No-Show to Appointment

Yulimar Rivero

Adelphi University

DSC 789: Strategic Capstone Project

Zahra Sadeghi Maman

May 16, 2023

INTRODUCTION

The no-show appointment problem relates to the serious issue that comes with patients either choosing to bail on an appointment or failing to remember that they had scheduled an appointment. This problem has led to tremendous economic and efficiency issues in the healthcare sector.

The approach to addressing this problem started with the Business understanding aspect. After analyzing previous analytical works, we discovered that there is a positive correlation between consultation length and patient no-shows. Other works concluded that age, gender, appointment type, and number of appointments in the previous year were significant contributors to determining the status of appointment no-show. The final work used simulation models to analyze the effects of different appointment policies. While doing this they found out that policies that allowed for overbooking and double-booking could significantly reduce the negative impacts on the healthcare industry.

After analyzing previous analytical works, we moved onto checking our dataset. Our dataset is comprised of 16 different variables. These variables include age, gender, scheduled day, appointment day, month, calling time, waiting time, financial aid, hypertension, diabetes, alcoholism, handicap, appointment reminder, time between appointments, prior no-show, and show-up. We will dive into how we changed each variable to better serve our dataset but, these methods included factoring, labeling, and other techniques.

Some research questions that we included were variables that held the most weight in determining whether a patient would show up or not. Other questions that arose while analyzing our dataset included which balancing, feature selection, and models would produce the best

performance metrics. For model building, we used G-mean to analyze performance due to the fact that it accurately defines how well the model is able to make predictions on the testing set.

METHOD

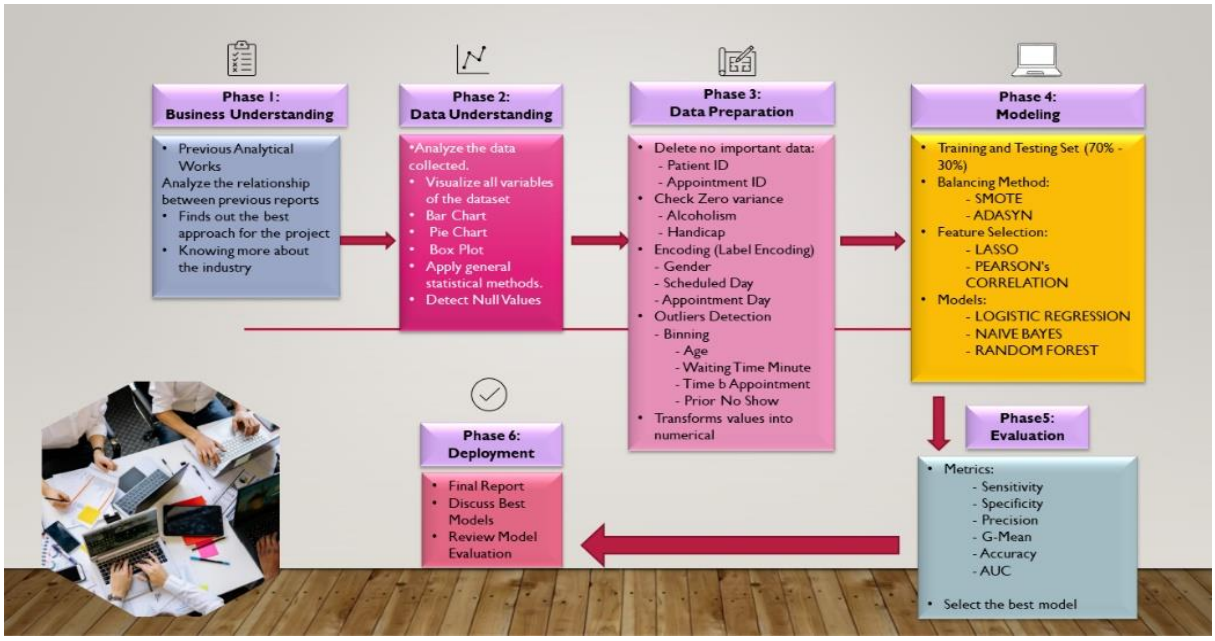
During the data understanding phase, we utilized Tableau to visualize all variables in the dataset, which provided valuable insights into the characteristics and distributions of each variable. By observing the statistical reports generated for every variable, we were able to gain a thorough understanding of their basic properties. This critical step allowed us to develop a clear idea of the necessary actions to be taken in the subsequent stages of the data processing pipeline. The visualization and exploration of the dataset through Tableau ensured that our decisions in the following phases, such as data preparation, feature selection, and model building, were grounded in a solid understanding of the underlying data structure and characteristics.

