STAT 482 Final Project Report: Medical Appointment No-Show Prediction

Anoushka Jadhav

Spring 2025

1 Introduction

Missed medical appointments create inefficiencies in healthcare systems, leading to wasted resources, increased wait times, and potential negative health outcomes for patients. Hospitals and clinics often struggle with reducing no-show rates, as missed appointments disrupt scheduling and can result in financial losses. The goal of this study is to develop a predictive model to determine whether a patient will attend their scheduled medical appointment based on demographic, medical, and scheduling related factors. By identifying key variables that influence appointment attendance, healthcare providers can implement targeted interventions such as reminders, scheduling adjustments, or outreach programs to reduce no-show rates.

The dataset used in this study consists of 110,527 medical appointments and includes 14 variables related to patient characteristics, medical conditions, and scheduling details. The key question being addressed is whether machine learning models can effectively predict no-shows and identify the most important factors contributing to missed appointments. The target variable, "No-show," indicates whether a patient attended their appointment. Other variables include demographic details such as age, gender, and neighborhood, as well as medical history features such as hypertension, diabetes, and alcoholism. Additionally, the dataset includes the date the appointment was scheduled, the actual appointment date, and whether the patient received an SMS reminder.

Understanding the factors influencing no-shows is critical for improving healthcare efficiency. Studies estimate that missed medical appointments cost the U.S. healthcare system over 150 billion dollars annually (Chen, 2023). By leveraging machine learning techniques, hospitals can proactively predict patients at risk of missing appointments and take preventive measures. The findings from this study could help optimize scheduling systems, reduce financial losses, and ultimately enhance patient adherence to medical care.

2 Research Strategy

2.1 Data

The dataset used in this study comes from the Brazilian Public Health Service and contains medical appointment records from the city of Vitória, Brazil, for the years 2015-2016. It has been sourced from Kaggle, a well-known platform for open-source datasets. The dataset consists of 110,527 appointment records from multiple healthcare facilities in Vitória and includes information on patient demographics, medical conditions, and appointment details.

Each record includes key features such as age, gender, neighborhood, scholarship enrollment (Bolsa Família program status), hypertension, diabetes, alcoholism, disability status, and whether an SMS reminder was received. The dataset also provides temporal information, including the date the appointment was scheduled and the actual appointment date.

The target variable in the dataset is "No-show," a binary variable indicating whether a patient attended ("No") or missed ("Yes") their appointment. Given that factors such as location (neighborhood), medical conditions, and appointment scheduling times may influence no-shows, exploratory data analysis (EDA) will be conducted to assess relationships between variables.

Since the dataset contains both categorical and numerical variables, preprocessing steps will involve encoding categorical data, handling missing values if present, and creating new features such as waiting time (the difference between scheduling date and appointment date), appointment day of week information, and seasonality information. These engineered features may provide additional insights into patient behavior and no-show trends. Additionally, class imbalance in the "No-show" variable will be examined, as an uneven distribution of show vs. no-show cases can impact model performance.

2.2 Planned analysis

This study will apply decision trees, random forests, and XGBoost models to predict whether a patient will miss their appointment. The analysis will begin with exploratory data analysis (EDA) to visualize trends, assess variable distributions, and identify patterns related to appointment no-shows. Key visualizations will include histograms, box plots, bar graphs, and heat maps to examine the relationship between patient characteristics and attendance rates.

Feature engineering will be performed to improve model performance. New features such as waiting time between scheduling and the appointment, appointment day of week, seasonality, and interactions between medical conditions and demographic factors will be created. Categorical variables such as gender and neighborhood will be one hot encoded, and numerical variables will be standardized where necessary.

The machine learning models to be used include a decision tree classifier, a random forest classifier, and an XGBoost model. The decision tree model will provide an interpretable framework for understanding how different features impact no-show rates, while the random forest and XGBoost models will enhance prediction accuracy. XGBoost, an advanced gradient boosting algorithm, will be particularly effective at handling any imbalance in the dataset and identifying non-linear relationships. Regularization techniques will also be used when fitting these models to ensure accuracy and prevent overfitting.

