**Prediction of Show/No-Show**

**Background:** Over 110 thousand of medical appointments took place in Brazil, some patients did not show up for their appointments. The dataset contains 14 features including the present of the patient or not on his appointment date. The features can be classified into 4 groups, patient information (id, gender, age), appointment information (appointment id, appointment date, scheduled date, no show, sms received), health information (hypertension, diabetes, alcoholism, handicap) and social information (Neighborhood, Scholarship). **A data dictionary is provided**.

1. **Data Wrangling. Submit the EXCEL file or the Jupyter Notebook file with answers to the following questions.**
2. Create two new categorical variables called **WeekdayAppointment** and **WeekdayScheduled**. Obviously, the values that these variables can take are seven days of the week. Fiil in those variables for each appointment in the data.
3. Doctor’s office is closed during the weekend thus they do not schedule new appointments in the weekends (hint: consider the **ScheduledDay** variable). Are there appointments scheduled in the weekend in the data? If so, how many? What do you suggest to handle these incorrect entries?
4. Doctor’s office is closed during the weekend thus they do not see patients in the weekends (hint: consider **AppointmentDay** variable). Are there patients seen in the weekend in the data? If so, how many? What do you suggest to handle these incorrect entries?
5. Are there any inconsistent entries in any column? For example **ScheduledDay** should take place before **AppointmentDay**. Gender, handicap and scholarship should have a unique value per patient. For every patient, if the age is different in different appointments, then the appointment with older age should take place after the appointment with younger age.
6. Are there any incorrect values in any of the other columns? For example, negative ages? If so, how many illegitimate entry are there in each column and what do you suggest to handle them?
7. Are there any outliers in any column? For example, appointments with age more than 100? If so, remove those entries with outlier values.
8. **Exploratory Data Analysis and Feature Engineering. Submit the EXCEL file or the Jupyter Notebook file with answers to the following questions.**
9. How many unique patients exist in the data?
10. Consider categorizing any numerical variable. For example, age can be categorized as [0-18)->Kids, [18-37)->Young Adults, [37-65)->Older Adults, 65+ -> Elderly.
11. What percentage of patients have more than one appointment?
12. What percentage of patients with more than one appointment have gone to all appointments?
13. Is there specific weekday when most patients miss their appointments?
14. If a patient has missed appointments in the past, then it is more likely that he will miss an appointment. Create a variable called **NoShowHistory** to consider this information when building models to predict no-shows.
15. Which genders and age groups, patients are most likely not to show up to their appointments?
16. If an appointment is scheduled too early then the no show probability increases. Create a new variable called **AwaitingTime (**which should be equal to **AppointmentDay - ScheduledDay)** to consider this information. Categorize **AwaitingTime** variable as 0 days, 1-4 days, 5-15 days and 16+days. Note that the time portions have vanished from the **AppointmentDay**, because all appointment times were set exactly to 00:00:00.
17. What percentage of the patients schedule their appointments on the same day, and what percentage of those patients who schedule on the same day show up for their appointment?
18. Do SMS reminders help increase the show up rate for appointments in different **AwaitingTime** categories?
19. Is there any correlation between patients’ positive records in hypertension, diabetes, alcoholism and handicap and them not showing up to their appointments?
20. Which neighborhood has the most no-show rate? Are neighborhoods with more scholarship patients are more likely to not show up?
21. **Prediction and analysis of results. You can use one of the Jupyter Notebooks that I posted during the semester as a template to answer the questions in this section. Submit the Jupyter Notebook file with comments for each question.**
22. Is the data balanced? What is the percentage of no-shows in the data? If the data is not balanced, use one of the techniques we learned in the class to create a balanced data set.
23. Convert categorical variables into dummy indicator variables using get\_dummies function of pandas. (we did this in the class).
24. Drop columns which you think are unnecessary for prediction, e.g., patient id.
25. Split the data into testing and training sets using train\_test\_split function of sklearn.model\_selection library. (we did this in the class).
26. Use logistic regression and apply 5-fold cross validation on the training set. Report Accuracy, Recall, Precision, F1-score over all folds.
27. Comment on your results? Are they better than a naïve model that will predict every appointment as “show up”? If not, what might the problem? What kind of data might be helpful to improve your model?
28. Use the logistic regression model in the testing set and report AUC.
29. Comment on your results. Do you consider the result as good AUC performance?