**Executive Summary**

**XGBoost Classification Analysis of Comorbidities for Predicting Mortality from COVID**

**J. Amanda Terzin**

**Western Governors University**

**Master of Science in Data Analytics Capstone**

**November 26, 2022**

**Statement of the Problem and the Hypothesis**

The recent Covid-19 pandemic created an unprecedented strain on hospitals and healthcare organizations. The need for beds outnumbered the available beds (French et al., 2021). The inpatient staff was stretched thin while the outpatient staff was furloughed due to clinic closures (Office of the Assistant Secretary for Planning and Evaluation, 2022). Shortages of supplies and medications for patients and protective equipment for staff were caused by both increased use and supply chain issues (Hick, 2021). The fiscal impact was devastating as well. Hospitals faced a paradox of increased business with less revenue. Reimbursement was down due to an increase in uninsured and Medicaid patients laid off from their jobs. Elective surgeries and procedures that are a major revenue source were halted during surges (Balser, 2021).

 Healthcare organizations need a predictive tool to inform ways to minimize the impact of Covid on operations. The hypothesis of this study is that the comorbidities of patients diagnosed with Covid-19 can be used to predict the risk of mortality with statistical significance.

**Summary of the Data Analysis Process**

**The Data Set**

The data set used for this study is publicly available synthetic data generated using The MITRE Corporation’s SyntheaTM Synthetic Patient Population Simulation. The data set was created to facilitate modeling Covid data without privacy and security risk to patients (Walonoski et al., 2020). The set is comprised of 16 csv files of which 4 were used.

**Table 1**

*Summary of Original Data use*

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **csv** | **Rows** | | **Columns** | | **Information** |
| **Original** | **Kept** | **Original** | **Kept** |
| patients | 124,150 | 71,329 | 25 | 3 | Id, DOB, Gender |
| observations | 1,621,9969 | 132,648 | 8 | 4 | Id, Covid test, Cause/date of death,  Smoking Status |
| conditions | 1,143,900 | 114,349 | 6 | 3 | Id, Medical Conditions |
| encounters | 1,881,954 | 21,929 | 15 | 6 | Id, Hospital Days, ICU Days |

**Data Extraction and Preparation**

**The 4 csv files were read into Python in separate data frames. The needed data were extracted from each. The strategy to tackle the large amount of data contained in the files was to start with the patients’ data frame, structured with one line per patient, and use it as the anchoring base. Once the data from the remaining 3 data frames were extracted and processed, it was added to the patients’ data frame by patient Id. Observations, conditions, and encounters had more than a million rows with multiple rows per patient. The data was extracted and prepared as follows:**

**Table 2**

***Summary of Extracted Data***

|  |  |  |
| --- | --- | --- |
| **Data Frame** | **Column** | **Information** |
| observations | Covid test | -Identify Covid patients, drop non-Covid patients in all 4 data frames  -Compute age at the time of diagnosis and add the Age column to the patients’ data frame  -Drop pediatric patients in all 4 data frames |
| Cause/date of death | -Identify patients who died from Covid  -Create target variable Mortality in the patients’ data frame |
| Smoking Status | -Extract the value for each patient  -Pivot values to 3 separate columns one-hot encoded  -Concatenate to the conditions data frame |
| conditions | Medical Conditions | -18 out of 170 conditions dated before Covid diagnosis were extracted for each patient  - 2 pairs of similar conditions consolidated for a total of 16 conditions  - Pivot conditions to 16 separate columns one-hot encoded  -Concatenate to the patients’ data frame |
| encounters | Hospital Days  ICU Days | -Total days for hospital admissions and ICU stays for diagnosis of Covid computed  -Added to the patients’ data frame as columns |

**The final data frame contains 71,091 rows, representing 1 patient per row, and 23 columns.**

**Table 3**

***Summary of Final Variables***

|  |  |  |  |
| --- | --- | --- | --- |
| **Field** | **Data Type** | **Variable Type** | **Use** |
| Id | Categorical | Independent | Index |
| ICU\_days | Continuous | Independent | Demonstration |
| Total\_Hosp\_days | Continuous | Independent | Demonstration |
| Age | Continuous | Independent | In model |
| Gender | Categorical | Independent | In model |
| Alzheimer`s\_disease | Categorical | Independent | In model |
| Asthma | Categorical | Independent | In model |
| Obesity | Categorical | Independent | In model |
| Chronic\_congestive\_heart\_failure | Categorical | Independent | In model |
| Chronic\_kidney\_disease | Categorical | Independent | In model |
| Chronic\_obstructive\_bronchitis | Categorical | Independent | In model |
| Coronary\_Heart\_Disease | Categorical | Independent | In model |
| Diabetes | Categorical | Independent | In model |
| History\_of\_myocardial\_infarction | Categorical | Independent | In model |
| Hyperlipidemia | Categorical | Independent | In model |
| Hypertension | Categorical | Independent | In model |
| Hypertriglyceridemia | Categorical | Independent | In model |
| Prediabetes | Continuous | Independent | In model |
| Pulmonary\_emphysema | Categorical | Independent | In model |
| Stroke | Categorical | Independent | In model |
| Current\_smoker | Categorical | Independent | In model |
| Former\_smoker | Categorical | Independent | In model |
| Mortality | Categorical | Dependent | In model |