To evaluate model performance, several metrics will be used, including accuracy, precision, recall, F1 score, and the area under the ROC curve (AUC-ROC). Since predicting no-shows is an imbalanced classification problem, precision and recall will be prioritized to ensure that the model correctly identifies patients likely to miss their appointments without generating excessive false positives. Additionally, SHAP (Shapley Additive Explanations) analysis will be conducted to interpret the feature importance in the XGBoost model, providing insights into which variables have the greatest impact on patient attendance.

In addition to predictive modeling, Bayesian Causal Forests (BCF) will be used to estimate the causal impact of key variables on no-show rates. Unlike standard machine learning models that focus on correlations, BCF allows for the estimation of heterogeneous treatment effects, identifying whether interventions such as SMS reminders have a true causal effect on appointment attendance. By adjusting for confounding variables, BCF will help isolate the direct impact of factors such as waiting time and patient demographics on no-show rates. This causal inference approach will provide actionable insights for healthcare providers to design targeted interventions aimed at reducing missed appointments.

The final stage of the analysis will combine predictive insights from machine learning models with causal estimates from BCF to provide actionable recommendations for reducing no-show rates. The most important predictive factors will be highlighted, and recommendations will be made for healthcare providers to reduce no-show rates through targeted interventions, improved scheduling strategies, and better patient communication methods. The findings from this study could provide valuable guidance for optimizing hospital operations and improving patient outcomes.

3 Exploratory Data Analysis

3.1 Data Preparation

Before conducting exploratory analysis, several preprocessing steps were necessary to clean and transform the dataset. First, the dataset was checked for missing values, and no missing data was identified. To standardize and simplify analyses, the variable Handcap, originally indicating the number of disabilities, was converted into a binary indicator (0 for no disabilities, 1 for at least one disability). Additionally, the No-show variable was encoded into a binary numeric format, assigning ”Yes” (missed appointment) to 1 and ”No” (attended appointment) to 0.

During initial exploration, a few records with negative ages were discovered.

These instances, likely representing data entry errors, were removed to preserve the dataset's integrity. Another important step was creating the variable WaitingDays, calculated as the difference in days between the scheduled date and the appointment date. Some records initially had negative WaitingDays, indicating appointments recorded as occurring before they were scheduled. These rows were also removed due to their logical inconsistency.

After these cleaning steps (removing records with negative ages and negative waiting days) the final cleaned dataset comprised 110,521 appointments. This means that only six total records were removed from the original dataset of 110,527 entries, indicating minimal data loss and maintaining the integrity of the dataset.

Feature engineering further enhanced the dataset, including the creation of new variables such as AppointmentWeekday and ScheduledWeekday, which capture the day of week of appointments and scheduling, respectively. To capture temporal seasonality, AppointmentMonth was extracted from the appointment date. Furthermore, a WeatherPattern variable was introduced based on Vitória's tropical climate. Months were categorized into "Rainy" (January-March, October-December) and "Dry" (April-September) periods, allowing for analysis of how weather conditions may influence patient attendance behavior. However, since all appointments in the dataset occurred during the dry season months, both AppointmentMonth and WeatherPattern were later removed from the analysis due to lack of variability. These engineered variables allowed for deeper exploration into potential scheduling patterns influencing patient no-show behavior.

3.2 Analysis & Results

Exploratory data analysis (EDA) was conducted to identify key trends and patterns within the dataset. Initially, categorical variables were analyzed to assess their impact on no-show rates. From the plots shown in 1, gender appears to influence attendance, as females had higher absolute counts of no-shows compared to males, although this likely reflects the higher overall proportion of females making appointments. Scholarship enrollment, indicating financial assistance, also showed a visible impact; patients with scholarships displayed proportionally higher no-show rates compared to those without scholarships. Conversely, medical conditions such as hypertension, diabetes, alcoholism, and disabilities (Handcap) had minimal noticeable differences in no-show rates, suggesting that these conditions alone might not significantly predict appointment attendance.

Continuous variable analysis revealed strong insights visualized in 2. The waiting days between scheduling and the actual appointment were particularly impactful: patients with longer waiting periods showed distinctly higher no-show frequencies. Age distribution plots highlighted younger patients as having higher no-show proportions compared to older individuals, indicating age related trends in attendance behavior.

Further insights emerged from examining appointment scheduling patterns.

