Classification Project

Predict COVID-19 Hospitalization Based on Patient Basic Info

By William (Wei-Lun) Mai

Abstract

COVID-19 has changed everyone's lives drastically since 2020. We are most afraid of our loved ones being hospitalized due to Covid, and what saddens us the most is that it sometimes leads to the death of our loved ones. What if we can predict whether a person will be hospitalized when a confirmed case of Covid shows up? Then we can act proactively to care and provide guidance to the person, and hopefully, it will prevent the person from being hospitalized. On the other hand, the measure can also help ease the hospital bed utilization rate, especially in areas with few empty beds.

Design

With the patient-level detail of each Covid-19 confirmed case reported, we use it to train various classification models after proper feature engineering is conducted on the dataset. The models are compared based on their recall and F1 scores. The model's target is to predict whether a person will be hospitalized when the case is first reported, usually before they are hospitalized. After identifying the people with higher risk, the administration will provide better care and guidance and eventually help them avoid hospitalization.

Data

The dataset contains patient-level data for all Covid-19 cases reported in the US gathered by CDC. Currently, it stands at 37.5 million rows. We use the data from August to October 2021 to train our models. The dataset contains some basic info of each confirmed case, such as age group, gender, location, race, hospitalized or not, etc. Various feature engineering techniques are used to ensure the ideal performance of our models is reached. The additional dataset containing the FIPS code of US states and territories is used to group the location of patients into four regions.

Algorithms

Various classification models are used to evaluate their effectiveness for this project. Mainly recall and F1 are used to assess the performance of the models. We use logistic regression to establish baseline metrics. Later we compare scores from tree classifier, random forest, gradient boosted trees (XGBoost), and Naive Bayes. We then tackle the class imbalance issues while training with the random forest model. Last but not least, we conduct hyperparameter tunning to find the best random forest model for this project.

Tools

- Numpy and Pandas for manipulating data
- Matplotlib and Seaborn for plotting
- Sklearn and xgboost for classification modeling

Communications

Presentation pdf, write-up, and Excel files can be found on the following GitHub page:
https://github.com/zyzzyva1423/Classification_Project