**Exploratory Data Analysis**

Exploratory data analysis was performed on the prepared data. **A comparison of hospital utilization between deceased and recovered patients revealed both total hospital days and ICU days are significantly higher for the mortality group. Bivariate visualizations of the categorical variables were not particularly revealing due to the imbalanced nature of the data.**  The target for prediction is the minority by almost 20:1. Graphs were dominated by the negative class, making the positive class visually negligible. A distribution plot of the one continuous variable, Age, demonstrated a distinct difference in mean between the two classes, but no linear relationship. **A correlation heatmap of the predictors showed significate multicollinearity between some of the independent variables.**

**Statistical Tests**

**The Chi-squared test for independence between the independent categorical variables and the target showed the relationship was statistically significate, p-value < 0.5, for all of them except Asthma.**

**Analysis**

**The model was created using an eXtreme Gradient Boosting (XGBoost) classifier. This method was chosen for its robustness to multicollinearity and its ability to identify non-linear relationships (Gupta, 2020).** The data was divided into a train and test set with an 80/20 stratified split. **Hyperparameter tuning was accomplished with the HyperOpt library that utilizes Bayesian optimization (**Koehrsen, 2018)**. Mitigation measures for the imbalanced data: 1)** Cross-validation with sklearn’s RepeatedStratifiedFold to maintain a consistent distribution of the minority class with each fold. 2) The appropriate evaluation metric, the area under the precision-recall curve (AUC-PR), does not include the large true-negative group in its equation allowing a greater focus on the minority true-positive class. 3) Tuning the scale\_pos\_weight parameter to balance the weights (Mumtaz, 2020). The AUC-PR was also chosen as the best metric for this study because identifying all of the positive-class, patients at high risk of mortality, is considered a priority over balancing the true positive and true negative rates as the ROC-AUC does (Brownlee, 2020).

The model was trained on the training set using the best parameters and predictions made on the test set. The predictions were evaluated with a classification report, a confusion matrix, a ROC-AUC curve, and an AUC-PR curve. Further interpretation of the results with the Python SHAP package was done to look at the features’ contribution to the model.

**Outline of the findings**

**Result Statistics**

1) Classification Report and Confusion Matrix

Graphical user interface, application

Description automatically generated

Graphical user interface, application

Description automatically generated

**2) Performance Curves**

Graphical user interface, application

Description automatically generatedGraphical user interface, application

Description automatically generated

By the usual metrics for evaluating classification tasks, accuracy and AUC, the model performed reasonably well with 0.75 and 0.83 respectively. Interpreting the results in the context of the goal of identifying all of the true positives, it performed relatively well with an AUC-PR of 0.22 which is 4.5 times the baseline of 0.049, or the random chance of choosing the right class.

Given these results, the comorbidities of patients diagnosed with Covid-19 can be used to predict the risk of mortality with statistical significance. However, the fact it missed 166 out of the 619 true positives and falsely labeled 3,387 as positive, shows that there are other factors affecting mortality.

**Feature Importance**

The SHAPS package allows us to look at the features’ contribution to the model.

Ranked Importance- the top 11 conditions that impacted the predictions.

Graphical user interface, text, application

Description automatically generated

Contribution by Feature Value- how the specific values of the features contributed to the predictions.

Graphical user interface, application

Description automatically generated

**Limitations**

One of the limitations of the study is that the data is synthetic. Better insights might be found in a database of actual patients. Additionally, the comorbidities are defined somewhat generally. Diabetes, for instance, covers a wide range of severity. Quantifying the severity with tools such as the Diabetes Severity Score (DISSCO) could improve the predictive value (Zghebi, 2020). Additionally, because the data is synthetic, the model will have to be validated with actual patient data before the results are actionable by the stakeholders.

**Proposed Actions**

1. Apply the model to an actual patient database to evaluate performance.
2. Consider including clinical measurements for each patient at the time of diagnosis. A patient who has been sick for days before coming in to be seen may already be too late to benefit from treatment as well as a patient diagnosed at the first sign of sickness (Sun et al. 2020).
3. Include Covid vaccination status to see if it affects mortality. At the time of this study, Covid vaccinations were not documented in the dataset.

**Expected Benefits of the Study**

Once validated with actual patient data, the results will offer insight to healthcare organizations to develop protocols for managing patients and resources during Covid surges.

1. Prevention- Maximise precautions for at-risk populations with education and vaccination.
2. Early intervention- Early treatment once diagnosed reduces the severity of the infection reducing the need for hospitalization.
3. Outpatient treatment- Identifying lower-risk patients allows for treatment safely as an outpatient, reducing hospital volume.

