ORIE 4741 Midterm Report; Medical NoShow

Kathleen Zhu (kz233) & Alice Li (al957)

1. Background

Due to social reforms that occurred throughout the 1970s in Brazil and changes made to its constitution, the Brazilian National Health Care System provides free healthcare to all its citizens. Despite this, however, the country is plagued by socioeconomic inequalities and as a result, healthcare varies tremendously by social class. On top of that, corruption and waste run rampant in a system that is already grossly underfunded. Our project hopes to address one of the sources of waste and inefficiency in the Brazilian healthcare system - medical appointment no-shows - by creating a model that will predict whether or not a particular patient with a specific appointment time will be a no-show. In doing so, we hope to motivate health care providers to take vigilant measures to follow-up with patients who are potential no-shows not only to reduce administrative wastes, but also to minimize the number of individuals who miss their care.

2. Raw Data

The dataset is obtained from Kaggle, posted by Joni Hoppen, who obtained it from the Municipality of Vitória. The raw data contains 15 features and 299806 observations. Each observation corresponds to a different appointment within the public healthcare system in the city of Vitória, Espírito Santo of Brazil between the years 2013 and 2015. The 15 features associated with each observation in the dataset are of a variety of data types. These features include the age and gender of the patient, information about the time and date the appointment was made, information about the time and date of the appointment itself, and some health conditions of the patient. The response variable is status, a binary indicator variable for whether or not the patient was a no-show.

3. Data Cleaning and Feature Engineering

Overall the data is of good quality, as there were no missing values; however, we removed 343 duplicate rows. We made minor tweaks to the data described as follows. There were several negative Ages, so we restricted **`Age`** to nonnegative values. For ease of computation and interpretation we converted **`WaitTime`**, the number of days between a patient making the appointment to the actual date of the appointment, from a negative value to a positive one. We convert all binary variables to one-hot encoding. For instance, for **`Status`** we changed the values ("No-show" or "Show-Up") to 0 and 1, respectively. For **`Handicap`** and **`SMS_Reminder`**, binary indicator variables, there were several instances where a positive indicator took a value greater than 1, (such as 2 or 4) but we restrict these values to 1. We also added a dummy variable **`NoShow`** which takes value 1 if a patient is a No-Show and 0 otherwise. Also we add dummy variables for gender: **`Male`** and **`Female`**. We create variables **`RegDate`, `ApptDate`, `ApptMonth`,** to track specific features of the dates.

4. Descriptive Statistics

Status (1 for a show-up and 0 for a no-show) shows that among the appointments within our dataset, approximately 30% were a no-show. The average wait time is nearly 14 weeks. Among patients who made an appointment, 57% received an SMS reminder regarding their appointment. The average age of the patients within this dataset was 38 years old. Gender is 1 for female and 0 for male. Of the patients in our data, 67% were female. Approximately 10% receive government stipends.

	Status	WaitTime	Sms_Reminder	Age	Gender	Scholarship
mean	0.697612	13.844876	0.572129	37.814644	0.668441	0.096916
std	0.459293	15.688672	0.494771	22.808208	0.470774	0.295844
min	0	1	0	0	0	0
25%	0	4	0	19	0	0
50%	1	8	1	38	1	0
75%	1	20	1	56	1	0
max	1	398	1	113	1	1

5. Data Visualization

We begin by investigating the relationship between time and No-Shows. The time plots below show the relationship between the 1) proportion of No-Shows and the date appointments took place and 2) the proportion of No-Shows and the date appointments were made. We observe that there doesn't seem to be a trend over time.

