Marcus Woo

BrainStation

Application of Machine Learning to predict Covid-19 mortality risk

Background and Problem Statement

The outbreak of the Covid-19 pandemic has placed a huge burden on our health care system. Early prediction of mortality risk among infected individuals can decrease mortality rate by ensuring efficient allocation of health resources and better implementation of health interventions. Lots of research have been done on applying different machine learning and deep learning models to predict the outcome of infected patients and identify risk factors related to Covid, and these models have shown promising results in the early prediction of death among Covid-19 patients. However, many of these datasets contain confidential information about the patients and are not available to the public. The main objective of this project is to apply and compare three different machine learning models in predicting covid mortality risk using publicly accessible data.

Data

The dataset is downloaded from the City of Toronto's open data portal as a CSV file. It contains demographic, geographic information, and hospital records related to Covid for all confirmed cases that are reported to Toronto Public Health since January of 2020. Other information on average household income and population density is downloaded from the 2016 census through Wellbeing Toronto and is incorporated with the dataset.

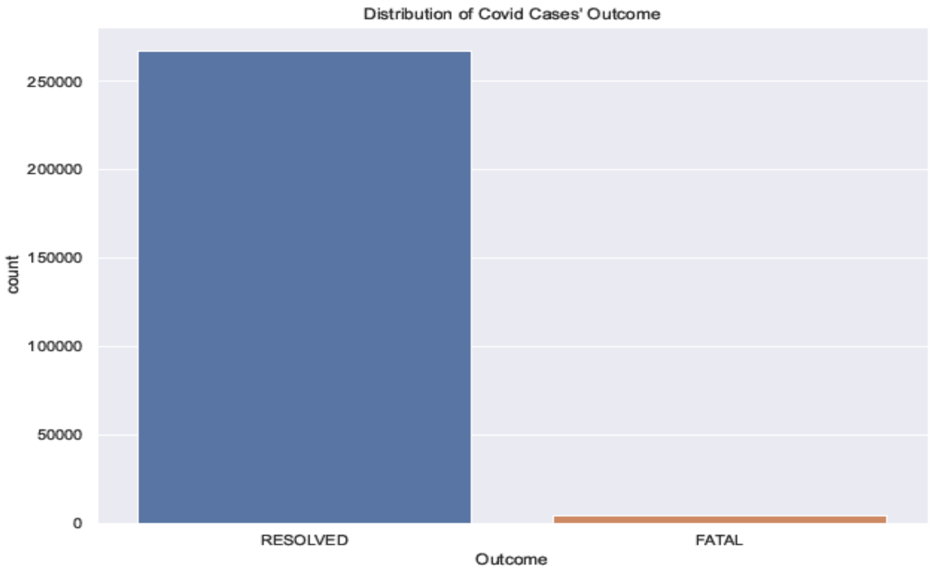
Preprocessing, feature engineering, and EDA

The dataset had 277473 rows and 21 columns, and each row of data represents an individual patient. The target variable is the patient's outcome of either `Resolved` or `Fatal`. This is a binary classification project. Independent variables included different demographic characteristics such as age, gender, income, and population density. Other features include episode date, which best estimates when the disease was acquired by the patients, and information on the patient's hospitalization record from Covid.

The virus is evolving and the risk of severe illness from Covid changes over time, so a time variable would be a good predictor of mortality risk. Month and Year are extracted from `episode date`. Another variable `days elapsed` is also created by calculating the difference between `episode date` and the date of the first case reported on January 21, 2020, to try and capture the magnitude of the virus in different time periods.

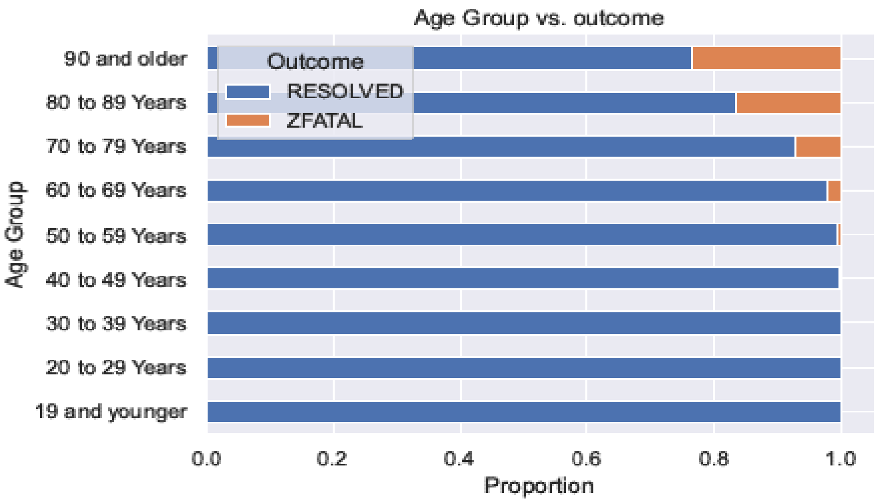
Missing values and columns that contain redundant information are removed. The final dataset has 270962 rows of data and 13 features.