(Balser, 2021)

**References**

Balser, J., J. Ryu, M. Hood, G. Kaplan, J. Perlin, and B. Siegel. (2021). Care Systems COVID-19 Impact Assessment: Lessons Learned and Compelling Needs. *NAM Perspectives.* Discussion Paper, National Academy of Medicine, Washington, DC. [Care Systems COVID-19 Impact Assessment: Lessons Learned and Compelling Needs - National Academy of Medicine (nam.edu)](https://nam.edu/care-systems-covid-19-impact-assessment-lessons-learned-and-compelling-needs/?gclid=Cj0KCQiAyMKbBhD1ARIsANs7rEGXsk8MFS5gaVbbD4Gt1aQY0w-y2tzLRcEaG-0HMORwJBJ0a71F7IEaAjtXEALw_wcB)

Brownlee, J. (2021). Tour of Evaluation Metrics for Imbalanced Classification. Retrieved November 10, 2022, from [Tour of Evaluation Metrics for Imbalanced Classification - MachineLearningMastery.com](https://machinelearningmastery.com/tour-of-evaluation-metrics-for-imbalanced-classification/)

French G, Hulse M, Nguyen D, et al. (2021). Impact of Hospital Strain on Excess Deaths During the COVID-19 Pandemic — United States, July 2020–July 2021. *MMWR Morb Mortal Wkly Rep*, *70,* 1613–1616. [Impact of Hospital Strain on Excess Deaths During the COVID-19 Pandemic — United States, July 2020–July 2021 | MMWR (cdc.gov)](https://www.cdc.gov/mmwr/volumes/70/wr/mm7046a5.htm#suggestedcitation)

Gupta, S. (2020). Pros and cons of various Machine Learning algorithms. Retrieved November 10, 2022, from [Pros and cons of various Machine Learning algorithms | by Shailaja Gupta | Towards Data Science](https://towardsdatascience.com/pros-and-cons-of-various-classification-ml-algorithms-3b5bfb3c87d6)

Hick, J. L., D. Hanfling, M. Wynia, and E. Toner. (2021). Crisis Standards of Care and COVID-19: What Did We Learn? How Do We Ensure Equity? What Should We Do? *NAM Perspectives.* Discussion, National Academy of Medicine, Washington, DC. [Crisis Standards of Care and COVID-19: What Did We Learn? How Do We Ensure Equity? What Should We Do? - National Academy of Medicine (nam.edu)](https://nam.edu/crisis-standards-of-care-and-covid-19-what-did-we-learn-how-do-we-ensure-equity-what-should-we-do/)

Koehrsen, W. (2018). A Conceptual Explanation of Bayesian Hyperparameter Optimization for Machine Learning. Retrieved November 15, 2022, from Retrieved November 15, 2022, from [A Conceptual Explanation of Bayesian Hyperparameter Optimization for Machine Learning | by Will Koehrsen | Towards Data Science](https://towardsdatascience.com/a-conceptual-explanation-of-bayesian-model-based-hyperparameter-optimization-for-machine-learning-b8172278050f)

Mumtaz, A. (2020). How to Effectively Predict Imbalanced Classes in Python. Retrieved November 08, 2022, from [How to Predict Imbalanced Classes in Python | Towards Data Science](https://towardsdatascience.com/how-to-effectively-predict-imbalanced-classes-in-python-e8cd3b5720c4)

Office of the Assistant Secretary for Planning and Evaluation. (2022). Impact of the COVID-19 pandemic on the hospital and outpatient clinician workforce: challenges and policy responses. (Issue Brief No. HP-2022-13). U.S. Department of Health and Human Services. [aspe-covid-workforce-report.pdf (hhs.gov)](https://aspe.hhs.gov/sites/default/files/documents/9cc72124abd9ea25d58a22c7692dccb6/aspe-covid-workforce-report.pdf)

Sun, Q., Qiu, H., Huang, M., & Yang, Y. (2020). Lower mortality of COVID-19 by early recognition and intervention: experience from Jiangsu Province. *Annals of intensive care, 10(1)*, 33. https://doi.org/10.1186/s13613-020-00650-2

Walonoski, J., Klaus, S., Granger, E., Hall, D., Gregorowicz, A., Neyarapally, G., Watson, A., & Eastman, J. (2020). Synthea™ Novel coronavirus (COVID-19) model and synthetic data set. *Intelligence-based medicine, 1*, 100007. [Synthea™ Novel coronavirus (COVID-19) model and synthetic data set - ScienceDirect](https://www.sciencedirect.com/science/article/pii/S2666521220300077?via%3Dihub)

**Zghebi SS, and others. (2020). Development and validation of the Diabetes Severity Score (DISSCO) in 139626 individuals with type 2 diabetes: a retrospective cohort study. *BMJ Open Diabetes Res Care.;8*:e000962**