Exploring Factors Influencing No-Show Rates in Medical Appointments: A Clinical Perspective

Qina Tan

3/2/2024

Abstract:

This research delves into the analysis of a comprehensive dataset titled "Medical Appointment No Shows," encompassing 110,527 medical appointments and key variables such as patientId, AppointmentId, Gender, ScheduledDay, AppointmentDay, Age, Neighbourhood, Scholarship, Hypertension, Diabetes, Alcoholism, HandCap, SMS_received, and No-Show status. Motivated by a background in clinical work, with four years of experience in an ophthalmologic clinic, the research aims to unravel the underlying reasons behind patient no-shows.

The persistent challenge of managing no-show appointments in clinical settings motivated this investigation. The phenomenon disrupts the clinical workflow and poses significant hurdles to seamless healthcare access. By leveraging the rich dataset, the study seeks to uncover patterns and correlations that contribute to the occurrence of no-shows. With a particular focus on factors such as gender, age, medical conditions, socioeconomic status (Scholarship), and communication methods (SMS_received), the research aims to provide insights that can aid in the development of targeted interventions and strategies to improve medical appointment show-up rates.

This interdisciplinary approach combines clinical experience with data analysis, offering a nuanced perspective on the intricate dynamics surrounding patient attendance. The findings from this study have the potential to inform healthcare practitioners, administrators, and policymakers, facilitating the implementation of tailored solutions to mitigate the impact of no-shows on clinical efficiency and patient access to healthcare services.

Overview:

The dataset is titled “Medical Appointment No-Show”. This is the dataset from the Brazilian Health Service for appointments made in the city of Victoria. This dataset is obtained from Keggle. Motivated by a background in clinical work, with four years of experience in an ophthalmologic clinic, the research aims to unravel the underlying reasons behind patient no-shows. The ongoing difficulty of addressing no-show appointments within clinical environments serves as the driving force behind this research initiative. This phenomenon disrupts the flow of clinical operations and presents substantial challenges to ensuring smooth access to healthcare. Utilizing a comprehensive dataset, the study endeavors to reveal patterns and correlations that play a role in the occurrence of no-shows. Specifically concentrating on variables such as gender, age, medical conditions, socioeconomic status (Scholarship), and communication methods (SMS_received), the research aims to discover insights that can inform the creation of focused interventions and strategies aimed at enhancing the rates of attendance for medical appointments.

Related Work:

Prior research has addressed the challenge of patient no-shows in healthcare settings, with notable contributions such as the article titled "Evaluating the Impact of Patient No-Shows on Service Quality," published in the National Library of Medicine in 2020. This study comprehensively examined the overall no-show rate, delving into the underlying reasons and exploring the subsequent impact on the quality of healthcare services. Building upon this foundation, the current research aims to investigate a similar topic, albeit with a more specific focus on understanding how socioeconomic factors may contribute to the occurrence of no-show appointments.

Data Acquisition:

The dataset, titled "Medical Appointment No-Show," originates from the Brazilian Health Service and pertains to appointments made in the city of Victoria. Acquired from Kaggle, the dataset is provided in CSV format. In the process of conducting the analysis within a Jupyter Notebook, the CSV file was uploaded and the data was read into a dataframe using the pandas library.