The data preparation process involved several steps to ensure a clean and reliable dataset for further analysis. First, the dataset "noshow.csv" was imported, and its structure was inspected, including checking for missing values. Unimportant columns "PatientId" and "AppointmentID" were removed as they were not significant for the analysis. Zero-variance variables were identified as "Alcoholism" and "Handicap" which indicated that they did not contribute significantly to the model. Categorical variables (Gender, ScheduledDay, and AppointmentDay) were encoded by using label encoding. Outliers were detected in variables such as "Age", "Calling_time..hour.in.a.day.", "Waiting_time..minute.", "Time_b_appointment..day.", and "Prior_noshow" using boxplots and were treated through binning method.

Finally, the structure of the cleaned and preprocessed dataset was re-inspected, missing values were checked again, and the data types of the variables were converted to numeric before proceeding to the next stages of the project, such as data balancing, feature selection, and model training. This comprehensive data preparation process ensured that the dataset was of high quality for building accurate and reliable models.

In the modeling phase of our project, we employed balancing techniques, such as Synthetic Minority Over-sampling Technique (SMOTE) and Adaptive Synthetic (ADASYN), to address the imbalanced nature of our dataset. This step ensured a more reliable and robust modeling process. For feature selection, we utilized both Lasso regression and Pearson correlation to identify the most relevant and significant variables contributing to the show-up prediction.

Various classification algorithms were tested, including Logistic Regression, Naïve Bayes, and Random Forest, to build our predictive models. By comparing their performance using metrics such as G-Mean, it aimed to identify the most suitable model for predicting the show-up outcome in the given dataset. The combination of these techniques and models led to a comprehensive understanding of the data and the creation of an effective predictive model for the show-up output.



Methodology Framework

RESULTS

This section presents the detailed interpretation of findings from the results tables and visualization graphs. The evaluation metrics used to assess model performance include sensitivity, specificity, precision, G-Mean, accuracy, and AUC.

Scenario	Model	Encoding approach	Imputation	Outlier handling method	Balancing approach	Feature selection	Sensitivity	Specificity	Precision	Gmean	Accuracy	AUC
1	LR	Label Encoding	NA	Binning	SMOTE	Lasso	0.647	0.798	0.457	0.719	0.767	0.807
2	NB	Label Encoding	NA	Binning	SMOTE	Lasso	0.726	0.613	0.329	0.667	0.636	0.669
3	RF	Label Encoding	NA	Binning	SMOTE	Lasso	0.528	0.894	0.566	0.687	0.818	0.711
4	LR	Label Encoding	NA	Binning	SMOTE	Pearson	0.615	0.795	0.440	0.670	0.758	0.788
5	NB	Label Encoding	NA	Binning	SMOTE	Pearson	0.715	0.628	0.334	0.670	0.646	0.671
6	RF	Label Encoding	NA	Binning	SMOTE	Pearson	0.507	0.900	0.572	0.676	0.819	0.704
7	LR	Label Encoding	NA	Binning	ADASYN	Pearson	0.654	0.773	0.431	0.711	0.748	0.787
8	NB	Label Encoding	NA	Binning	ADASYN	Pearson	0.738	0.592	0.322	0.661	0.622	0.665
9	RF	Label Encoding	NA	Binning	ADASYN	Pearson	0.531	0.883	0.544	0.685	0.810	0.707

