**Robert Zell, Mario Alcaraz, & Hamza Yousaf**

**BDA 640**

**Dr. Reza Maihami**

**10/16/2023**

**Executive Summary:**

Problem statement:

Montanaro Hospital's 23-bed observation unit (OU), headed by Dr. Erin Kelly, is confronting capacity and utilization challenges. Created to handle short-term stays of under 48 hours, a significant portion of patients from the medicine service exceed this timeframe. A concerning 45% of these prolonged stays end up switching to inpatient status, indicating a potential misjudgment in initial patient placement. The hospital employs an "exclusion list" to identify cases unsuitable for the OU, but its current form isn't effectively filtering patients. Dr. Kelly, recognizing the financial and operational implications, aspires to revamp the list using advanced data analytics.

Methodology we’ve used to address the problem:

Utilizing the CRISP-DM framework, our process began by understanding Montanaro Hospital's challenge with patient placement. During data analysis, we visualized balanced distributions and addressed minor missing values. For identifying patients transitioning from "observation" to "inpatient", logistic regression provided superior accuracy compared to other models. Meanwhile, for predicting the duration of patient stays, the FGLS model was chosen, as it adeptly managed heteroskedasticity issues present in our data. Together, these techniques enabled us to offer actionable insights to refine the hospital's patient placement procedure.

Results of your analysis:

Upon analyzing the performance of multiple classification models for predicting patient "flips" in Montanaro Hospital's observation unit, logistic regression emerged as the most effective tool. It achieved an accuracy of 63.64% and displayed superior distinction abilities with an AUC of 0.6922. While Decision Tree and Random Forest closely followed with accuracies of 62.73%, XG Boost lagged slightly at 60.61%. These findings provide a foundation for informed decision-making, with logistic regression being recommended for optimal patient placement predictions.

Recommendation:

We recommend that Montanaro Hospital prioritizes the integration of the logistic regression model and FGLS model into its patient management system, given its superior accuracy. Continuous monitoring of model performance and data quality is crucial, necessitating monthly reviews. Staff training is paramount to ensure effective utilization of the model's predictions in daily operations. It's also essential to establish a feedback mechanism from frontline staff to continuously refine the model. While relying on data-driven insights, a balance with human clinical judgment ensures comprehensive patient care. Lastly, exploring advanced predictive models will position the hospital at the forefront of healthcare analytics.

**Problem Statement:**

Dr. Erin Kelly supervises the 23-bed observation unit (OU) at Montanaro Hospital, a 260-bed facility situated in a medium-sized U.S. city. After a little over a year of the OU’s operations, Dr. Kelly is concerned about the efficiency of the unit due to two primary factors:

1. The extended average length of stay for patients in the OU.
2. A significant number of patients whose care designation switched from “observation” to “inpatient.”

**Key Problems Identified:**

1. Lengthy Average Stay: The mean duration of stay for patients in the OU, especially for the two-thirds tagged as medicine service patients, exceeds 60 hours. This surpasses the intended short-term observation period, implying inefficiencies, and potential misuse of the OU.
2. Transition from Observation to Inpatient: A considerable percentage (46%) of the medicine service patients initially classified under observation status get transitioned to inpatient status, suggesting possible issues in initial triaging or assessment.
3. Suboptimal Use of the OU: Despite the existence of the OU, a large number (115 weekly on average) of observation-level patients are placed in inpatient beds, either due to the OU's full capacity or because of an exclusion list dictating which preliminary diagnoses shouldn't be housed in the OU.
4. Data Challenges and System Inefficiencies: The endeavor to obtain relevant clinical data for analysis from Montanaro's electronic health records (EHR) became a significant obstacle due to bureaucratic and system inefficiencies.
5. Operational and Economic Impacts: With inpatient wards operating near 100% capacity, using these beds for observation-level patients is a sub-optimal allocation of resources. Furthermore, almost 1,900 patients left the emergency department (ED) without seeing a provider in the past year, representing a potential loss of approximately $1.33 million in revenue.

**Objective:**

The core aim is to refine the operations of the OU by ensuring that patients are accurately and efficiently triaged. By enhancing the precision of the exclusion list through data-driven methods, the hospital aspires to maximize the OU's capacity, improve patient flow, and eventually increase hospital revenue by minimizing the number of patients who leave without being attended to. Furthermore, we intended to arm staff members with immediate access to these predictive analytics via an interactive R Shiny dashboard. This shows staff the predicted output for the likelihood of flipping, expected length of stay, and the current distribution and summary statistics of patients for each day by expected length of stay.