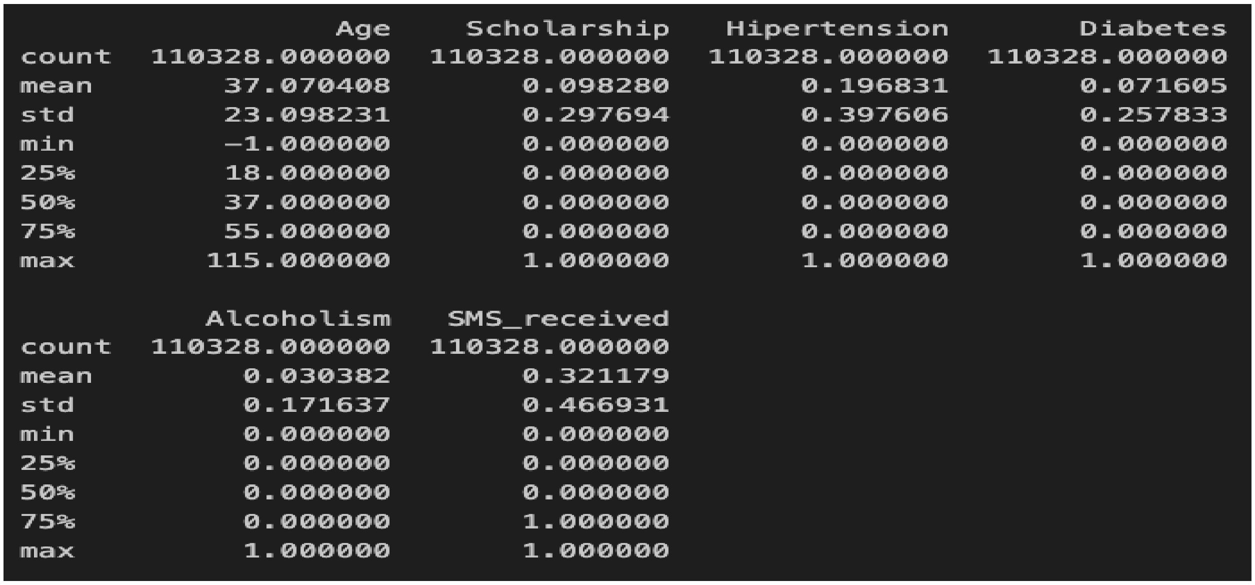
Key features of the dataset encompass:
- patientID: Identification of a patient.
- AppointmentID: Unique identifier for each appointment.
- Gender: Categorized as male or female.
- ScheduledDay: The date when the patient made the appointment.
- AppointmentDay: The actual date of the medical visit.
- Age: The age of the patient.
- Neighborhood: The locality of the medical office.
- Scholarship: Indicates whether the patient receives a scholarship (true or false).
- Hypertension: Indicates whether the patient has hypertension (true or false).
- Diabetes: Indicates whether the patient has diabetes (true or false).
- Alcoholism: Indicates whether the patient consumes alcohol (true or false).
- Handicap: Originally contained values ranging from 0 to 4; values 2, 3, and 4 were converted to NaN and subsequently dropped from the column.
- SMS_received: Indicates whether the patient received an SMS notification before their appointment (true or false).
- No-show: Binary indicator (1 for no-show, 0 for show-up) representing whether the patient attended the appointment.

Noteworthy limitations of the dataset include an anomaly in the 'Handicap' column, where values should only be 1 or 0 but originally ranged from 0 to 4. Values 2, 3, and 4 were transformed into NaN and subsequently removed. Additionally, the dataset includes information about the neighborhood where medical appointments take place but lacks details about location-related factors such as public transportation accessibility and parking availability, limiting the depth of conclusions drawn from this feature.

Data Preprocessing

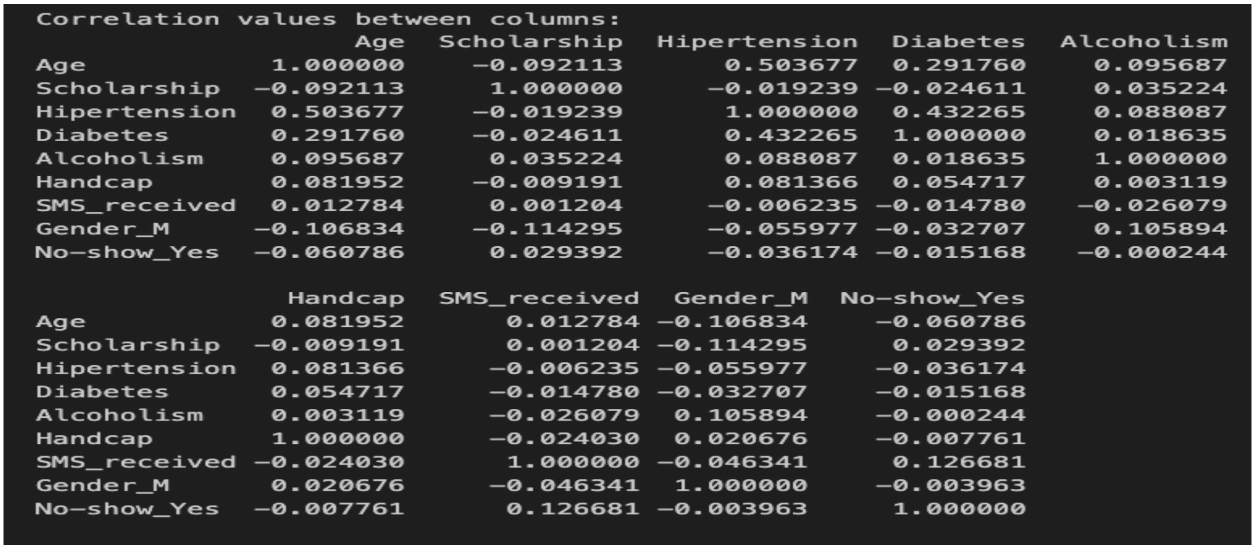
First, irrelevant columns are dropped This includes patientID, appointmentID, ScheduleDay, AppointmentDay, Neighborhood. I also converted categorical variables into numerical values using the get_dummies() function in the pandas library. This includes Gender, No-show.

