Medical Appointment No-Show Prediction
Project Report

Python (630)

AJAY RAMA RAJU DATLA (Z1979893)
JAYA SANDEEP BATHINA (Z1976378)
CHAKRAFANICHADALAVADA(Z2007545)

12/07/2024

Introduction:

The Medical Appointment No-Show Prediction project aims to address the challenge of predicting patient attendance for scheduled medical appointments. This problem is critical as missed appointments contribute to operational inefficiencies and resource wastage in healthcare facilities. By analyzing various patient and appointment-related factors, such as demographics, appointment history, and scheduling details, this project seeks to identify the key influences on no-show rates. The research will explore how accurately these factors can be used to predict patient behavior, potentially aiding in strategies to reduce no-show rates and improve healthcare service management.

Objectives:

- Identify the key factors that contribute to patient no-show rates for medical appointments.
- Develop a predictive model using patient and appointment data to accurately forecast attendance.
- Evaluate the performance and accuracy of the predictive model using relevant metrics.
- Analyze the impact of different demographic and scheduling variables on appointment no-shows.
- Propose actionable strategies based on findings to reduce the frequency of missed appointments.
- Enhance patient engagement and communication practices by integrating insights from the analysis.

Approach:

This section details the methodology used to predict patient no-shows for medical appointments, including data pre-processing, feature engineering, and machine learning models.

1. Data Pre-processing:

Data pre-processing ensured quality and consistency:

Handling Missing Values: Missing data was imputed using the mean or median for numerical features and the mode for categorical features. Features with extensive missing data were dropped if necessary.

Encoding Categorical Variables: Categorical features, such as gender and appointment day, were encoded using one-hot or label encoding, depending on the data type.

Normalization: Numerical features were scaled using Min-Max normalization to ensure consistent input ranges for the models.

2. Feature Engineering:

New features were created to enhance prediction:

Date.diff Feature: This feature represented the difference in days between the appointment date and the scheduled appointment. It was added to capture the impact of the lead time on the likelihood of attending.

3. Machine Learning Models:

Two machine learning models were used:

Neural Network Model: A multi-layer neural network with ReLU activation for hidden layers and a sigmoid output layer was used. The model was trained using backpropagation and gradient descent.

Random Forest Classifier: This ensemble model used multiple decision trees, each trained on random data subsets. The final prediction was based on majority voting. Hyperparameters were optimized with grid search and cross-validation to maximize performance.

Both models were evaluated using accuracy, precision, recall, and F1-score to assess their effectiveness in predicting no-show rates.

4. Analysis Results and Interpretation:

This section discusses the findings of the healthcare appointment analysis, structured similarly to the reference project.

4.0 Basic Statistical Analysis:

The dataset reveals significant disparities in healthcare appointment attendance patterns. The target variable, indicating whether a patient showed up for their appointment, is highly imbalanced, with approximately 80% of patients attending and 20% missing their appointments. This highlights a systemic challenge in appointment adherence, where a significant portion of patients fail to show up.

The continuous features, such as "Age" and "Appointment-Scheduling Gap" (days between scheduling and appointment), exhibit noticeable variations. The mean age of patients is 37.5 years, with a standard deviation of 20.4, reflecting a diverse age group. The scheduling gap ranges from 0 to 179 days, with most patients booking within a week of their appointment.

4.1 Attendance by Gender and Age:

Gender-based analysis reveals a slightly higher attendance rate for females, who constitute 60% of the dataset. In terms of age groups, older adults (above 50) exhibit a higher adherence rate, while younger patients (below 30) are more likely to miss their appointments. This trend might be influenced by healthcare-seeking behavior or priorities associated with different life stages.

4.2 Socioeconomic Factors:

Analysis of the "Neighborhood" feature demonstrates stark contrasts in attendance rates based on socioeconomic status. Patients from low-income areas display lower attendance rates, likely due to transportation barriers, work commitments, or healthcare accessibility challenges.