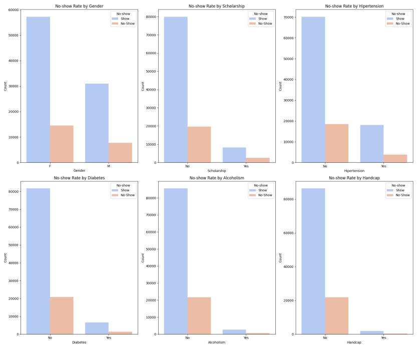

Figure 1: No-show Rates by Patient Demographics and Medical Conditions

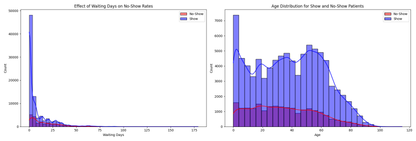

Figure 2: Distributions of Waiting Days and Patient Age by No-show Status

The appointment weekday analysis plot indicated that no-show rates varied across the week, with Saturday appointments experiencing the highest no-show rates, contrary to typical expectations. Thursday appointments exhibited the lowest no-show rates, suggesting that scheduling during midweek could potentially improve attendance rates.

Analysis of SMS reminders provided additional interesting results. Counterintuitively, patients who received SMS reminders showed higher no-show rates than those who did not. This finding might indicate that SMS reminders were being selectively sent to patients already perceived at higher risk of nonattendance or reflect other behavioral factors influencing patient response.

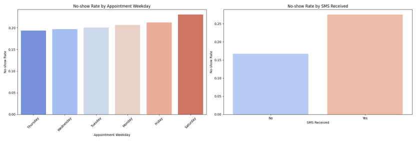

Figure 3: No-show Rates by Appointment Weekday and SMS Reminders

A correlation heatmap supported previous findings, clearly highlighting WaitingDays as positively correlated with no-shows, reinforcing its importance. Age was mildly negatively correlated with no-shows, reaffirming that older patients are somewhat more reliable in attending appointments. Other medical conditions displayed weak correlations, further diminishing their predictive importance.

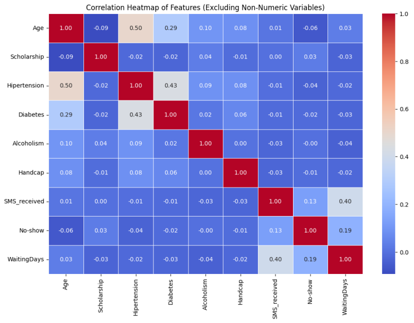

Figure 4: Correlation Heatmap of Numeric Features

Finally, interaction analyses provided more detailed insights into how variables behaved in relation to each other. As seen in 5, it was revealed that younger patients facing longer waiting periods were notably more likely to miss appointments, highlighting a critical combination of age and waiting time as predictive of higher no-show rates. The impact of SMS reminders across waiting times confirmed previous patterns that patients receiving reminders still had high no-show rates, indicating limited effectiveness of reminders alone. Please note that in 5, ”Shows” are indicated in blue and ”No-shows” indicated in orange.

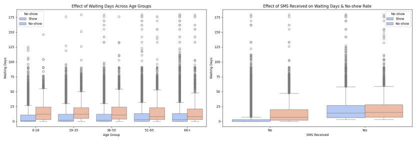

Figure 5: Interaction Effects: Waiting Days by Age Group and SMS Reminders

In summary, these exploratory findings underscore the importance of waiting times and patient age, while highlighting the limited effectiveness of SMS reminders. Healthcare providers might achieve greater attendance improvements by optimizing scheduling practices, particularly by reducing waiting days and carefully considering appointment days within the week.

4 Statistical Models

4.1 Methods

To predict whether a patient would miss their scheduled medical appointment, three supervised machine learning classification models were implemented: Decision Tree, Random Forest, and XGBoost. These models were selected due to their ability to handle nonlinear relationships and provide either interpretability or high predictive performance. Before modeling, the dataset was cleaned to remove irrelevant variables such as patient and appointment IDs, raw date fields (ScheduledDay and AppointmentDay), and the derived AgeGroup variable, since Age was already included as a continuous predictor. Categorical variables, including Gender, Neighbourhood, AppointmentWeekday, and ScheduledWeekday, were transformed using one-hot encoding to convert them into a numeric format suitable for machine learning algorithms.