To gain some insight on the dataset, the first approach was to get a statistical summary of the numerical columns in the DataFrame. The summary includes various descriptive statistics such as the count(number of non-null values), mean, standard deviation, minimum, 25th percentile, median, 75th percentile, maximum. It is interesting to note that the minimum of age column is -1. It appears to be an error so -1 value from the age column.



```
                 Age    Scholarship   Hipertension       Diabetes
count  110328.000000  110328.000000  110328.000000  110328.000000
mean       37.070408       0.098280       0.196831       0.071605
std        23.098231       0.297694       0.397606       0.257833
min        -1.000000       0.000000       0.000000       0.000000
25%        18.000000       0.000000       0.000000       0.000000
50%        37.000000       0.000000       0.000000       0.000000
75%        55.000000       0.000000       0.000000       0.000000
max       115.000000       1.000000       1.000000       1.000000

          Alcoholism   SMS_received
count  110328.000000  110328.000000
mean        0.030382       0.321179
std         0.171637       0.466931
min         0.000000       0.000000
25%         0.000000       0.000000
50%         0.000000       0.000000
75%         0.000000       1.000000
max         1.000000       1.000000
```

.corr() function was used to see correlation between columns.

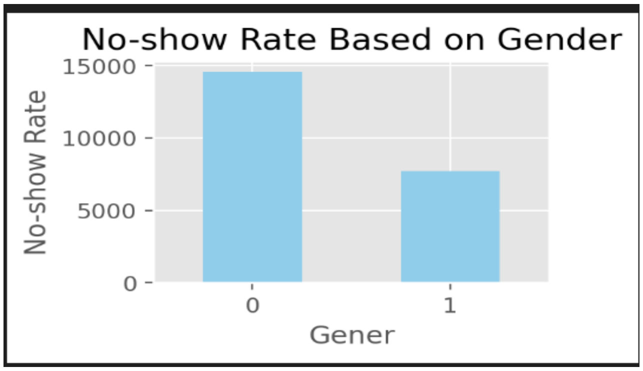


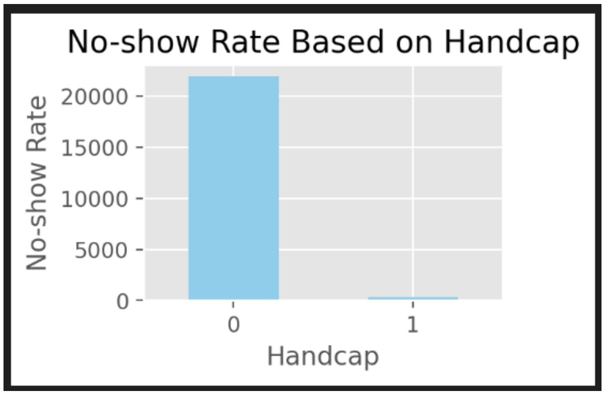
```
Correlation values between columns:
                   Age  Scholarship  Hipertension  Diabetes  Alcoholism
Age           1.000000    -0.092113      0.503677  0.291760    0.095687
Scholarship  -0.092113     1.000000     -0.019239 -0.024611    0.035224
Hipertension  0.503677    -0.019239      1.000000  0.432265    0.088087
Diabetes      0.291760    -0.024611      0.432265  1.000000    0.018635