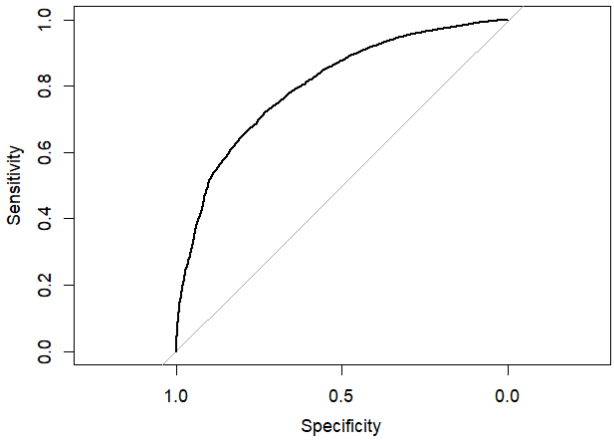
A thorough analysis of the performance metrics for each scenario revealed that Scenario 1 stands out as the most effective approach. In this scenario, the train data was balanced using SMOTE, feature selection was performed using LASSO, and a Logistic Regression model was employed for analysis. The G-Mean value of 0.719 demonstrated a well-balanced model that effectively addresses sensitivity and specificity.

Sensitivity (also known as recall or true positive rate) measures the proportion of true positives correctly identified by the model. In Scenario 1, the sensitivity was 0.647, indicating that 64.7% of actual positive cases were correctly identified. Specificity, on the other hand, measures the proportion of true negatives correctly identified. With a specificity of 0.798, the model in Scenario 1 correctly identified 79.8% of actual negative cases.

Precision, another important metric, assesses the proportion of true positive cases among those predicted as positive by the model. In Scenario 1, the precision value was 0.457, suggesting that 45.7% of cases predicted as positive were indeed positive. Accuracy, a commonly used metric, measures the overall proportion of correct predictions. The accuracy in Scenario 1 was 0.767, which implies that 76.7% of all predictions made by the model were correct.

Lastly, the AUC (Area Under the Receiver Operating Characteristic Curve) is a comprehensive metric that evaluates the overall performance of the model across all classification thresholds. The AUC value in Scenario 1 was 0.807, which is considered good, further supporting the effectiveness of this approach.

