Predict COVID-19 Hospitalization Based on Patient Basic Info

William Mai

December 15th 2021

Introduction

- Predict whether a confirmed Covid-19 patient will be hospitalized
- Provide timely and dedicated care to prevent hospitalization
- The result can also be used to predict the demand for hospital beds related to Covid

Data

37.5M rows of patient-level data for all Covid-19 cases reported in the US gathered by CDC

- Case month (recent 3 months is picked)
- State, county
- Age group
- Sex
- Race
- Known exposure
- Symptom status
- Hospitalized

Training Workflow

1 Feature Engineering

- Grouping categories
- One Hot Encoder for categorical features

2 Establish Baseline

- Train on Logistic Regression model

3 Try Different Models

- Tree Classifier
- **Random Forest**
- Gradient Boosted Trees (XGBoost)
- Naive Bayes

4 Class Imbalance

- Resampling (over, **under**, SMOTE)
- Class Weight

5 Hyperparameter Tuning

- # of estimators
- Max depth
- Min samples split
- Min samples leaf

Feature Engineering

Exposure

- Yes
- Missing
- Unknown
- NaN

Race

- White
- Black
- Asian
- Other
- American Indian/Alaska Native
- And more...

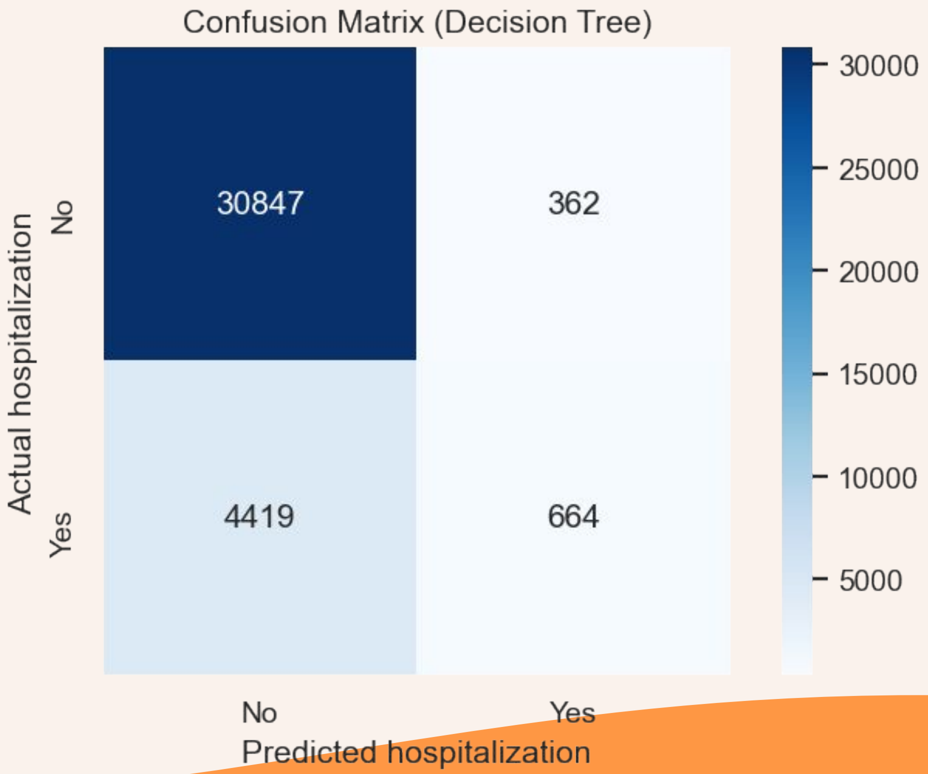
Location

Group states/territories into four regions based on the official FIPS code

Result - Preliminary Training

	Logistic	Tree	Random Forest	XGBoost
Precision	0.6101	0.6472	**0.6362**	0.6526
Recall	0.1161	0.1306	**0.1328**	0.1216
F1	0.1950	0.2174	**0.2197**	0.2050

Result - Confusion Matrix



Result - Class Imbalance Solution

	Oversampling	Undersampling	SMOTE	Class Weight (1:5)
Precision	0.3193	**0.3135**	0.3151	0.3197
Recall	0.6681	**0.6785**	0.6754	0.6685
F1	0.4321	**0.4288**	0.4297	0.4325

