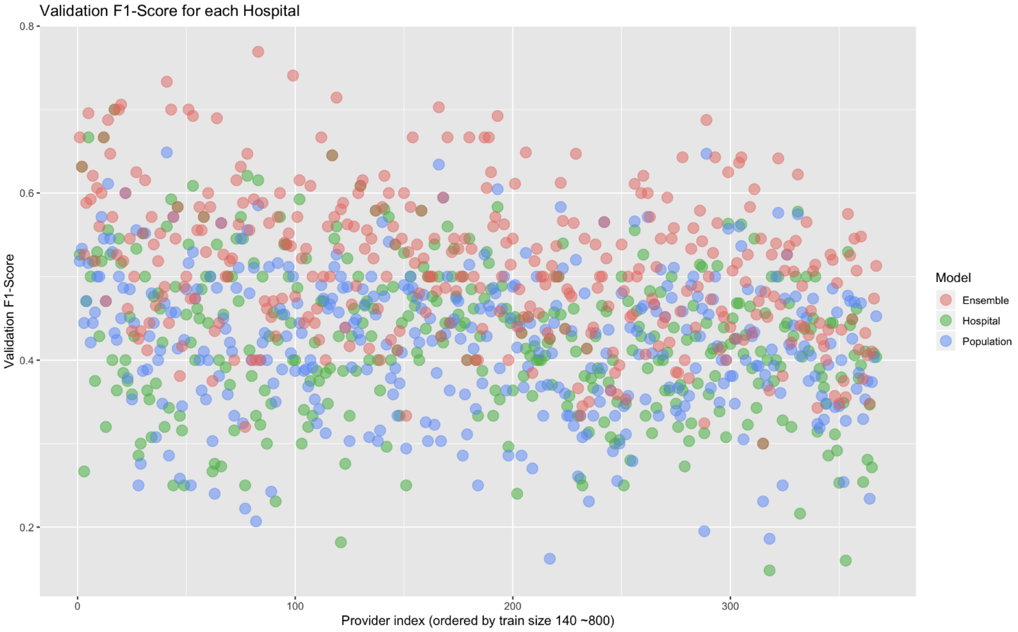
Selected Hospitals' Size Distribution



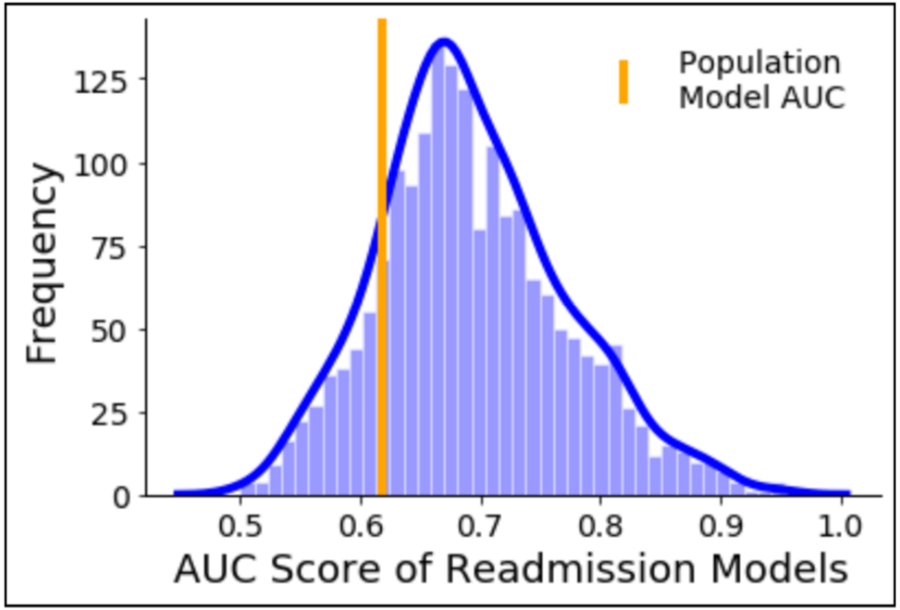
COLUMBIA UNIVERSITY
Data Science Institute

Johnson & Johnson

Appendix - Comparison - Hospital Level (order by size)



