**Data Analytics Capstone Topic Approval Form**

**Student Name:** Petra Bier

**Student ID:** 011093668

**Capstone Project Name:** Random Forest Regression Model to Determine Hidden Hypoxemia

**Project Topic**: Random Forest model to detect hidden hypoxemia

**This project does not involve human subjects research and is exempt from WGU IRB review.**

**Research Question:** Can a random forest regression (RFR) model predict the difference between a pulse oximeter (SpO2) and arterial oxygen saturation (SaO2) reading using the given data set?

**Hypothesis**:

**Null hypothesis**- A random forest model cannot detect the difference between the SpO2 and the SaO2 with an RMSE of less than 3 percentage points.

**Alternate Hypothesis**- A random forest model can detect the difference between the SpO2 and the SaO2 with an RMSE of less than 3 percentage points.

**Context:**

The contribution of this study to the field of data analytics and the Western Governors University MSDA program is to create a predictive model that can predict the SpO2 – SaO2 gap to determine the accuracy of the SpO2 measurement. SpO2 is a convenient, non-invasive measurement of the blood's oxygen levels. A major benefit of SpO2 reading is that it can be done continuously. Pulse oximetry is only required to be accurate within 2-3% of SaO2 levels and only two-thirds of the time, according to FDA requirements (Wong et al., 2021). The gold standard for blood oxygen saturation is SaO2, which is collected via arterial blood gas (ABG). This is an invasive procedure where blood is drawn from an artery with inherent risks, such as bleeding, infection, and blood clots. Since it is not a continuous measurement, multiple draws are often required. Research has demonstrated that SpO2 does not always correlate accurately with SaO2, especially at lower levels (Sjoding et al., 2023). Hidden hypoxemia, which is when SpO2 is > 88% yet the SaO2 is <88%, may be present and remain undetected if solely relying on SpO2 readings, leading to inaccurate diagnosis and delayed treatment of hypoxemia. A Random Forest regression model will be used to analyze non-linear relationships from the chosen predictor variables to predict SpO2 accuracy in alignment with the FDA error requirements.

**Data:**

The data used for this analysis was sourced from the Physionet BOLD data set (Goldberger et al., 2000; Matos et al., 2023a). The data was accessed after completing the required training and data usage agreement(DUA). While Physionet does have data that can be considered identifiable, this particular data set was de-identified for public usage. The blood-gas and oximetry linked dataset (BOLD) foundSpO2 readings that were aligned with SaO2 readings taken within a five-minute time frame. Variables across the three datasets were standardized, with only those available across all databases being used to reduce the number of null values. A link to the data set can be found here:

<https://physionet.org/content/blood-gas-oximetry/1.0/>

Data Gathering: The data was collected from three Electronic Health Record (EHR) databases (MIMIC-III, MIMIC-IV, eICU-CRD). These are large, publicly available data sets that have retrospectively gathered data from ICUs at large hospitals, and all patient information had been de-identified within each database.

The original data set had 49,093 rows and 142 variables. It contained approximately 1,719,765 NaN values out of 6,971,206 total values (24.67%). After cleaning and feature engineering, the final data set will contain 49,093 rows and 38 columns. The sparsity of the cleaned data frame is 19.15%.

|  |  |  |  |
| --- | --- | --- | --- |
| Variable | Variable Type | Feature Engineering | Variable Type |
| admission\_age | Numeric | No | Independent |
| sex\_female | Categorical | No | Independent |
| race\_ethnicity | Categorical | No | Independent |
| spo2 | Numeric | No | Independent |
| *gap* | *Numeric* | *Yes* | *Dependent* |
| vitals\_heart\_rate | Numeric | No | Independent |
| vitals\_sbp\_ni | Numeric | No | Independent |
| cbc\_hemoglobin | Numeric | No | Independent |
| cbc\_hematocrit | Numeric | No | Independent |
| o2\_carry | Numeric | Yes | Independent |
| cbc\_mch | Numeric | No | Independent |
| cbc\_mchc | Numeric | No | Independent |
| cbc\_mcv | Numeric | No | Independent |
| cbc\_rbc | Numeric | No | Independent |
| cbc\_rdw | Numeric | No | Independent |
| bmp\_sodium | Numeric | No | Independent |
| bmp\_potassium | Numeric | No | Independent |
| bmp\_bicarbonate | Numeric | No | Independent |
| bmp\_bun | Numeric | No | Independent |
| bmp\_creatinine | Numeric | No | Independent |
| bmp\_glucose | Numeric | No | Independent |
| bmp\_anion\_gap | Numeric | No | Independent |
| bun\_cr\_ration | Numeric | Yes | Independent |
| bmp\_calcium | Numeric | No | Independent |
| bmp\_lactate | Numeric | No | Independent |
| hfp\_albumin | Numeric | No | Independent |
| weight\_winz | Numeric | No | Independent |
| height\_winz | Numeric | No | Independent |
| bmi\_winz | Numeric | No | Independent |
| pbw | Numeric | Yes | Independent |
| resp\_winz | Numeric | No | Independent |
| mbp\_winz | Numeric | No | Independent |
| temp\_winz | Numeric | No | Independent |
| wbc\_winz | Numeric | No | Independent |
| alt\_winz | Numeric | No | Independent |
| alp\_winz | Numeric | No | Independent |
| ast\_winz | Numeric | No | Independent |