Best from 1:1 to 1:10

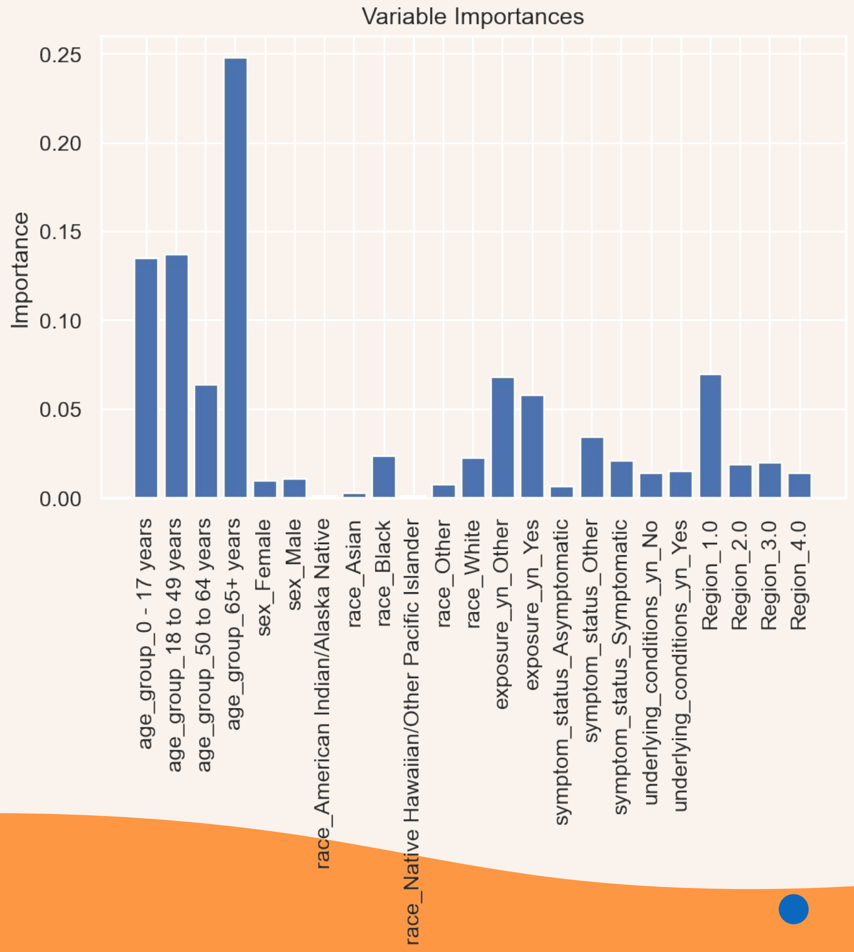
Recall = 0.6783

Final Model

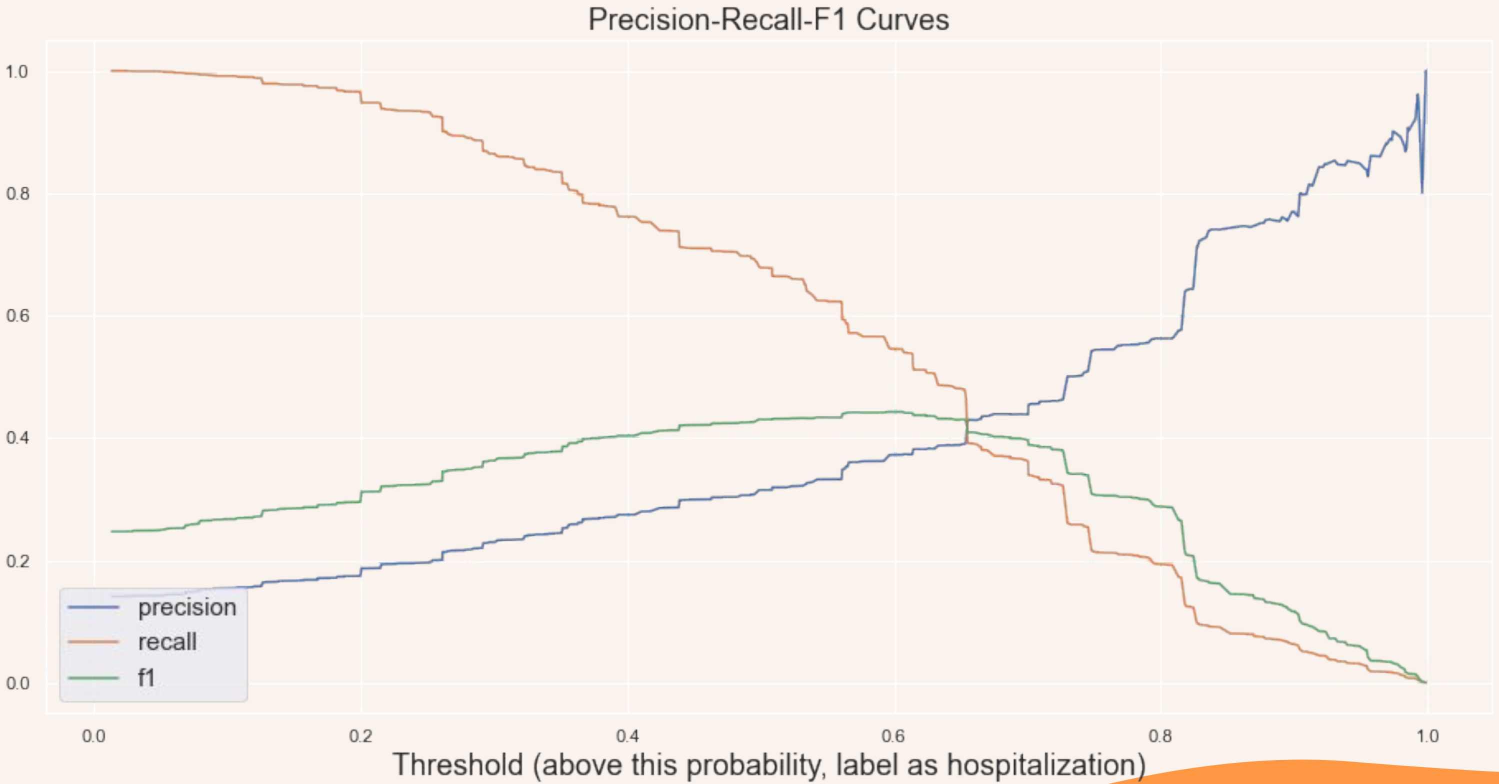
- Random Forest
- Undersampling (⅕ of 0)
- n_estimators = 100
- min_samples_split = 5
- min_samples_leaf = 4
- And more...