**Potential Benefits:**

1. Increased Hospital Capacity: With a more efficient OU, more patients can be accommodated, improving the throughput.
2. Enhanced Patient Care: By ensuring that patients are placed in the appropriate care setting from the outset, the quality of care can be augmented, reducing chances of medical errors due to unnecessary transfers.
3. Economic Gains: By refining the OU operations, the hospital can potentially realize significant financial gains, both from increased ED attendances and optimized resource allocation.

**Methodology:**

Our methodology, rooted in the CRISP-DM framework, allowed us to move systematically from understanding the business challenge to delivering actionable, data-driven insights. Our selected models, especially logistic regression for classification and FGLS for regression, provided the most robust results, guiding the hospital's decision-making process.This section elucidates the steps taken.

3.1 Business Understanding:

As previously highlighted, our primary goal was to streamline the patient placement process in the observation unit (OU) of Montanaro Hospital. We aimed to use predictive modeling to determine the likelihood of a patient switching from "observation" to "inpatient" status and predict their length of stay.

3.2 Data Understanding:

a) Data Visualization:

Our initial exploration entailed visualizing the distributions of our variables, revealing patterns, anomalies, or potential outliers. For continuous variables, we inspected histograms and box plots. For categorical variables, we assessed frequency distributions to ensure no significant imbalances. The distributions were relatively balanced, and therefore, classification models wouldn't face undue biases.

b) Missing Data & Correlation Analysis:

We identified 20 missing observations and chose to omit them, considering the negligible impact on the overall dataset size. Our correlation analysis showed no high correlations, allowing us to retain all our variables. Furthermore, Principal component analysis was conducted to look for opportunities to reduce the number of features, however the output suggests a lack of opportunity in this regard.

3.3 Data Preparation:

Most of our Data cleaning efforts were focused on recoding variables. We transformed characters into numeric values where necessary and renamed certain factor levels for clarity.

3.4 Modeling:

Two primary modeling tasks were identified:

a) Classification Model (Patient Flipping):

The objective was to determine if a patient would transition from "observation" to "inpatient." We tried various models:

Logistic Regression: Proved to be the most effective, delivering superior performance with respect to accuracy and area under the curve (AUC).

Random Forest, Classification Trees, and XGBoost: These were evaluated but did not outperform logistic regression in terms of accuracy or AUC.

b) Regression Model (Predict Length of Stay):

We embarked on predicting the length of stay of patients:

Linear Regression: Our baseline models.

FGLS (Feasible Generalized Least Squares): Due to detected heteroskedasticity in the residuals, we leveraged the FGLS approach, which effectively weighted the residuals, enhancing the model's performance. This model boasted the lowest Mean Absolute Error of 55.91, and highest R-squared value across all models of 0.37.

Super Learner: This stacked ensemble method was evaluated and came out to have the lowest RMSE of the models tested at 74.66 but lacked effectiveness as FGLS better compensated for unequal variation in the error term across observations. This resulted in lower overall explanatory power, but also less interpretability. Furthermore, the model itself was less amenable to implementation due to longer prediction times and more complex refinement processes.

3.5 Performance Criteria:

Our model performance was gauged through multiple metrics:

a) Sensitivity vs. Specificity: Ensuring a balance between these was crucial. Misclassifying someone who requires intensive care or overloading the inpatient units would both be detrimental. Hence, both sensitivity and specificity were crucial in evaluating our models.

b) Accuracy: This metric primarily guided our model selection process for the classification task.

c) AUC (Area Under the Curve): It provides a robust measure for classification model performance, indicating its ability to distinguish between the classes effectively.

d) R-squared: Used for our regression models to understand the proportion of variance in the dependent variable that's predictable from the independent variables.

e.) Mean Absolute Error (MAE): This measures the sum of errors in absolute values, and is a common performance metric for regression models. Our FGLS model performed best by this measure.

f.) Root Mean Square Error (RMSE): Measures the average distance between predicted and actual values as the square root of the average error. SuperLearner performed best by this metric.

