Raag Patel / DS 7333 – Quantifying the World / Fall 2023 / Case Study 2

Logistic Classification on Hospital Readmittance Rates

Data Introduction:

This study introduces a dataset comprising information of 101,766 patient records across 48 attributes, aimed at unraveling the underlying factors influencing hospital readmittance. These attributes span patient demographics such as race, age, and gender, as well as more medical characteristics like time spent in the hospital, number and type of diagnoses, laboratory procedures undertaken, medications administered, and more.

The dataset provides insights into various aspects of a patient's hospital journey, including the type and source of their admission, the specialty of the attending medical practitioner, and their discharge disposition. Crucially, the data also captures specifics about diabetes medications, shedding light on potential relationships between medication regimes and readmittance rates. The ultimate objective of this exploration is to employ a logistic regression classifier to predict the likelihood of a patient's readmittance based on these diverse features. This has the potential to inform clinical decision-making as well as driving administrative strategies to optimize patient care and resource allocation.

Methods:

Null Values: Missing data poses significant challenges in data analysis. It can bias outcomes, reduce the statistical power of a study, and undermine the generalizability of results. The choice of imputation method should consider the nature of the data, the amount of missingness, and the context of the research.

Logistic Regression: Logistic regression is a statistical model utilized primarily for predicting binary outcomes by modeling the relationship between one or multiple independent variables and a binary dependent variable.

Model Evaluation: Accuracy, precision, recall, and the F1 score are metrics in classification problems, each offering a distinct perspective on model performance. Accuracy measures the proportion of correctly predicted classifications in the total predictions made, serving as a

general indicator of the model's correctness. Precision gauges the accuracy of positive predictions, specifically quantifying how many of the predicted positive cases were indeed positive. Recall, or sensitivity, assesses the model's capacity to identify all positive cases, demonstrating its efficacy in capturing true positives amidst the actual positive cases. F1 score encapsulates a balance between precision and recall in a single value, dividing the two.

ROC Curve: The Receiver Operating Characteristic (ROC) curve is a graphical representation that showcases the diagnostic capability of a binary classification system as its discrimination threshold varies. By plotting the True Positive Rate (TPR) against the False Positive Rate (FPR) at different thresholds, the ROC curve provides a comprehensive view of a model's performance across varying classification thresholds. The area under the curve (AUC) provides a singular measure, with a value of 1 indicating perfect prediction and 0.5 suggesting no better accuracy than random guessing.

Feature Cleaning:

Feature	Percentage
Race	2.23%
Weight	96.86%
Payer_Code	39.56%
Medical_Specialty	49.08%
Diag_1	.02%
Diag_2	.35%
Diag_3	1.40%
Max_Glu_Serum	94.75%
A1Cresult	83.28%

This table shows the amount of missing values in columns who had any missing values

The following decisions were made for each variable with missing data in the diabetes dataset:

1. Race:

Decision: Imputed with mode.

The missingness for race was relatively low (2.23%). Mode imputation was chosen due to its simplicity for categorical variables with minimal missing values. This method introduces minimal bias, especially when missingness is low.

2. Weight:

Decision: Column dropped.

With 96.86% missing values, it's impractical and potentially misleading to impute such a large proportion. Given the magnitude of the missing data, any imputation would be a major source of introduced bias.

3. Payer_code:

Decision: Imputed with 'UNKNOWN'.

The missing percentage (39.56%) is substantial. Rather than imputing with a mode, which might be misleading, labeling it as 'UNKNOWN' serves as a placeholder that signifies missing data

4. Medical_specialty:

Decision: Imputed with 'UNKNOWN'.

Given the high missing rate (49.08%), a placeholder of 'UNKNOWN' was deemed appropriate. This offers transparency, indicating the lack of information rather than introducing potential bias by imputing with the mode or another method.

5. Diag_1, Diag_2, Diag_3:

Decision: Imputed with 0.

As diagnostic codes, it's crucial to maintain the integrity of these columns. The missing percentages were low (0.02%, 0.35%, and 1.40%, respectively), allowing us to inpute 0, so that

we can maintain integrity, while removing missing values. 0 is not a registered code, and does not show up in any of the columns.

6. Max_glu_serum & A1Cresult:

Decision: Columns dropped.