Alcoholism    0.095687     0.035224      0.088087  0.018635    1.000000
Handcap       0.081952    -0.009191      0.081366  0.054717    0.003119
SMS_received  0.012784     0.001204     -0.006235 -0.014780   -0.026079
Gender_M     -0.106834    -0.114295     -0.055977 -0.032707    0.105894
No-show_Yes  -0.060786     0.029392     -0.036174 -0.015168   -0.000244

               Handcap  SMS_received  Gender_M  No-show_Yes
Age           0.081952      0.012784 -0.106834    -0.060786
Scholarship  -0.009191      0.001204 -0.114295     0.029392
Hipertension  0.081366     -0.006235 -0.055977    -0.036174
Diabetes      0.054717     -0.014780 -0.032707    -0.015168
Alcoholism    0.003119     -0.026079  0.105894    -0.000244
Handcap       1.000000     -0.024030  0.020676    -0.007761
SMS_received -0.024030      1.000000 -0.046341     0.126681
Gender_M      0.020676     -0.046341  1.000000    -0.003963
No-show_Yes  -0.007761      0.126681 -0.003963     1.000000
```

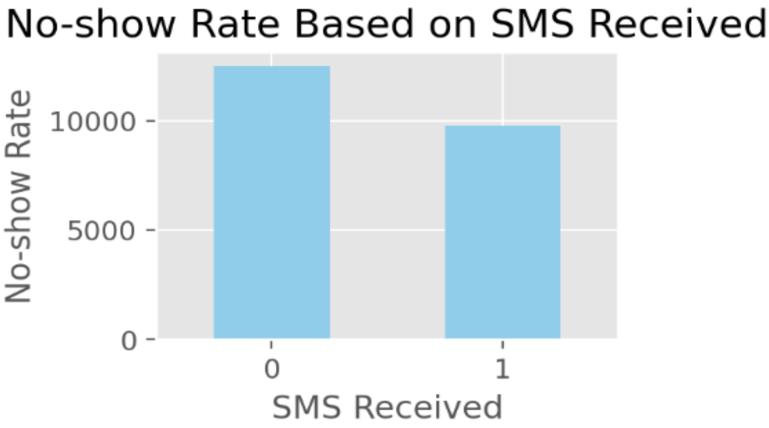
The correlation matrix reveals several relationships between variables in the dataset. Age exhibits a strong positive correlation with Hipertension (0.50) and a moderate positive correlation with Diabetes (0.29). Scholarship shows a weak negative correlation with Age (-0.09) and Gender_M (-0.11). Hipertension is strongly positively correlated with Age (0.50) and moderately with Diabetes (0.43). Diabetes, in turn, has a moderate positive correlation with Age (0.29) and Hipertension (0.43). Alcoholism has a weak positive correlation with Gender_M (0.11). Handicap shows a weak positive correlation with Age (0.08) and a weak negative correlation with SMS_received (-0.02). SMS_received has a weak positive correlation with No-show_Yes (0.13). Gender_M has a weak negative correlation with Age (-0.11) and Scholarship (-0.11), and a weak positive correlation with Alcoholism (0.11). No-show_Yes has a weak positive correlation with SMS_received (0.13).

