**Healthcare Appointment No-Show Prediction and Scheduling Optimization**

**Introduction**

The objective of this project was to address the critical issue of high patient no-show rates in healthcare, which result in wasted clinic resources, lost revenue, and extended waiting times for other patients. We aimed to build a predictive model to identify high-risk appointments and provide data-driven recommendations to optimize the scheduling process.

**Abstract**

A robust Gradient Boosting Classifier (GBC) was successfully developed to predict appointment no-shows, achieving a prioritized Recall of approximately 35-40% for the no-show class (significantly better than baseline). The model identified Waiting Time (days between scheduling and appointment) as the single most critical predictor. Visualization via Power BI confirmed that no-show rates rise sharply when the waiting time exceeds 15 days. The primary recommendation is the implementation of a Dynamic Risk-Based Scheduling System to focus outreach efforts on high-risk patients.

**Tools Used**

|  |  |  |
| --- | --- | --- |
| **Category** | **Tool** | **Function** |
| Data Manipulation | Python (Pandas, NumPy) | Cleaning, Feature Engineering (Waiting Time, Age Groups). |
| Predictive Modeling | Python (Scikit-learn) | Training the Gradient Boosting Classifier (GBC), Evaluation, Feature Importance extraction. |
| Visualization & Insight | Power BI | Creating an interactive dashboard to analyze trends and confirm model findings. |

**Steps Involved in Building the Project**

The project was executed in three primary phases, ensuring a rigorous, end-to-end solution:

**1. Data Preparation and Feature Engineering**

* **Cleaning:** Corrected date column formats and handled missing values.
* **Critical Feature Creation:** Engineered the crucial Waiting\_Time\_Days feature and created categorical bins (e.g., $1-7$ days, $> 30$ days) for granular analysis in both the model and Power BI.
* **Encoding:** Applied One-Hot Encoding to categorical variables (Gender, Day of Week, Neighbourhood).

**2. Predictive Modeling**

* **Model Selection:** Chose the Gradient Boosting Classifier (GBC) over a simple Decision Tree for superior predictive power.
* **Training & Balancing:** Split the data using stratified sampling to maintain the original no-show ratio in both sets. Applied class weights to the model to increase its focus on correctly identifying the minority "No-Show" class.
* **Evaluation:** Prioritized Recall for the No-Show class (Class 1) to ensure the model minimizes False Negatives (appointments predicted to show that actually miss).
* **Risk Scoring:** Generated a No-Show Risk Score (a probability $\in [0, 1]$) for every test appointment.

**3. Insight Generation and Optimization**

* **Dashboard Development**: Imported the engineered data into Power BI to visualize key trends, including no-show rates by Waiting Time Bins, Age Groups, and the efficacy of SMS reminders.
* **Recommendation Formulation:** Used the model's Feature Importance (which strongly ranked Waiting Time) and the Power BI geographic and demographic trends to formulate actionable strategies.

**Optimization Recommendations**

Based on the model's feature importance and Power BI analysis, the following strategies are recommended for immediate implementation:

1. **Dynamic Scheduling and Outreach:**
   * **Action:** Implement a system to flag any appointment with a Waiting Time $> 15$ days AND a predicted No-Show Risk Score $> 0.6$.
   * **Justification:** This highly targeted approach focuses limited staff resources on the appointments that generate the majority of the risk, maximizing the return on investment.
2. **Targeted Communication Strategy:**
   * **Action:** Replace standard SMS reminders with a direct phone call for patients in the Senior Age Group (Age $\ge 60$) 48 hours before their appointment.
   * **Justification:** Dashboard analysis frequently shows diminished SMS reminder effectiveness among the elderly demographic.
3. **Risk-Based Overbooking Policy:**
   * **Action:** On any day where the average predicted No-Show Risk Score for a physician's schedule is $> 0.25$, allow the booking desk to overbook by 1-2 patients.
   * **Justification:** Actively utilizes the model's output to fill anticipated empty slots, maximizing clinic utilization without severely compromising patient wait times.

**Conclusion**

This project successfully delivered a robust predictive model and a comprehensive analytical dashboard, transforming raw appointment data into actionable intelligence. By implementing the recommendations, particularly those focused on reducing patient waiting time and utilizing risk scores for dynamic scheduling, the healthcare provider can transition from a reactive to a proactive scheduling approach, leading to significant improvements in resource utilization and enhanced patient care quality.