Limitations: The dataset has two significant class imbalances that will present challenges: Patients with hidden hypoxemia represent only 3.5% of the dataset, and patients with a SpO2-SaO2 gap greater than 3% are only present in 23% of the cases. While the SpO2- SaO2 were collected within a five-minute window, the other lab values were drawn within 24 hours of the SpO2 reading. In an ICU setting, many labs are redrawn within four to six hours so that changes over time can be trended and any changes addressed promptly. This data has values that may not be relevant in relation to the time the SaO2 was drawn, which may reduce the predictive power of some features.

Delimitations: The data will be delimited by removing any values collected from arterial blood collection. Arterial blood draws are often painful and can lead to bleeding problems and complications (Bijapur et al., 2019). Arterial lines for measurement of blood pressure require a specialized skill set, as well as carry an increased risk of infection. The venous lab measurements and non-invasive blood pressure measurements will create a model using data readily available in small hospitals and rural clinics with fewer resources (Harrington et al., 2020). The SaO2 will be used in the project's exploratory data phase but not as a variable for the model.

**Data Gathering:** The data will be downloaded from the Physionet website in a CSV format and uploaded into a Python Jupyter notebook. Many variables will be removed, including all patient stay identifiers, all hospital characteristics, admission data, invasive vital measurements, delta time measurements, and Sequential Organ Failure Assessment (SOFA) scores (U.S. Department of Health and Human Services, 2020). All variables with Nan values greater than 50% will be removed. Feature engineering will include adding the target variable column of the SpO2-SaO2 gap and medical calculations frequently used for patient assessment, including the BUN/ creatine ratio, the oxygen-carrying capacity, and Predicted Body Weight (PBW). The data will be treated for outliers using winzorization. Categorical variables will be one hot encoded with all columns retained if the column contains less than five variables; otherwise, target encoded will be utilized. The Nan values will be imputed using Basian Ridge imputation. Graphs, charts, and heat maps will be presented in an industry-standard format such as but not limited to PowerPoint.

**Data Analytics Tools and Techniques**:

Design of the Study: The technique for this analysis will be to build a Random Forest Regression (RFR) model. This type of analysis is an ensemble machine-learning technique that creates many decision trees and then combines the results. The data was checked for normality using Q-Q plots, although RFR models do not have an assumption of normality and can handle non-linear data well (Sahai, 2023). The model will be split into training and test sets to validate the model at an industry standard of 80/20, respectively (Baheti, 2021). Data leakage was addressed by splitting the data early and performing any transformations separately on the training and test sets. Hyperparameter tuning will be done using an Optuna study to find the best fit for the RFR model (Preferred Networks, Inc., 2024).

**Justification of Tools/Techniques:** The RFR analysis will be run using Python in a Jupyter Notebook in Pycharm. Python was also chosen for its ease of use and the large number of open-source libraries. It is easily scalable for large and small projects, and is strong in general software engineering, allowing analysis tools to be easily integrated into applications and websites (McKinney, 2022). Python is easily accessible, while SAS generally used by large corporations that warehouse data making it difficult to access for individual users and even smaller companies (User, 2022).