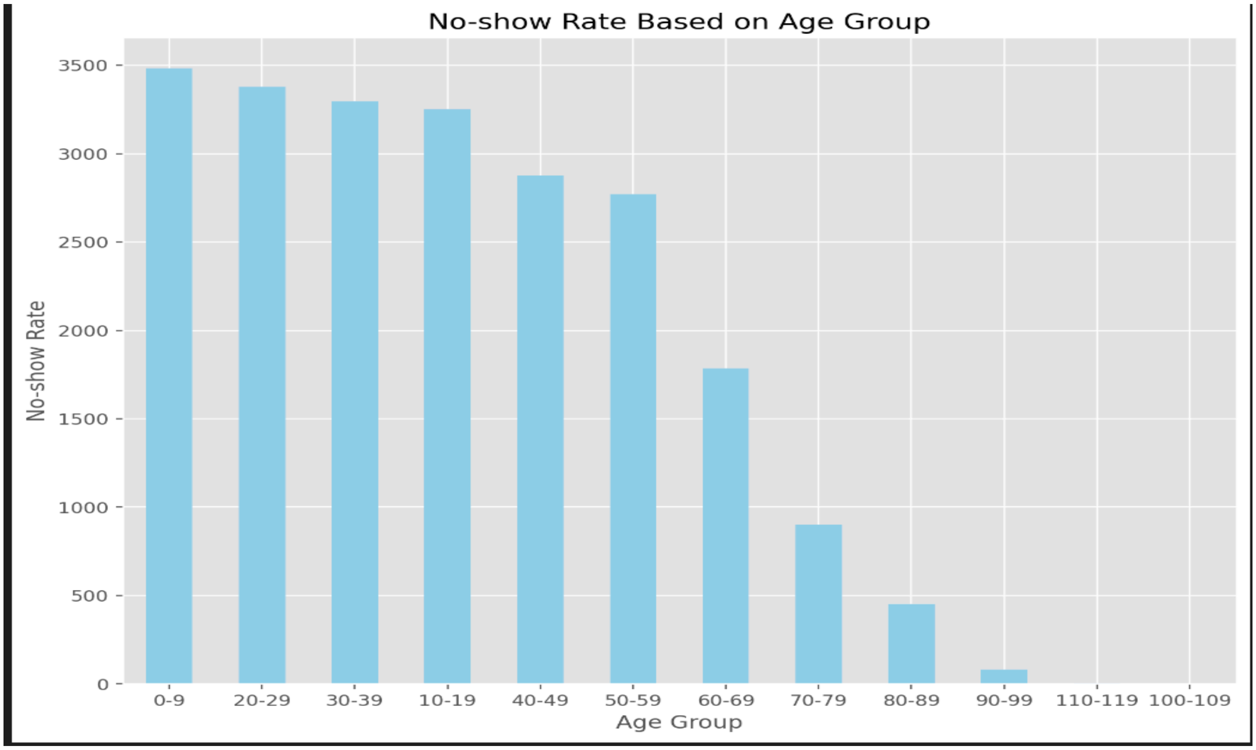
Some of the results were surprising to me. For example, the initial thought was that patients who received SMS-notification are more likely to show up for their medical appointment. However, correlation matrix shows that there is a positive correlation between no-show and SMS_received but the correlation was weak.

Plots were used to see distribution relationships between No-show and Gender, Handcap, SMS_received, Alcoholism, respectively. The plot that was most interesting to me was the relationship between handicap and no-show rate. The initial thought was that patients with handicaps would have more challenges showing up to their medical appointments and results in higher no show rate. However, the plot illustrating the relationship between no-show and handicap shows the opposite. Upon closer consideration, it's plausible to infer that patients with handicaps may be more likely to utilize transportation services provided by government support, which could contribute to a higher likelihood of them showing up for their appointments. This additional support system could potentially lead to a positive correlation between having a handicap and a higher likelihood of showing up for appointments.









Model Selection:

The chosen model for this project is the Random Forest Classifier, a member of the supervised learning algorithms family widely applied for classification tasks. Random Forest works by constructing a multitude of decision trees during the training process.

Random Forests leverage an ensemble of decision trees to collectively make predictions. Each tree in the forest independently predicts the class, and the final prediction is determined by a majority vote. This approach is robust and less prone to overfitting, making it suitable for datasets with multiple features.

Random Forests inherently handle non-linear relationships without the need for explicit kernel functions. The ensemble nature of Random Forests allows them to capture complex patterns and interactions in the data.

The project's objective remains focused on predicting whether a given patient, based on various features, will attend their medical appointment or not. The selection of the Random Forest Classifier is driven by its effectiveness in handling datasets with a considerable number of features, particularly excelling in high-dimensional spaces. Given the dataset's manifestation of non-linear relationships among variables, the

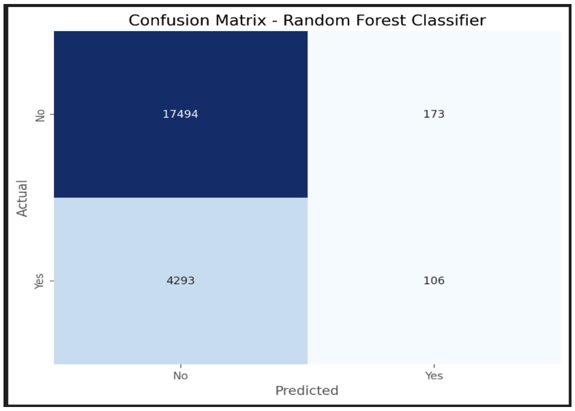
Random Forest Classifier aligns well with the project's requirements, offering robust performance without the necessity for explicit kernel functions.