Feature Importances

“Age group” appears to be the most important feature, followed by “Known Exposure.”



Result - Precision vs Recall



Conclusion

Predictability

The model can identify 68% of people who will likely get hospitalized later

Adaptability

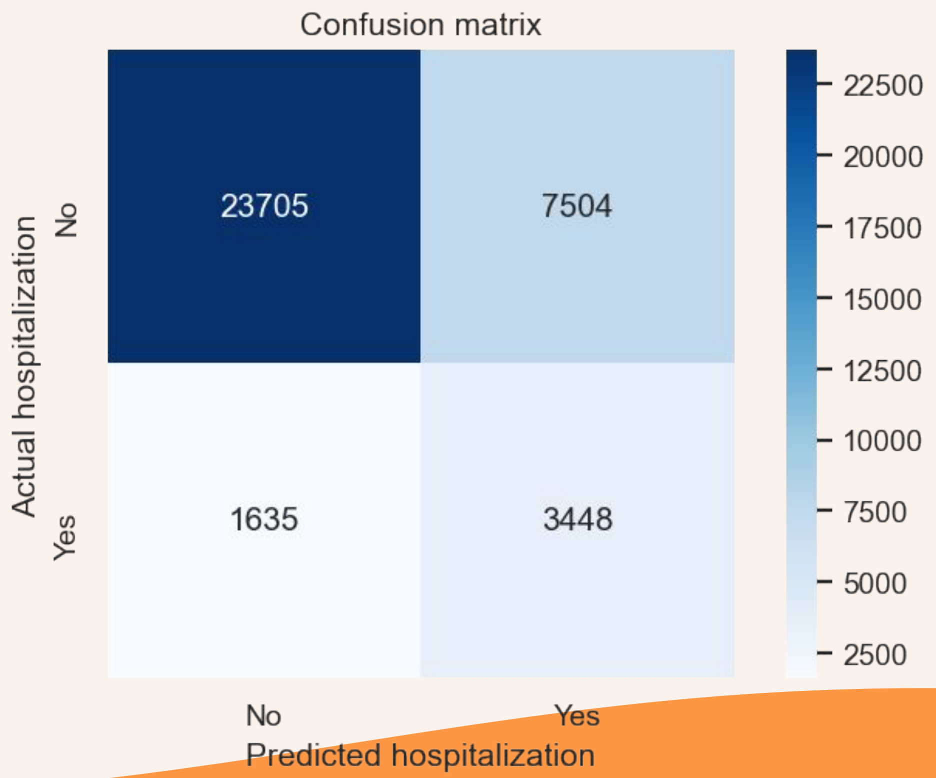
The model can be adjusted for an ideal true positive rate based on the current capacity of the healthcare system

Thanks!

Do you have any questions?

Appendix

Confusion Matrix - Final Model



Result - ROC Curve