After preprocessing, the dataset was split into training and test sets using a 70/30 ratio, with stratification based on the target variable to maintain the proportion of no-show cases in both subsets. The target variable, No-show, was binary, with 1 indicating that a patient did not attend the appointment and 0 indicating that they did. Given the class imbalance in the data where most patients did show up, special attention was given to evaluation metrics that account for imbalance. For the Decision Tree and Random Forest models, class weights were set to ”balanced” to give more importance to the minority class. XGBoost was configured using the log loss evaluation metric and a scaleposweight of 1 to provide a baseline for handling class imbalance.

Model performance was assessed using a combination of accuracy, precision, recall, F1-score, and the area under the ROC curve (AUC-ROC). These metrics allowed for a well-rounded evaluation of each model's ability to correctly classify no-show appointments, especially focusing on recall and AUC as key metrics due to the importance of identifying patients likely to miss their appointments. To further interpret the predictions made by the best-performing model, SHAP (Shapley Additive Explanations) was used to quantify the impact of each feature on the model's output. SHAP values were computed using the XGBoost model, and both summary and bar plots were generated to visualize global feature importance.

4.2 Results

The results of the three models showed that each had strengths and weaknesses in predicting medical appointment no-shows. The Decision Tree model achieved a ROC AUC score of 0.722 and had a strong recall of 0.8337 for the no-show class, meaning it was able to correctly identify a large proportion of actual no-shows. However, the precision for the no-show class was only 0.2993, indicating that many of the model's no-show predictions were false positives. The overall accuracy of the model was 57.25%, reflecting a moderate ability to generalize to new data. The Random Forest model showed slightly better performance with a ROC AUC score of 0.724 and a no-show recall of 0.8675. Like the decision tree, it suffered from low precision (0.2946) for no-shows but benefited from ensemble learning, which helped improve the model's ability to capture complex patterns. Its overall accuracy was 55.39%, similar to the decision tree, but it offered improved generalization through multiple decision paths.

The XGBoost model achieved the highest overall performance based on AUC, with a ROC AUC score of 0.740 and an accuracy of 79.93%. However, while the model's precision for the no-show class was higher at 0.5225, its recall dropped significantly to 0.0713, indicating that it correctly identified very few actual no-shows. This suggests that the model prioritized minimizing false positives at the cost of missing true no-show cases. Given the imbalance of the target variable and the real world importance of identifying likely no-show patients, this low recall undermines the practical utility of the model despite its high AUC.

To better understand the inner workings of the XGBoost model, SHAP analysis was conducted. The SHAP summary plot in 6 revealed that WaitingDays, the number of days between when the appointment was scheduled and when it occurred, was the most influential feature in predicting no-shows. Longer waiting periods were associated with a higher likelihood of missing appointments. Age was the second most important feature, with younger patients more likely to be no-shows. Interestingly, the feature SMS received also had a notable impact, but in a counterintuitive direction: patients who received SMS reminders were often more likely to miss their appointments. This may indicate that reminders were disproportionately sent to patients already identified as high-risk, or that the reminders alone were insufficient to alter behavior. Other features such as Gender, Scholarship, and Hypertension had smaller but still measurable contributions to the model's predictions. The SHAP bar plot 7 confirmed these findings by showing that WaitingDays, Age, and SMS received had the highest average impact on the model's output. A SHAP dependence plot for Age, shown in 8, further demonstrated that younger patients, particularly those under 30, tended to have positive SHAP values, meaning they were more likely to miss appointments. As age increased beyond 30, the SHAP values gradually decreased, indicating that older patients were less likely to no-show. A slight increase in SHAP value variability was observed among patients aged above 90, although this was likely due to fewer observations in that age range. Overall, the dependence plot reinforced the trend that advancing age was associated with a reduced likelihood of missing medical appointments. These results provide insight into key factors that influence appointment adherence and suggest possible areas for targeted intervention.