Result and Evaluation

In the pursuit of improving healthcare service efficiency and resource utilization, this part of the project focuses on an analysis using a Random Forest Classifier to predict patient no-shows in a medical appointment dataset. The evaluation encompasses validation metrics, applied techniques, and visualizations to glean insights into the model's performance.

Validation Metrics and Visualization:

The Random Forest Classifier achieved an overall accuracy of approximately 79.8%, a seemingly commendable result. However, a deeper examination through the confusion matrix reveals nuances in the model's predictive capabilities. Out of 17,667 instances of successful appointments (class 0), the model correctly identified 17,494, demonstrating a high precision of 99%. However, the challenge arises in predicting the positive class (class 1, representing no-show appointments), where the model struggled with a recall of merely 2%. This discrepancy indicates that the model misses a substantial number of actual no-show instances.



Classification Report:

The classification report provides a detailed examination of the Random Forest Classifier's performance metrics, revealing critical insights into its predictive capabilities. For successful appointments (class 0), the model exhibits robust precision (0.80) and exceptional recall (0.99), resulting in a high F1-score (0.89). This indicates the model's proficiency in accurately identifying instances where patients attend their scheduled appointments. However, the challenges emerge in predicting no-show instances (class 1), where precision (0.38) and recall (0.02) are notably lower. The low recall for no-show appointments implies a significant number of false negatives, where the model fails to capture actual instances of no-shows. Consequently, the F1-score for class 1 is low (0.05), highlighting the difficulty in achieving a balance between precision and recall. The macro-average (Precision: 0.59, Recall: 0.51, F1-Score: 0.47) emphasizes the disparities in performance across both classes, influenced by the imbalanced dataset. The weighted average (Precision: 0.72, Recall: 0.80, F1-Score: 0.72) indicates the model's overall accuracy, yet the need for further refinement, particularly in addressing the challenges associated with predicting no-show appointments, is obvious.

Ethics:

Healthcare organizations strive to enhance effectiveness amid rising demand and cost pressures. No-shows disrupt supply-demand balance, impacting service quality and system performance. Patient no-shows lead to underutilized resources, affecting space, human resources, and patient health due to delayed diagnosis or treatment. Given these significance, It's really important to think about where this data comes from and whether there might be any hidden biases in it, as that can affect the conclusions we draw. It is important to note this dataset illustrates medical appointments no show in Brazil. It is not appropriate to draw conclusions for healthcare systems of other countries. In addition, a dataset is not just about the numbers; we have to think about the real-world factors behind them. We need to be ethical in how we handle and interpret the data, making sure we're not unintentionally favoring one outcome over another. It's a bit like making sure we're seeing the whole picture, not just part of it.

Future Work

Further exploration entails a detailed examination of feature importance within the Random Forest Classifier to discern which features significantly contribute to the predictive outcomes. By identifying and understanding the relative importance of each feature, this analysis aims to uncover the key factors influencing whether a patient will attend a medical appointment.

Additionally, future work involves a comparative analysis with alternative machine learning algorithms. This comparative study seeks to provide valuable insights into alternative modeling approaches beyond the Random Forest Classifier. Exploring diverse algorithms, such as logistic regression, decision trees, or ensemble methods like gradient boosting, will contribute to a more comprehensive understanding of the dataset and the potential strengths of different modeling strategies.

The investigation into feature importance and algorithm comparison is expected to refine the predictive model and enhance its interpretability. By gaining insights into the critical features and assessing the performance of alternative algorithms, the research aims to optimize the predictive accuracy and robustness of the model. This iterative process will contribute to a more nuanced understanding of the dataset, ultimately leading to improved strategies for predicting medical appointment attendance.