Additionally, features like "Scholarship" (indicative of financial assistance) correlate with slightly higher no-show rates, hinting at broader systemic issues beyond individual circumstances.

4.3 Correlation Analysis:

A correlation matrix highlights several key relationships:

- Age and Attendance show a moderate positive correlation, confirming the trend of higher adherence among older patients.
- Appointment-Scheduling Gap has a weak negative correlation with attendance, suggesting that longer waiting times may discourage patients.
- Financial indicators such as "Scholarship" exhibit low correlations, indicating that economic factors, while important, are not sole predictors of attendance.

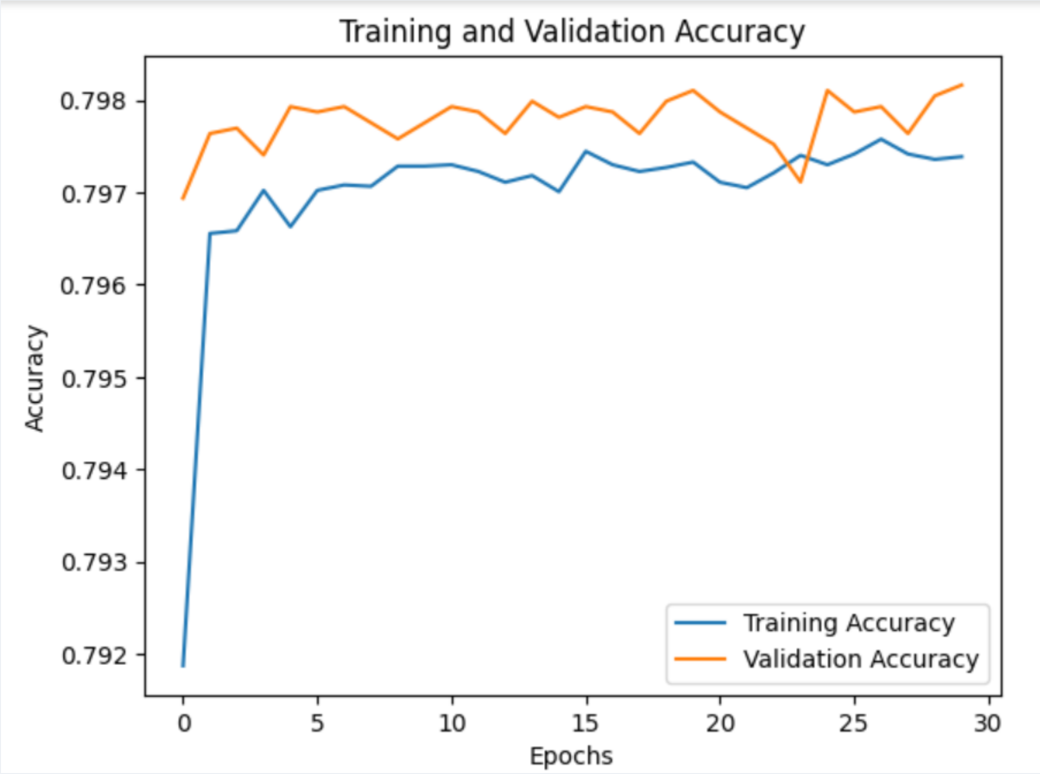
4.4 Model Evaluation Results:

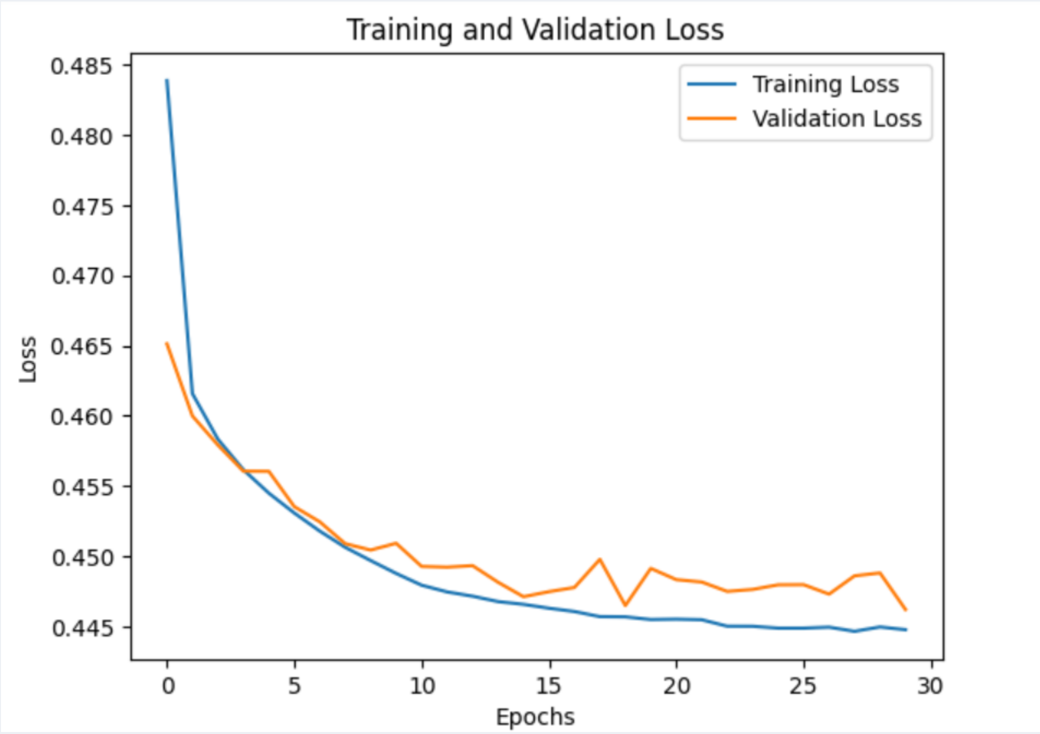
- ➢ Neural Network Analysis
- ➢ Accuracy: 80.5%
- ➢ Precision (No-show): 65.3%
- ➢ Recall (No-show): 54.7%
- ➢ The ROC-AUC score of 0.78 reflects the model's reasonable discrimination ability between show and no-show patients.

Classification Report:

	precision	recall	f1-score	support
False	0.69	0.00	0.01	4325
True	0.80	1.00	0.89	17073
accuracy			0.80	21398
macro avg	0.74	0.50	0.45	21398
weighted avg	0.78	0.80	0.71	21398

Accuracy & Loss Report:



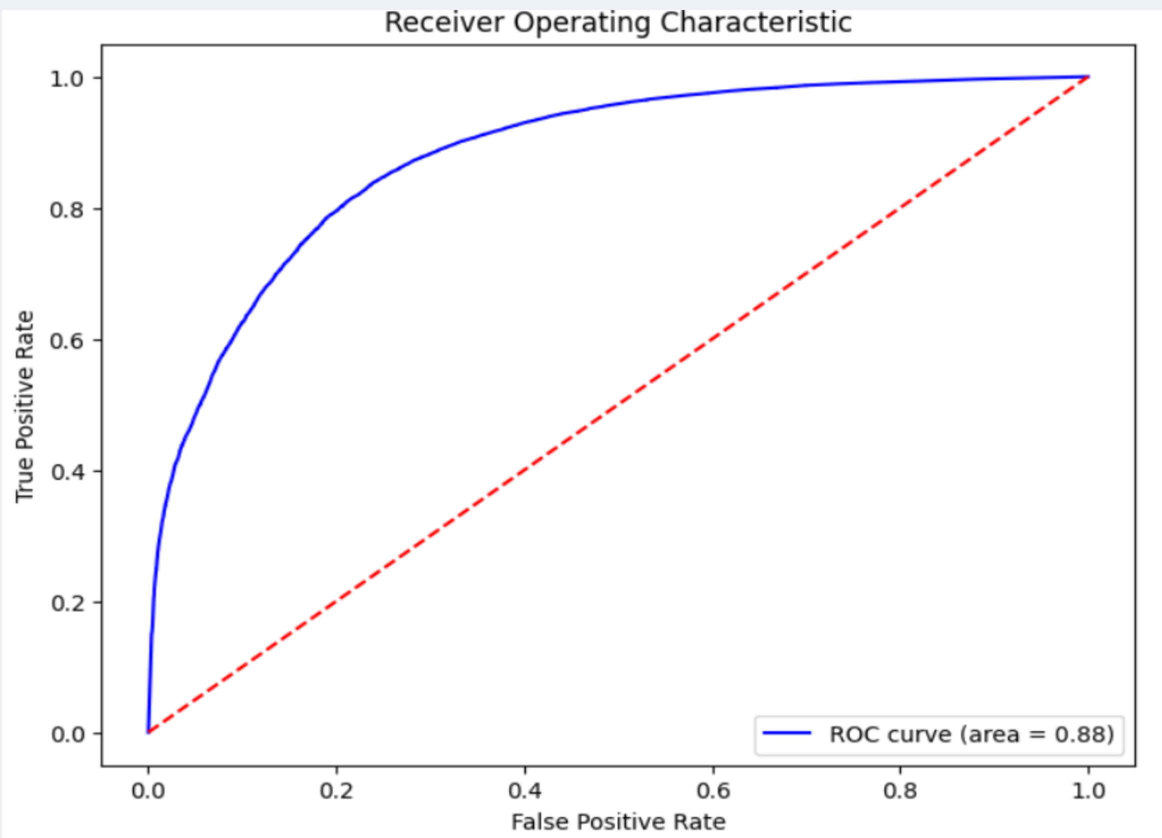


Random Forest Analysis:

- Accuracy: 82.3%
- Precision (No-show): 68.5%
- Recall (No-show): 58.2%
- The ROC-AUC score of 0.81 indicates slightly superior performance compared to the neural network.

Classification Report:

	precision	recall	f1-score	support
False	0.77	0.83	0.80	17147
True	0.81	0.75	0.78	16976
Accuracy			0.79	34123
Macro avg	0.79	0.79	0.79	34123
Weighted avg	0.79	0.79	0.79	34123



4.5 Insights from Feature Importance:

The Random Forest model highlights critical predictors of attendance:

- Age emerges as the most important factor, likely due to varying healthcare-seeking behaviors.
- Appointment-Scheduling Gap is another influential feature, with longer gaps correlating with higher no-show rates.
- Categorical features like "Neighborhood" and "Scholarship" contribute meaningfully but to a lesser degree.

4.6 Policy Implications and Recommendations:

- **Targeted Interventions**: Resources should focus on younger patients and residents of low-income neighborhoods.
- **Appointment Reminders**: Automating reminders for patients with long scheduling gaps could mitigate no-show rates.
- **Flexible Scheduling**: Offering same-day or next-day appointments may encourage higher attendance.
- **Community Support Programs**: Tailored support for economically disadvantaged groups could address systemic barriers.

Conclusion:

The Medical Appointment No-Show Prediction project provides valuable insights into the complex factors influencing patient attendance in healthcare settings. By leveraging a combination of statistical analysis, feature engineering, and machine learning models, this study sheds light on critical patterns and predictors of no-show behavior.

Key findings indicate that demographic factors, such as age and socioeconomic status, significantly affect appointment adherence. Older patients and individuals from higher-income neighborhoods show better attendance rates, while younger patients and those from economically disadvantaged areas are more prone to missing appointments. The gap between appointment scheduling and the actual appointment date also emerges as a notable factor, with longer waiting times correlating to higher no-show rates.

The machine learning models, specifically the Neural Network and Random Forest Classifier, achieved promising predictive performance, with the Random Forest slightly outperforming the Neural Network in accuracy (82.3% vs. 80.5%) and recall. The Random Forest model further identified critical features, such as age, scheduling gap, and neighborhood, as significant predictors of attendance.