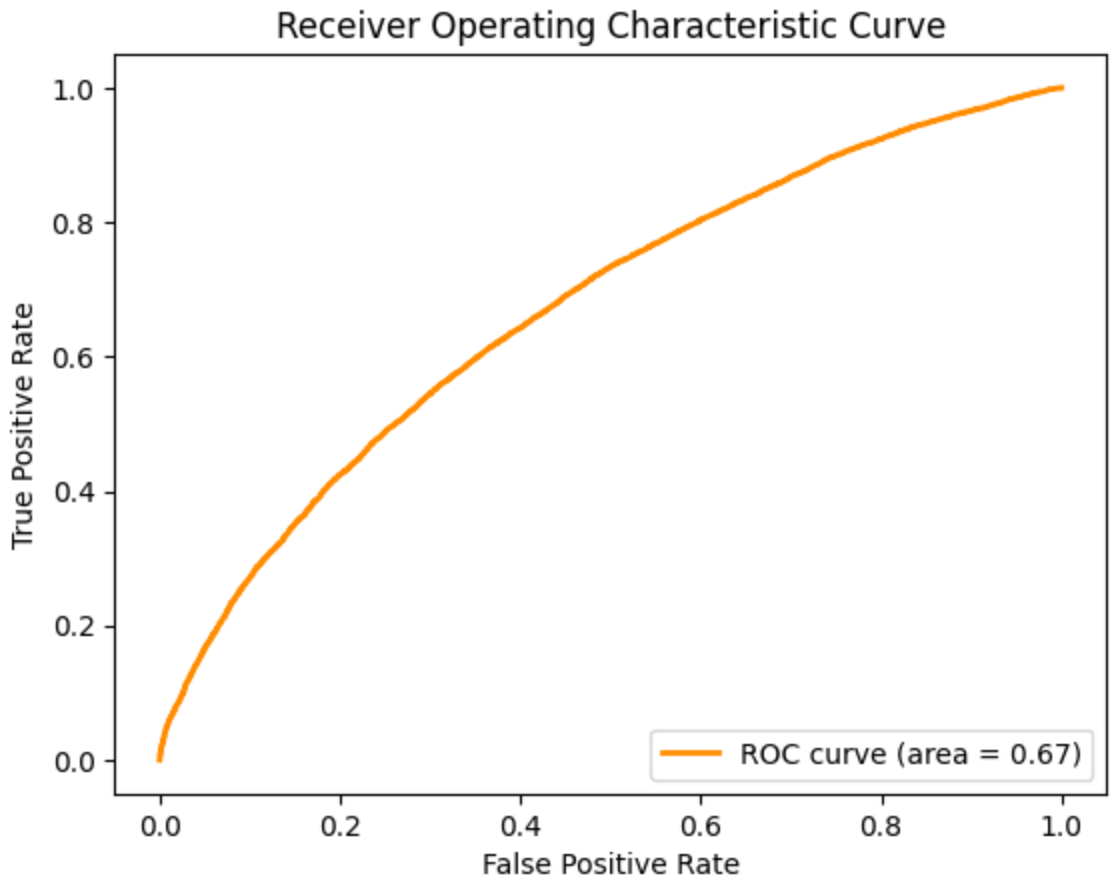
Both columns had exceedingly high missing values (94.75% and 83.28%, respectively). Given their high degree of missingness, imputation would not be representative and would introduce undue bias. Therefore, the columns were removed.

Mathematical models, including logistic regression, are fundamentally designed to interpret numerical inputs. Consequently, incorporating categorical variables requires a transformation. A commonly adopted technique is "one-hot encoding". This transformation ensures that our models can seamlessly incorporate and process the categorical data, ultimately enhancing the accuracy and reliability of the predictions.

Simple Model:

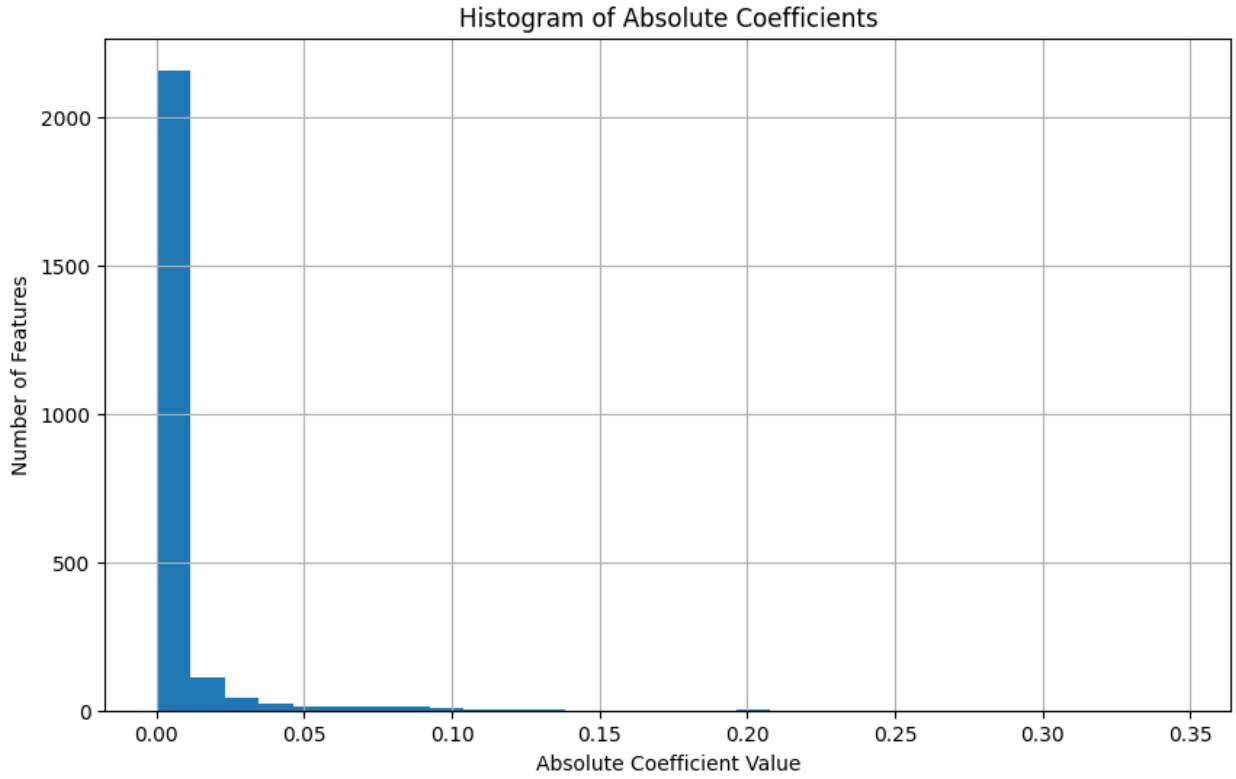
In our study, we employed logistic regression to predict hospital readmittance based on a set of patient attributes. The objective was to discern patterns in patient data that are indicative of a higher likelihood of readmittance. By fitting the model on historical patient data, we were able to derive coefficients for each attribute which, in turn, allowed us to estimate the probability of readmittance for new patients. The outcomes were binary, indicating either a readmission ("Yes") or no readmission ("No"), and the model provided insights into which factors were most predictive of a return to the hospital.

The first model used all the features. Evaluating the model's predictive performance, we found an accuracy of approximately 62.62%, suggesting that it correctly predicts readmittance in just over 6 out of 10 cases. The model's precision, standing at 63.93%, indicates that when the model predicts a readmittance, it is correct in nearly 64% of the instances. However, the recall value of 43.28% suggests that the model identifies only about 43% of the actual readmittance cases. The F1-Score, which harmonizes precision and recall, was determined to be 51.62%. These metrics offer a promising start, feature selection may enhance the model's accuracy in future iterations.



Observing the ROC curve that is slightly pulled towards the y-axis but remains closer to the diagonal suggests a model that performs better than random guessing, but not considerably so. With an ROC AUC of 0.67, the model does demonstrate some capability in distinguishing between the positive and negative classes. However, there remains substantial room for improvement. The pull towards the y-axis signifies that at some thresholds, the model has a decent sensitivity without escalating the false positives drastically. Yet, the proximity to the diagonal implies that, overall, the separation between the classes is not distinct.

Feature Selection:

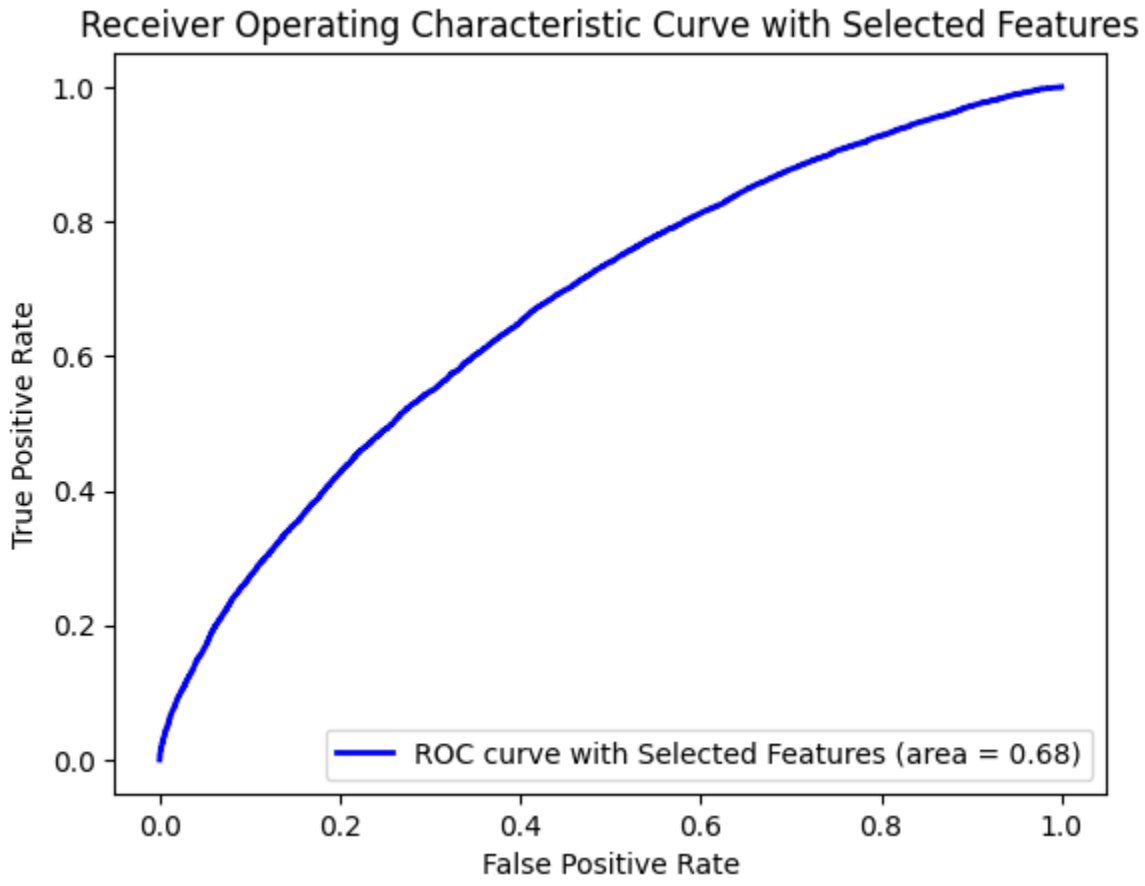


Due to “one-hot encoding’, categorical variables with multiple categories exacerbated the size of the data, leading us to about 2500 features. The above plot shows the distribution of the absolute value of the coefficients, differed from the simple model, to showcase it’s strengths.

The absolute coefficients spanned a range from an absolute minimum of 0 to a maximum of 0.3464, with a mean value of approximately 0.0065 and a standard deviation of 0.0211. Notably, the median absolute coefficient sat at a modest 0.001146, suggesting that half of the features exert a relatively low influence on the model's predictions. The features from the top quartile, specifically those with an absolute coefficient surpassing the 0.0035 mark, represent the 75th percentile of the distribution, which is used as our cut off metric for the final model. By capitalizing on the most impactful features, we can maximize the model's accuracy, while concurrently minimizing potential distractions and overfitting risks presented by less influential variables.

Final Model:

By focusing solely on the most influential features derived from the top quartile of absolute coefficients, a marked improvement in predictive metrics was observed. The accuracy, from the simple model's 0.6262 increased to 0.6306 with the refined feature set. Additionally, the precision, indicating the fraction of positive identifications that were actually correct, received a slight enhancement from 0.6392 to 0.6362. More significantly, the recall, which denotes the fraction of actual positives our model successfully identified, surged from 0.4328 to 0.4628, capturing a broader subset of patients at risk of readmittance. Furthermore, the F1-Score, a mean of precision and recall, advanced from 0.5162 to 0.5358, underscoring a balanced improvement in both these metrics.



In relation to the Receiver Operating Characteristic (ROC) curve, the model achieved an Area Under the Curve (AUC) of 0.68. The ROC curve, in essence, provides a visualization of the trade-off between the model's true positive rate and false positive rate. An AUC value of 0.68

suggests that the model possesses a moderate ability to discern between those patients who will be readmitted and those who won't. Our model demonstrates a level of discernment that is significantly better than random guessing, though with room for enhancement.

Conclusion:

The initial logistic regression model was developed with a broad spectrum of features to predict hospital readmittance. While this simple model offered insights, refining it by focusing on the top 25% of influential features, as gauged by their absolute coefficient values, led to an enhanced model performance. The refined model not only demonstrated improved accuracy, precision, and recall metrics but also was leaner and potentially more interpretable. However, it's important to address certain limitations in our approach. Our dataset lacked certain critical features such as 'A1Cresult' and 'weight'. The inclusion of these attributes could potentially elevate the model's performance, as they play pivotal roles in health scenarios. Specifically, A1C results, providing insights into a patient's average blood sugar levels over time, and weight, a fundamental health metric, can be instrumental in predicting hospital readmittance. In future iterations, it would be beneficial to integrate these vital indicators to achieve a more comprehensive and accurate predictive model.