Hospital Appointment No Show predication

Abstract

Hospital appointment no show is a comment issue that cause problems to health organization and health practices in both government and privet sector. It is effected on the hospital slot scheduling and the hospital revenue.The goal of this project is to use classification models for Noshow predication in order to help in this issue.

Design

The predication of the patient who will not attend their appointment or the number of the no show would help the organization in the decision making and to reduce the no show rate. Also it would help to set a plan to mange the hospital appointments. The data will be exploring to answer following question:

- What is the most factor that cause the patient not attend their appointment?
- Is the patient with chronic disease like hypertension and diabetic not attending their appointment?
- Is The day of Week affected on patient attendee?
- Is the long waiting time to get the appointment affected on patient attendee?

Data

- Dataset used in this project is a patient appointment data with the status of the patient as show or no show
- The source of the data is from kaggle. https://www.kaggle.com/joniarroba/noshowappointments
- The data consist of 110,527 records with 13 features. Scholarship, Hypertension Diabetes, Alcoholism, Handicap, SMS_received are the feature with value of 0 and 1.
- Patient ID and appointment ID is a database column that will not be included as a feature.
- Schedule day and appointment day is a date time type and it will be used to get an interested feature which is the weekday and the waiting time that can affected on the show or no show of the patient.
- The remaining features are gender, neighborhood and patient age which are interested as well and it can show a relation with the no show of the patient.

Algorithms

❖ **Data Preprocessing**

As data preprocessing following actions have been applied:

- Convert date columns “ScheduledDay” and “AppointmentDay” form float to date date type.
- Remove Age value = -1
- Adding Waiting time and Appointment Week day to the data
- Waiting time contains minus data, this is happening when ScheduledDay grater than AppointmentDay. As data cleansing for these cases the ScheduledDay replaced by the AppointmentDay and vice versa. Assuming this is data entry mistake.
- Remove Patient ID and Appointment ID since they are a database column and no meaning to have them in the data
- Remove date columns: “ScheduledDay” and “AppointmentDay” after convert them to Day, month and year from model better performance

❖ **EDA**

Some of exploratory data analysis applied

  - A graph for total Show & No Show cases

- A graph for No show cases by Diabetes, hypertension, day of the week and waiting time to find the relation between NoShow and these features.

❖ Model

- Data is imbalance and oversampling technique has been applied “SMOTE”.
- Random forest, Decision tree and Logistic regression were used.
- Logistic regression was best performance “high F1 score” than other techniques since in Noshow predication we need to balance between FP and FN.

❖ Model Evaluation and Selection

- The data was split into 70/30 training vs testing and below the sores
- Below the score of Logistic regression

    - **F1 Score:** 0.400
    - **Test Precision:** 0.366
    - **Test Recall:** 0.316

❖ Tools

- Numpy and Pandas for data manipulation.
- Scikit-learn for modeling.
- Matplotlib and Seaborn for plotting.