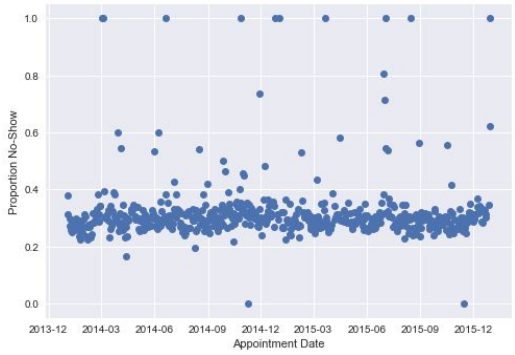

Figure 1: No-Show Proportion by Appointment Date

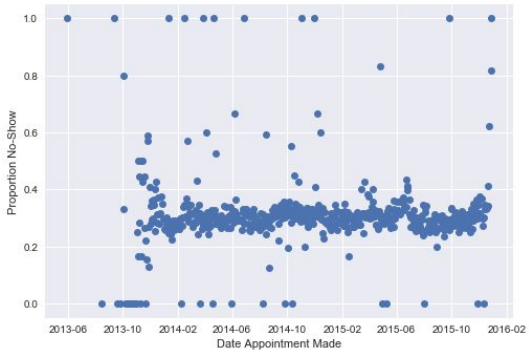

Figure 2: No-Show Proportion by Date Appointment made

Next we look at bar plots that show the proportion of No-Shows by the day of the week (very few appointments for Sunday and Saturday) and month. There only seems to be slight differences by weekday and by month.

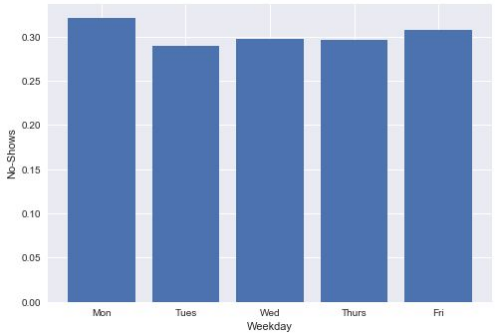

Figure 3: No-Show Proportion by Weekday

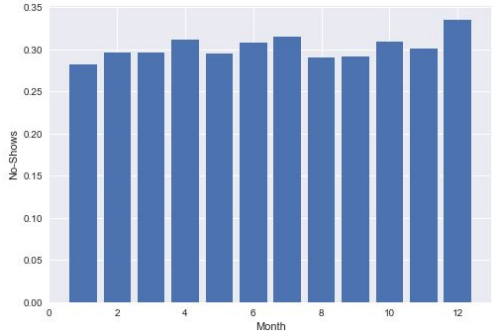

Figure 4: No-Show Proportion by Month

The correlation matrix below lets us quickly observe the relationships between No-Shows and the various binary indicator variables in the data. We observe that **`NoShow`** and **`Show-up`** are perfectly negatively correlated as are **`Male`** and **`Female`** by design. Variables like **`Smokes`** and **`Alcoholism`** are correlated, but there seems to be a lack of correlation between health conditions and showing up to appointments. There does, however seem to be a slight positive correlation between **`NoShow`** and **`Scholarship`**, and a slight negative correlation between **`NoShow`** and **`Hypertension`**. The barplot below shows that there is no significant difference in the proportion of No-Shows by gender, where red indicates No-Show.