AUC Scenario 1

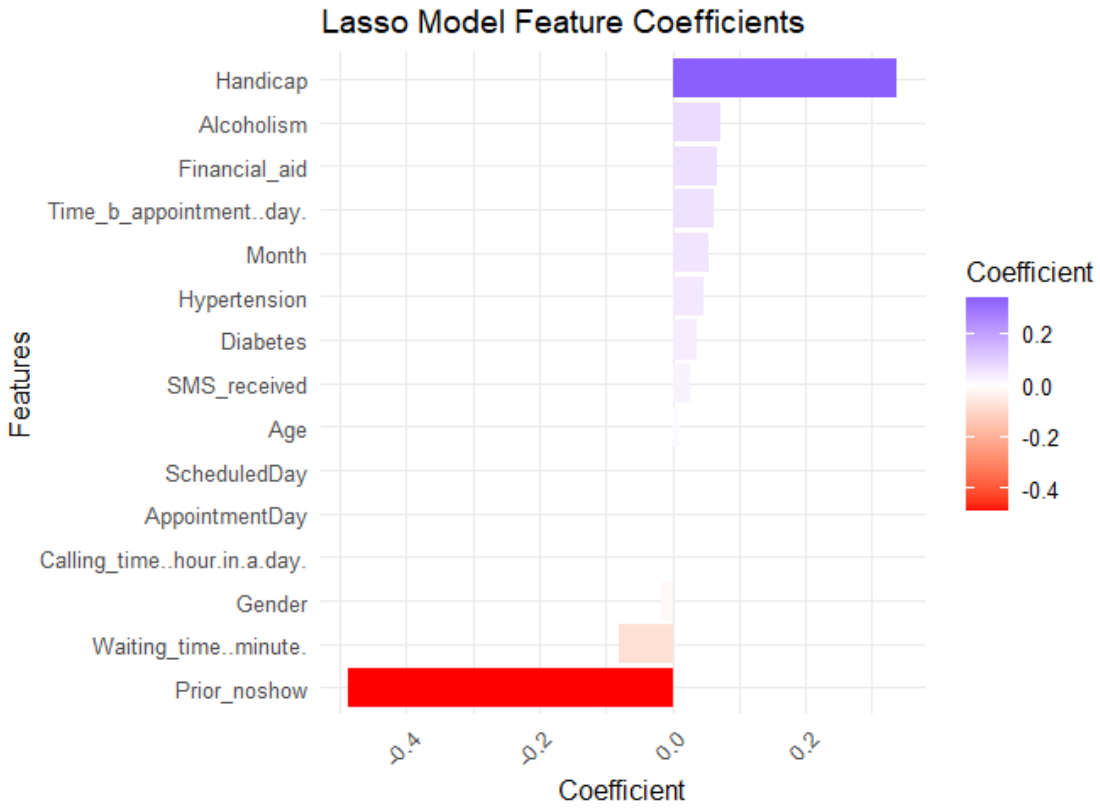


All these metrics led to Scenario 1 providing the best results among the tested scenarios, as evidenced by the highest G-Mean value of 0.719. The analysis of other evaluation metrics also supported the robustness of this approach, which combines SMOTE for data balancing, LASSO for feature selection, and Logistic Regression as the analytical model. This combination ensured a well-balanced model that effectively handles both sensitivity and specificity, making it suitable for addressing the challenges posed by imbalanced datasets.

Having selected the best scenario, the following evaluation was to determine which features were the most important for the model. Based on the results provided, we can assess the importance of each feature in the model from Scenario 1, which uses LASSO for feature selection. The coefficients given by LASSO represent the strength of the relationship between each feature and the target variable, with larger absolute values indicating a stronger relationship. Below is a list of features along with their respective coefficients:

Lasso Features Selection - Scenario 1

Feature	Coefficients
Age	0.011367586
Gender	-0.015367918
ScheduledDay	0.001028447
AppointmentDay	0.000657967
Month	0.055739646
Calling_time..hour.in.a.day.	0.000322619
Waiting_time..minute.	-0.079499915
Financial_aid	0.066590356
Hypertension	0.046239667
Diabetes	0.03796823
Alcoholism	0.072562827
Handicap	0.338156804
SMS_received	0.026935681
Time_b_appointment..day.	0.062889999
Prior_noshow	-0.487907185



The most important feature in the model, as indicated by the highest absolute coefficient value, was 'Prior_noshow' (-0.4879071846). This negative coefficient suggests that a prior history of not showing up for appointments is strongly associated with a higher likelihood of not attending future appointments.

The second most important feature is 'Handicap' (0.3381568039), with a positive coefficient, implying that the presence of a handicap is associated with an increased probability of attending an appointment.

Other features, such as 'Waiting_time..minute.' (-0.0794999147), 'Alcoholism' (0.0725628274), and 'Financial_aid' (0.0665903564), also demonstrated notable relationships with the target variable. However, their impact on the model's performance was not as substantial as 'Prior_noshow' and 'Handicap'.

Another important approach to analyze in Scenario 1 was to consider the presence of outliers in the model and observed the impact of them in the performance of the model. Here is a comparison of the metrics with and without outliers:

Metric with and without Outliers Scenario 1

Metrics	Outliers	No Outliers
Sensitivity	0.543	0.647
Specificity	0.854	0.798
Precision	0.501	0.457
G-Mean	0.681	0.719
Accuracy	0.788	0.767
AUC	0.787	0.807

From the results, it can be observed that the presence of outliers significantly impacts the model's performance. When comparing the model with outliers to the one without outliers, the sensitivity improved from 0.543 to 0.647, indicating that the model without outliers is better at identifying true positives.