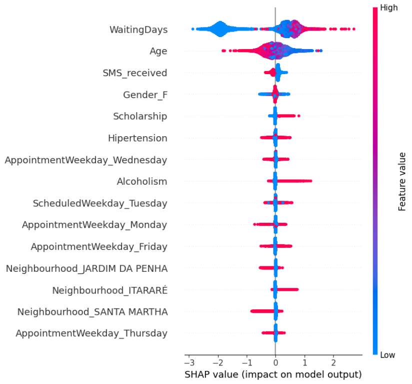

Figure 6: SHAP Summary Plot

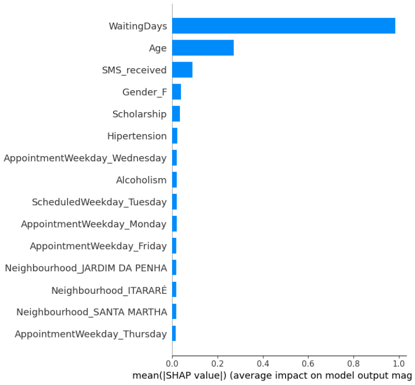

Figure 7: SHAP Importance Bar Graph

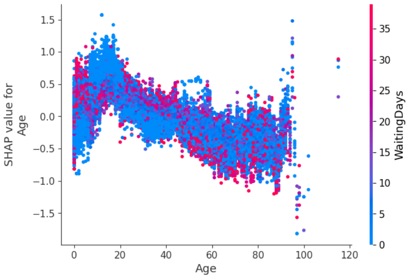

Figure 8: SHAP Dependence Plot for Age

4.3 Causal Inference Method & Results

To further investigate whether SMS reminders have a causal impact on patient attendance, Bayesian Causal Forest (BCF) analysis was conducted. In this framework, the treatment variable was defined as whether a patient received an SMS reminder, and the outcome variable was whether the patient missed their appointment. Covariates including demographic, medical, and scheduling features were used to adjust for potential confounding factors. The BCF model estimated an Average Treatment Effect (ATE) of -0.0298, indicating that receiving an SMS reminder reduced the probability of a no-show by approximately 3 percentage points on average. This suggests that SMS reminders have a modest but beneficial causal effect in improving appointment adherence.

Additionally, a distribution of the estimated individual-level treatment effects (CATEs) was plotted and is shown in 9. The histogram revealed that the majority of patients experienced a slight reduction in their no-show likelihood when receiving an SMS reminder. However, the magnitude of the causal effect varied across individuals, with some patients showing larger reductions and a small subset displaying no change or even a slight increase. Overall, the BCF

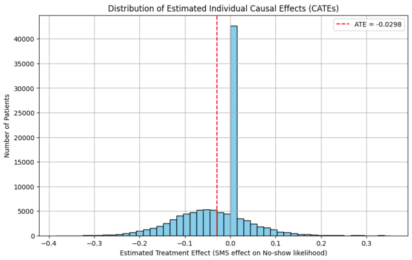

Figure 9: Estimated Individual Causal Effects (CATEs) for SMS Reminders

results reinforce the idea that SMS reminders can be an effective, though not universally impactful, intervention to improve patient attendance rates. While the XGBoost model associated SMS reminders with higher no-show rates, Bayesian Causal Forest analysis revealed that, after adjusting for confounding variables, SMS reminders causally reduced the probability of no-show. This discrepancy highlights the importance of using causal inference methods when evaluating interventions in observational data.

5 Conclusion

This study applied machine learning and causal inference methods to predict and understand medical appointment no-shows. Predictive modeling with Decision Trees, Random Forests, and XGBoost highlighted important factors such as waiting time, patient age, and SMS reminders. SHAP analysis revealed that longer waiting periods and younger age increased the likelihood of missing appointments, while receipt of an SMS reminder appeared to be associated with higher no-show rates. However, Bayesian Causal Forest analysis provided deeper insight, estimating that receiving an SMS reminder causally reduced the likelihood of no-show by approximately three percentage points after adjusting for confounding factors. This finding underscores the importance of distinguishing correlation from causation when evaluating interventions based on observational data. Overall, the results suggest that optimizing scheduling practices to reduce waiting times, targeting younger patients with more proactive reminders, and refining the use of SMS outreach could help healthcare providers improve appointment adherence rates and reduce resource inefficiencies.

References

Chen, A. M. (2023). Socioeconomic and demographic factors predictive of missed appointments in outpatient radiation oncology: an evaluation of access. *Frontiers in Health Services 3*(1).

6 Appendix

GitHub Repository Link: https://github.com/anoushkajadhav/medical-appointment-noshow-prediction.git