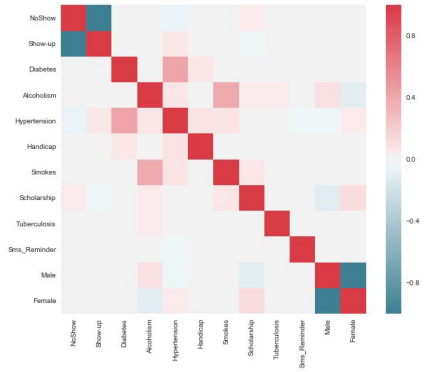

Figure 5: Binary Indicator Correlation Matrix

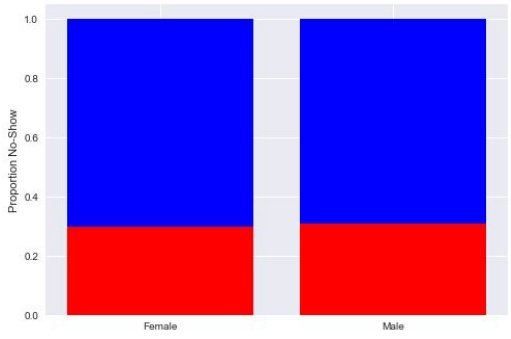

Figure 6: No-Show Proportion by Gender

Then we examine the relationship between Age and proportion of No-Show. In the stacked histogram below on the left, blue indicates Show-up and red again indicates No-Show. We see that older patients, those above 60, tend to have less No-Shows than younger patients. In the scatterplot below we look at the proportion of No-Show against wait times, which seems to have a slightly positive association. Therefore, **`Age`** and **`WaitTime`** seem to be a promising predictors.

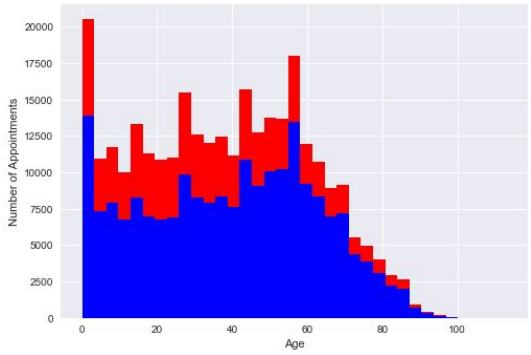

Figure 7: Distribution of Appointments by Age and No-Show

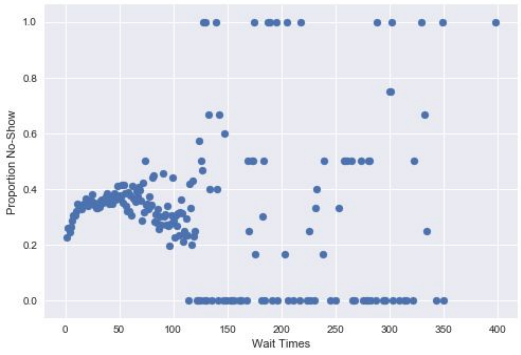

Figure 8: No-Show Proportion by Wait Times

6. Preliminary Model: Logistic Regression

As our goal is to solve a classification problem, for our first model we choose to implement a simple logistic regression to model the likelihood of a given appointment resulting in a No-Show. We begin by splitting our data into training and development sets, 80% and 20%, respectively. Then we trained a logistic regression model on the training set, without regularization. Based on the above descriptive, visual statistics, the features we chose to include in this initial regression were: **`Age, WaitTime, DayOfTheWeek, ApptMonth, Scholarship, Hypertension,`** and **`Smokes`**. The model is specified below, where the betas are the weights of the features, and the each of the days of the week has an associated dummy indicator variable, as it is encoded as a categorical feature.

$$logit\,(p(x)) = \beta_0 + \beta_{Age} * Age + \beta_{WaitTime} * WaitTime + \beta_{Mon} * I(Mon) + \ldots + \beta_{Sun} * I(Sun) + \beta_{ApptMonth} * ApptMonth + \beta_{Schol} * Scholarship + \beta_{Hyp} * Hypertension + \beta_{Smokes} * Smokes$$

After we trained the model on the training set, we tested it on our 20% held out development set. We obtained an initial accuracy score of: **0.69615**

7. Next Steps

- We want to further experiment with various combinations of features in our model, specifically with more attention to feature engineering
- We would like to use more rigorous validation methods by implementing cross-validation and k-fold validation. This way we can better ensure that our model performance isn't simply due to a fortuitous split of data.
- We want to improve our model performance by
    - Adding and experimenting with regularization terms
    - Developing a hybrid model, inspired by the work done by Alaeddini[1] et al.
    - Investigating other models, like decision trees and subsequently random forests

[1] Adel Alaeddini, Kai Yang, Pamela Reeves & Chandan K. Reddy (2015) A hybrid prediction model for no-shows and cancellations of outpatient appointments, IIE Transactions on Healthcare Systems Engineering, 5:1, 14-32