In EDA, it is found that our dataset has a highly unequal distribution of classes, so measuring accuracy score will not be a good metric for our dataset. We will adjust the class weights later at the modeling stage and look at AUC scores to evaluate the performance of our classification models.



- There is a huge imbalance between the two classes with 266941 `Resolved` cases and only 4021 `Fatal` cases.

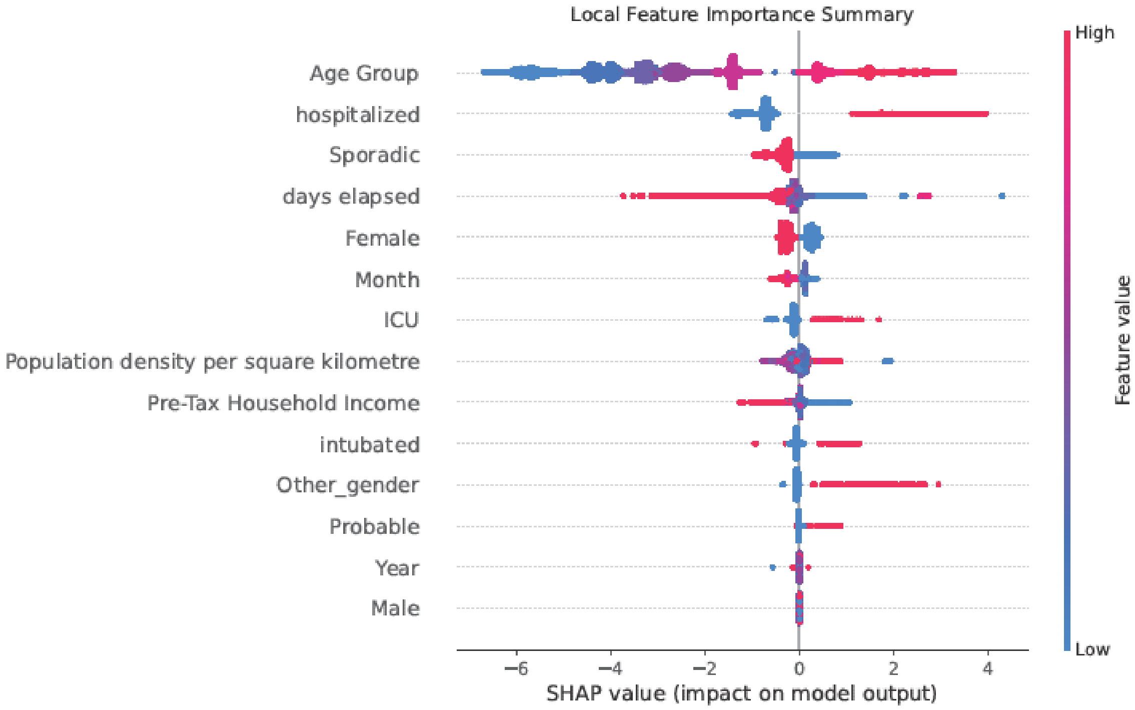
The distribution of patients' outcomes against different categorical variables is also explored in EDA. We see a relationship between Age and Covid mortality risk, where the older age group has a higher proportion of Covid deaths. This suggests 'Age Group' could be an important feature of our model output.



Modeling and Model Evaluation

Three machine learning models (LASSO regression, Random Forest, and XGBoost) are applied and GridSearchCV is performed for hyperparameter tuning. Class weights are adjusted when running the model to account for highly imbalanced classes. Both XGBoost and LASSO regression had very high AUC scores of 0.940 and 0.938, with XGBoost performing slightly better in terms of accuracy scores and F1 score. Although Random Forest performed the worst among the three models, it still had a high AUC score of 0.92.

For model interpretation, feature importance is computed using the SHAP package for our XGBoost model.



Features are ranked in order by calculating the mean absolute SHAP values with the most important feature at the top. ‘Age Group’ is determined to be the leading factor that influences the mortality risk of Covid-19 patients. The older age group (high feature value is represented with red) has a positive SHAP value and is associated with increased mortality risk, whereas younger age groups (low feature value is represented with blue) have negative SHAP values and have lower mortality risk.

Another important feature is ‘hospitalized’, this is a binary variable that takes values of 1 and 0. Patients who had been hospitalized (value of 1) are associated with increased mortality risk. `Days elapsed` is the four most important feature, this confirms time is indeed an important predictor of Covid mortality risk.

Conclusion

Machine learning models performed very well in identifying high-risk patients with Covid-19 using publicly available data. Variables such as age, gender, history of hospitalization, and time are shown to be good predictors of Covid mortality risk. With only a few easily accessible variables, our models had achieved great results, this lays important groundwork for applying machine learning to scale in fighting the pandemic.

Next Step

XGBoost model has been shown to be the better model and achieved the best results. By spending more time on fine-tuning and optimizing the model, I believe it can yield even better results. Another approach that can be taken to predict mortality is to train a neural network and see if it yields better performance than regression and tree-based models. I want to test my model with a larger population, as well as incorporate other demographic factors such as race, ethnicity, and education to predict Covid mortality risk.