The specificity, however, decreased from 0.854 to 0.798 when outliers are removed. This suggested that the model with outliers performs slightly better in identifying true negatives, but the improvement in sensitivity without outliers outweighs this advantage. The G-Mean, which measures the balance between sensitivity and specificity, increases from 0.681 with outliers to 0.719 without outliers. This indicated that the model without outliers provides a better balance between the two metrics, making it more suitable for handling imbalanced datasets. Accuracy improved slightly without outliers, from 0.788 to 0.767. Although the difference is not substantial, it is still indicative of better overall performance for the model without outliers. The AUC value increased from 0.787 with outliers to 0.807 without outliers. This improvement in AUC further supported the notion that the model without outliers offers better overall performance across all classification thresholds.

The analysis of the impact of outliers on the performance of the model in Scenario 1 demonstrated that removing outliers leads to a more balanced and better-performing model. While some metrics, such as specificity, might be slightly affected, the improvements in sensitivity, G-Mean, accuracy, and AUC ultimately made the model without outliers more suitable for addressing the challenges posed by imbalanced datasets.

Moving onto the impact of having different feature selection methods in the performance of the 9 scenarios, the results suggested that the choice of feature selection method does have an impact on the performance of the models. The models using LASSO feature selection (Scenarios 1-3) tend to have higher G-Mean and AUC values than their Pearson's Correlation counterparts (Scenarios 4-6). This indicated that LASSO is generally more effective at predicting the response variable across the three analytical models.

DISCUSSION AND CONCLUSION

In general, Scenario 1 stands out as the best-performing scenario among the nine tested, achieving the highest G-Mean value of 0.719. This scenario combines SMOTE for data balancing, LASSO for feature selection, and Logistic Regression as the analytical model. This well-balanced approach effectively managed both sensitivity and specificity, making it highly suitable for handling the challenges of imbalanced datasets.

The feature importance analysis for Scenario 1 identified 'Prior_noshow' and 'Handicap' as the two most influential features. While other features such as 'Waiting_time..minute.', 'Alcoholism', and 'Financial_aid' also exhibit relationships with the target variable, but their impact was not as significant.

The presence of outliers was another factor considered in the evaluation of Scenario 1. The comparison of models with and without outliers reveals that excluding outliers leads to a better-performing model. Although specificity is slightly lower, improvements in sensitivity, G-Mean, accuracy, and AUC demonstrated that the model without outliers is more suitable for addressing imbalanced datasets.

Overall, by following the same strategy this study took, it can be implemented a data-driven decision-making process that leverages the most effective feature selection method (LASSO) and analytical model (Logistic Regression) to accurately predict the response variable ‘Show Up’, and will enable informed decisions that address the challenges posed by imbalanced datasets and contribute to better outcomes. However, there are some suggestions that can be taken for future improvements and research directions divided into the following categories:

- Further Feature Engineering: Since 'Prior_noshow' and 'Handicap' had the strongest impact on the model, it may be beneficial to explore additional feature engineering techniques or

transformations that could better capture the relationship between these features and the target variable. This could potentially lead to even better model performance.

- Outlier Treatment: The results indicated that outlier removal leads to improved model performance in terms of sensitivity, G-Mean, accuracy, and AUC. Therefore, it is essential to develop a robust outlier detection and treatment strategy in the data preprocessing stage to enhance the performance of future models.
- Alternate Data Balancing Techniques: While the SMOTE technique has shown promising results in this analysis, it may be worth exploring alternative data balancing techniques such as cluster-based oversampling to evaluate their impact on model performance.
- Hyperparameter Tuning: The logistic regression model in Scenario 1 can potentially be further improved by tuning its hyperparameters, such as the regularization strength and type, using techniques like grid search or random search.
- Advanced Models: Consider exploring more advanced machine learning models such as gradient boosting machines (GBMs) or neural networks to determine if they can offer better performance for this problem.

By considering these suggestions, future research and improvements in this area can be directed toward developing more accurate and robust models for predicting appointment attendance, ultimately leading to better decision-making and resource allocation.