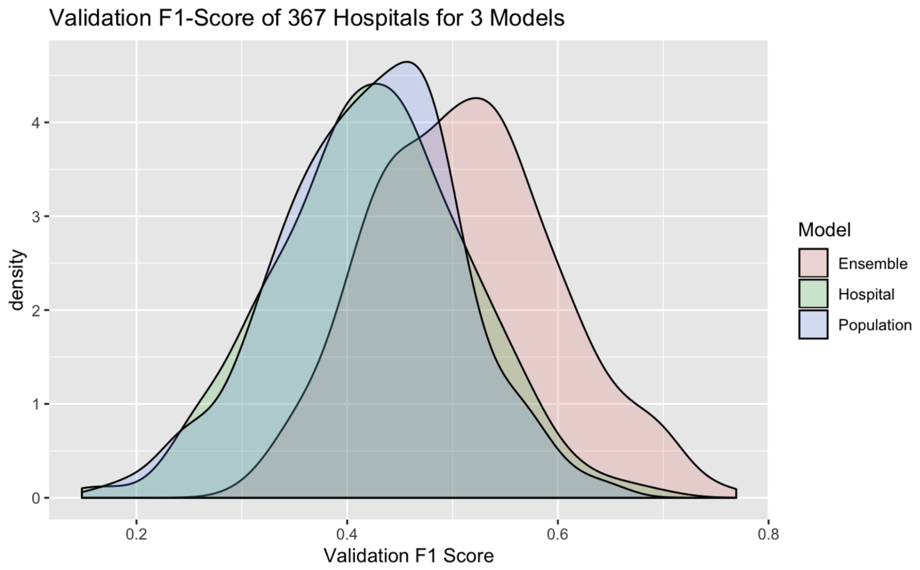
Appendix - AUC Last Year



Columbia University
Data Science Institute

Johnson & Johnson

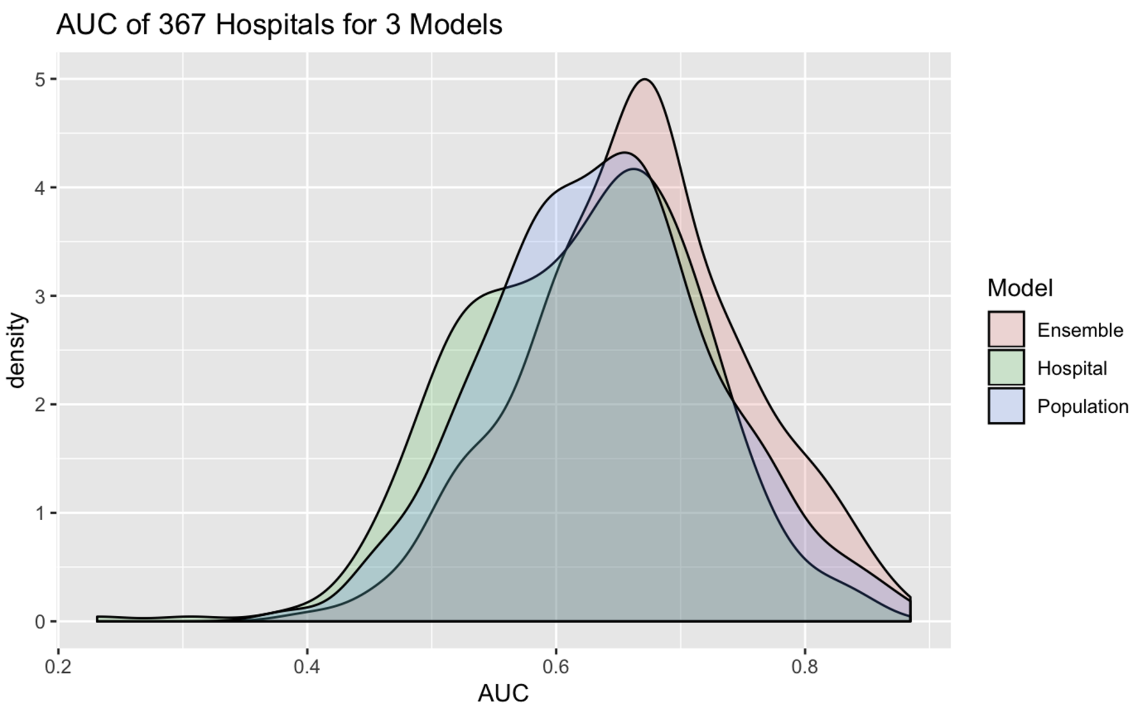
Appendix - F1 All Three



COLUMBIA UNIVERSITY
Data Science Institute

Johnson & Johnson

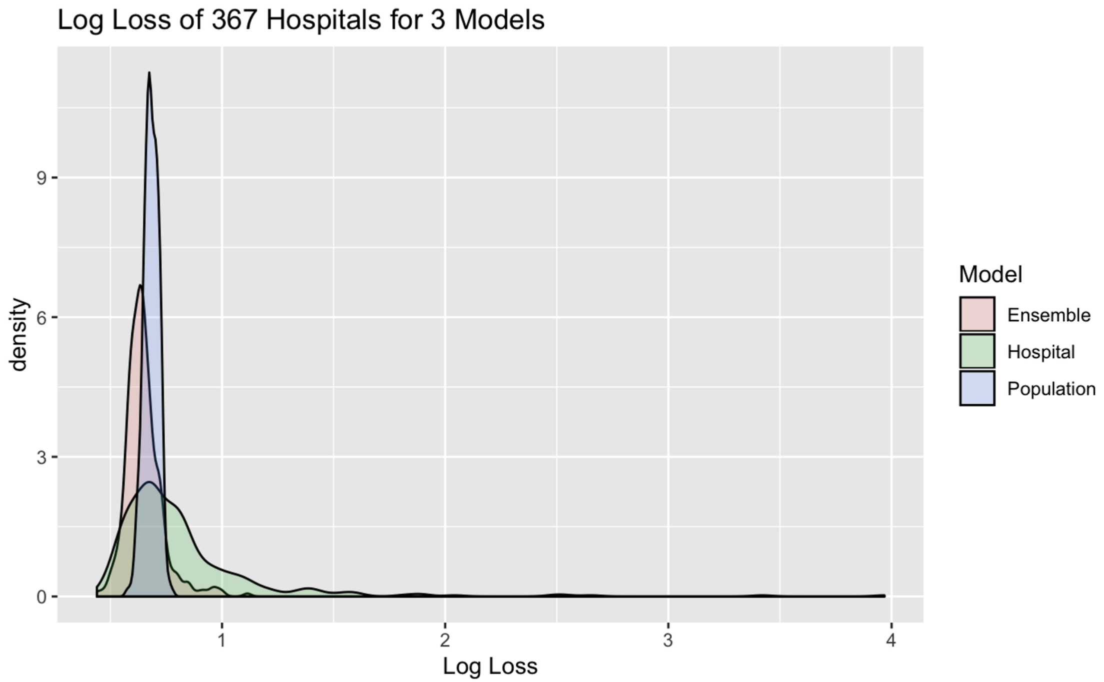
Appendix - AUC All Three



Columbia University Data Science Institute

Johnson & Johnson

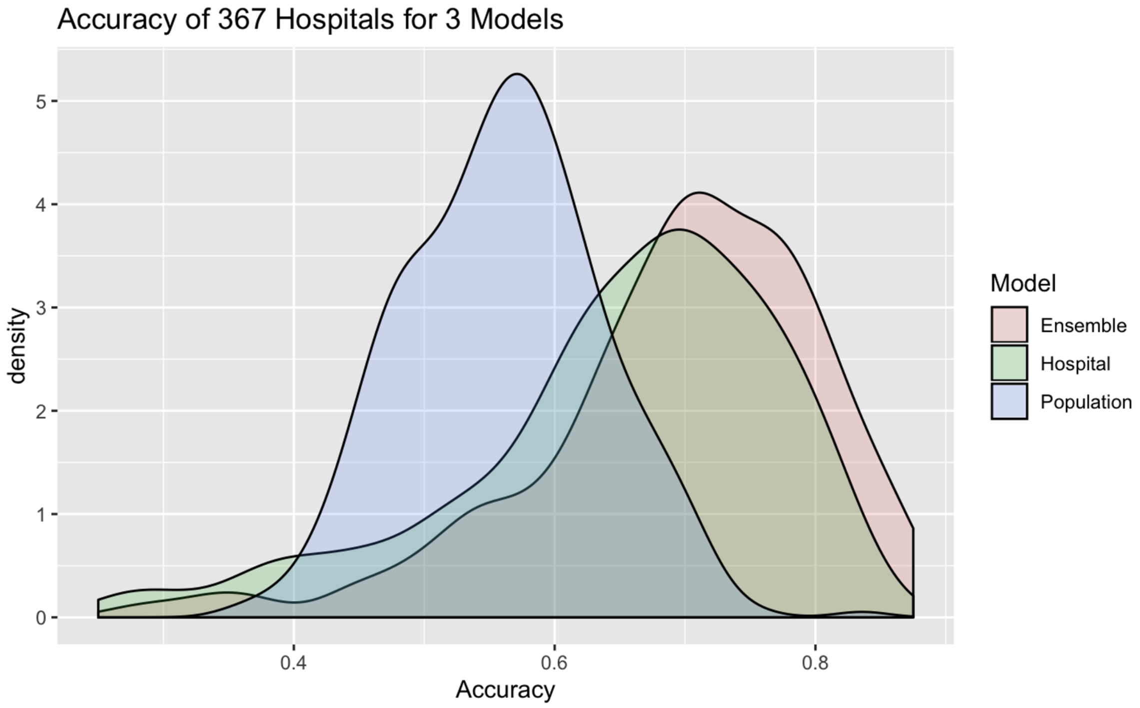
Appendix - Log Loss All three



Columbia University
Data Science Institute

Johnson & Johnson

Appendix - Accuracy



Appendix - Validation Results

Validation Scores							
	validation F1 Score	Log Loss	AUC	Accuracy	PRAUC	Precision	Recall
Population Model	0.4196	0.6822	0.6404	0.5623	0.3503	0.3054	0.6704
Hospital Model	0.4211	0.8201	0.6074	0.6442	0.3100	0.3418	0.5483
