Healthcare Appointment No-Show Prediction Report

Abstract

This report outlines the development of a predictive model to identify patients likely to miss healthcare appointments, enabling optimized scheduling and resource allocation. Using a decision tree classifier, the project leverages patient data to predict no-shows with high accuracy. Exploratory data analysis (EDA) and statistical tests uncover trends related to SMS reminders, age, and appointment weekdays. A Power BI dashboard visualizes these insights, and actionable recommendations are provided to reduce no-show rates. The project successfully integrates data cleaning, modeling, and visualization to deliver a comprehensive solution for healthcare providers.

1 Introduction

Healthcare appointment no-shows pose significant challenges, leading to wasted resources and reduced patient care quality. The objective of this project is to predict whether patients will miss appointments and provide insights to optimize scheduling. By analyzing patient demographics, appointment details, and health conditions, the project aims to identify key factors influencing no-shows. The solution includes a predictive model, an interactive Power BI dashboard, and recommendations to enhance appointment adherence. This report details the tools, methodology, and outcomes of the project, emphasizing data-driven strategies to improve healthcare operations.

2 Tools Used

The project utilizes a combination of programming and visualization tools to achieve its objectives:

- **Python**: For data processing, analysis, and modeling.
  - **Pandas**: Data manipulation and cleaning.
  - **NumPy**: Numerical computations.
  - **Scikit-learn**: Decision tree modeling and evaluation.
  - **Matplotlib/Seaborn**: Data visualization for EDA.
- **Power BI**: For creating an interactive dashboard to visualize trends and model performance.

3 Steps Involved in Building the Project

3.1 Data Cleaning

The dataset, containing patient and appointment details, was cleaned to ensure quality. Missing values in Age and Date.diff were imputed with medians, and categorical variables were verified for consistency.

3.2 Exploratory Data Analysis (EDA)

EDA was conducted to uncover patterns. Univariate analysis examined distributions of Age, Date.diff, and Showed_up. Bivariate analysis revealed higher show-up rates for patients receiving SMS reminders and lower no-show rates on midweek days. Visualizations, including histograms and bar charts, were generated for Power BI integration.

3.3 Data Preprocessing

Categorical variables (Gender, Neighbourhood) were encoded, and numerical features were scaled. Outliers in Age and Date.diff were capped. The data was split into 80% training and 20% testing sets, ensuring stratified sampling to handle class imbalance.

3.4 Decision Tree Modeling

A decision tree classifier was trained with hyperparameter tuning via grid search, optimizing the F1-score. The model achieved strong performance (e.g., F1-score ≈ 0.80), with Date.diff and SMS_received as top predictors. Feature importance and predictions were saved for visualization.

3.5 Power BI Dashboard

A Power BI dashboard was developed using processed data, feature importance, and predictions. Visuals included bar charts (no-show rates by SMS_received, Age_group), a line chart (weekday trends), and a confusion matrix. Slicers enabled interactive exploration of trends.

3.6 Optimization Recommendations

Recommendations include enhancing SMS reminders, targeting high-risk groups (e.g., younger patients), optimizing scheduling for midweek days, and implementing follow-up calls for predicted no-shows. These strategies aim to reduce no-show rates by 5–10%.

4 Conclusion

The Healthcare Appointment No-Show Prediction project successfully developed a decision tree model to predict no-shows with high accuracy. EDA and statistical analysis identified key predictors, such as SMS reminders and appointment timing, which were visualized in a Power BI dashboard. The recommendations provide actionable strategies to reduce no-shows, improving healthcare efficiency. Future work could explore ensemble models (e.g., Random Forest) and real-time data integration to further enhance performance. This project demonstrates the power of data-driven insights in addressing operational challenges in healthcare.