**Project Outcomes**: This analysis aims to create a model in which the most important lab values identified can be input into a model along with the SpO2 reading to determine an accurate value of a patient's SaO2 without the need to draw an ABG. This would reduce patient complications while also allowing for an accurate diagnosis of hypoxemia. This type of calculation could be done on a website or application such as MDCalc, a widely used medical reference (<https://www.mdcalc.com/>). Lessening the misdiagnosis of hypoxemia would lead to timely interventions and treatments of patients, in turn lessening complications and improving patient outcomes. The alternative hypothesis that a predictive RFR model can be constructed using the medical data set from the Physionet repository is supported by Matos (2023b).

**Projected Project End Date**: 04/15/2025

**Sources**:

Baheti, P. (2021, September 13). Train Test Validation Split: How To & Best Practices. V7. Retrieved February 1, 2025, from <https://www.v7labs.com/blog/train-validation-test-set>

Bijapur MB, Kudligi NA, Asma S. Central Venous Blood Gas Analysis: An Alternative to Arterial Blood Gas Analysis for pH, PCO2, Bicarbonate, Sodium, Potassium and Chloride in the Intensive Care Unit Patients. Indian J Crit Care Med 2019;23(6):258–262.

Goldberger, A., Amaral, L., Glass, L., Hausdorff, J., Ivanov, P. C., Mark, R., ... & Stanley, H. E. (2000). PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals. Circulation [Online]. 101 (23), pp. e215–e220.

Harrington, R. A., Califf, R. M., Balamurugan, A., Brown, N., Benjamin, R. M., Braund, W. E., Hipp, J., Konig, M., Sanchez, E., & Joynt Maddox, K. E. (2020). Call to Action: Rural Health: A Presidential Advisory From the American Heart Association and American Stroke Association. *Circulation*, *141*(10), e615–e644. Retrieved February 15, 2025, from <https://doi.org/10.1161/CIR.0000000000000753>

Matos, J., Struja, T., Gallifant, J., Nakayama, L. F., Charpignon, M., Liu, X., dos Santos Cardoso, J., Celi, L. A., & Wong, A. K. (2023a). BOLD, a blood-gas and oximetry linked dataset (version 1.0). *PhysioNet*. Retrieved February 1, 2025, from <https://doi.org/10.13026/phvt-3277>.

Matos, J., Struja, T., Restrepo, D. S., Nakayama, L. F., Gallifant, J., Weishaupt, L., Mullangi, N., Loureiro, M., Shapiro, S., Carrel, A., & Celi, L. A. (2023b). MIT Critical Datathon 2023: a MIMIC-IV Derived Dataset for Pulse Oximetry Correction Models (version 1.0.0). *PhysioNet*. Retrieved February 1, 2025, from <https://doi.org/10.13026/jfpc-pz79>.

McKinney, W. (2022). Python for Data Analysis (3rd ed.). O'Reilly.

Preferred Networks, Inc. (2024). Optuna - A hyperparameter optimization framework. Optuna. Retrieved February 15, 2025, from <https://optuna.org/>

Sahai, N. (2023, September 6). *Random Forest Regression – How it Helps in Predictive Analytics?* AnalytixLabs. Retrieved February 15, 2025, from https://www.analytixlabs.co.in/blog/random-forest-regression/

Sjoding, M. W., Iwashyna, T. J., & Valley, T. S. (2023). Change the framework for pulse oximeter regulation to ensure clinicians can give patients the oxygen they need. *American journal of respiratory and critical care medicine*, *207*(6), 661-664.

U.S. Department of Health and Human Services (2020, December 21). *TRACIE SOFA Score: What it is and How to Use it in Triage*. Technical Resources, Assistance Center, and Information Exchange (TRACIE). Retrieved March 10, 2025, from <https://files.asprtracie.hhs.gov/documents/aspr-tracie-sofa-score-fact-sheet.pdf>

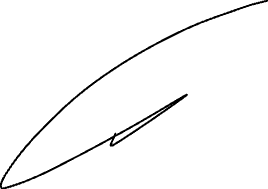
User, S. (2022, April 21). Comparison between SAS, R, and Python. Learn SAS Code. Retrieved February 10, 2025, from <https://learnsascode.com/sas-vs-r-vs-python/>

**Course Instructor Signature/Date:**

The research is exempt from an IRB Review.

An IRB approval is in place (provide proof in appendix B).

Course Instructor’s Approval Status: Approved



Date: 3/17/2025

Reviewed by:

Comments: Click here to enter text.