Ensemble Model	0.5007	0.6497	0.6603	0.6900	0.3499	0.4039	0.6585
Random Guessing	0.3149	1.0044	0.4929	0.4975	0.2311	0.2321	0.4893

Columbia University
Data Science Institute

Johnson & Johnson

Appendix - Test Results

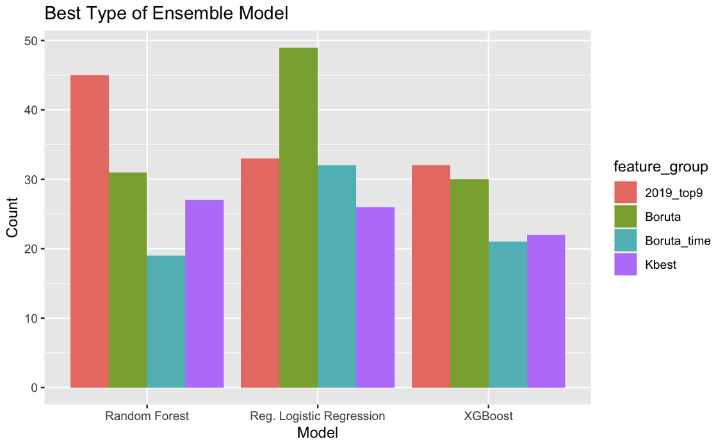
Test Scores							
	test F1 Score	Log Loss	AUC	Accuracy	PRAUC	Precision	Recall
Population Model	0.4179	0.6891	0.6326	0.5543	0.3389	0.3029	0.6737
Hospital Model	0.3331	0.8905	0.5479	0.5894	0.2685	0.2711	0.4319
Ensemble Model	0.3705	0.6784	0.5945	0.6079	0.2996	0.2994	0.4859
Random Guessing	0.3307	0.9892	0.5073	0.5076	0.2398	0.2441	0.5122

Columbia University Data Science Institute

Johnson & Johnson

Appendix- Ensemble Model Result

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



Columbia University Data Science Institute