**Results**

Upon evaluating multiple classification models, the results underscored logistic regression as the most potent tool for our specific challenge. Here's a breakdown of the findings:

Logistic Regression: This model topped the charts, showcasing an accuracy of 63.64%. Its Area Under Curve (AUC) of 0.6922 reflects its reliable performance, indicating a good distinction between the predicted classes. With a sensitivity of 61.43%, it predicts a majority of the "flips" correctly, while its specificity of 65.26% means it can also correctly identify a notable percentage of those who don't "flip".

Decision Tree: Achieving an accuracy of 62.73%, this model was slightly inferior to logistic regression. Its specificity was the highest among all models at 68.95%, signifying its strong capability to correctly classify non-flippers, but it compromised on sensitivity at 54.29%.

Random Forest: Mirroring the Decision Tree's accuracy, Random Forest too presented a 62.73% accuracy. Its balanced sensitivity and specificity percentages (61.43% and 63.68% respectively) show a moderate performance in both true positive and true negative predictions.

XG Boost: This model had the least accuracy at 60.61%. Though its sensitivity of 63.57% was commendable, its lowest specificity of 58.42% highlighted its challenge in classifying the true negatives accurately.

Assumptions:

Several assumptions were made during the modeling process:  
Independence of Observations: We presumed that each patient's data is independent of the others.  
Homoscedasticity: Especially relevant for linear-based models, we assumed constant variance of the residuals.  
Absence of Multicollinearity: We believed that no two independent variables were highly correlated. Our prior correlation analysis validated this.  
Linearity: For logistic regression, we assumed a linear relationship between the log odds of the outcome and predictor variables.  
No Information Bias: We operated on the premise that the collected data was devoid of any biases and truly represented the patient demographics and behaviors.

**Recommendations & Course of Action:**

1. Adopt Logistic Regression for Immediate Use:  
   Action: Integrate the logistic regression model into the hospital's patient management system. This model has proven to be the most accurate and should be the first line of prediction.  
   Consequence: This will enhance the accuracy of patient "flip" predictions and streamline patient placement, potentially reducing costs and improving patient care.
2. Continuous Data Monitoring & Model Refinement:  
   Action: Establish a dedicated team or task an existing one to monitor data quality and model performance. Monthly reviews should be done to ensure the model's predictive accuracy remains high.

Consequence: The model stays relevant and evolves with the changing dynamics of patient behavior, maintaining its utility and efficacy.

1. Staff Training & Implementation:  
   Action: Organize regular training sessions for staff to familiarize them with the model's predictions, its implications, and its use in decision-making.  
   Consequence: This ensures the model's insights are used efficiently and consistently across the board, maximizing the benefit derived from the predictive analytics.
2. Establish a Feedback Mechanism:  
   Action: Implement a system where hospital staff, especially those in the observation unit, can provide feedback on the model's predictions versus actual outcomes.  
   Consequence: Over time, this real-world feedback will be instrumental in refining the model, ensuring it remains relevant to on-ground realities.
3. Human Oversight & Decision-making:  
   Action: While the model provides valuable insights, it should be used as a decision-support tool. Important patient placement decisions should involve both model recommendations and clinical expertise.  
   Consequence: This balances data-driven insights with human judgment, ensuring comprehensive and safe patient care.
4. Exploration of Advanced Models:  
   Action: Dedicate a small team or external consultants to explore more advanced algorithms or techniques that could potentially enhance prediction accuracy further.  
   Consequence: The hospital stays at the forefront of technological advancements in predictive analytics, ensuring the best tools are being utilized.

**APPENDIX**

**Basic Statistics**

A screenshot of a computer

Description automatically generated

**Univariate Distributions**

A screenshot of a graph

Description automatically generated

**Bar Charts (w/ Frequency)**

A screenshot of a computer

Description automatically generated

**Correlation Analysis**

A screenshot of a computer screen

Description automatically generated

**Principal Component Analysis**

A graph with numbers and text

Description automatically generated with medium confidence

**Predictive Dashboard: User Interface**

A screenshot of a graph

Description automatically generated

**Classification Model Performance**

A screenshot of a computer

Description automatically generated

**Regression Model: R-Squared**

**Logistic Regression Model Output**

A screenshot of a computer screen

Description automatically generated

**Decision Tree Model**

A diagram of a number of numbers

Description automatically generated

**ROC Curves**

A graph of a graph

Description automatically generated

**FGLS Model Output**

A screenshot of a computer

Description automatically generated

**T-SNE Visualized K-Means Clustering**

A group of blue and orange dots

Description automatically generated

**Clustering Statistics**

A screenshot of a computer

Description automatically generated