Johnson & Johnson

Appendix -Patient Level Data to Hospital Level Data

Patient Level Data												
Patient ID	prvdr_num	age	sex	M_onehot	state	State target encoder	dgns_cd	dgns_cd.S	dgns_cd.M	dgns_cd.T	Readmision	Year
1	A	50	M	1	NY	0.7	S	1	0	0	1	2016
2	A	60	M	1	NY	0.7	T	0	0	1	0	2016
3	A	70	F	0	NY	0.7	M	0	1	0	0	2016
4	A	80	F	0	NY	0.7	S	1	0	0	0	2016
5	A	55	M	1	NY	0.7	S	1	0	0	1	2016
6	A	60	F	0	NY	0.7	T	0	0	1	1	2016
7	A	65	F	0	NY	0.7	S	1	0	0	0	2016
8	A	50	M	1	NY	0.7	S	1	0	0	1	2017
9	A	60	M	1	NY	0.7	M	0	1	0	0	2017
10	A	70	F	0	NY	0.7	M	0	1	0	0	2017

Hospital Level Data									
prvdr_num	age	Male Ratio	State	dgns_cd.S ratio	dgns_cd.M ratio	dgns_cd.T ratio	Year	Current Readmision Rate	Next Year Readmission rate
A	62.86	0.43	NY	0.57	0.14	0.29	2016	0.43	0.33
A	60.00	0.67	NY	0.33	0.67	0.00	2017	0.33	