1. **ABSTRACT**

**Objective:** Using binary classification systems, we trained, developed, and compared several models in predicting both the binary and probability of outcomes.

**Methods:** Data was collected from a clinic in the Rio De Janeiro of Brazil. The clinic services more than 80 neighborhoods in the region. The data was downloaded directly from Kaggle [1].

**Results:** After training our model, we managed an error rate of roughly 21% with a KNN Classifier for predicting no-shows with the dataset’s features. We found the best predictors to be the patient’s Neighborhood, Age, and lead time (days between scheduling and appointment date).

**Conclusion:** We found that the KNN model accuracy and ROC score to be convincing enough to be used in real-life applications. We also find that that sending SMS reminders to patients did improve appointment attendance and can be used in conjunction with this model.

1. **INTRODUCTION**

No-shows appointments have serious ramifications on a provider’s daily productiveness and revenue. Additionally, research shows that patients with frequently missed appointments demonstrated worse health care outcomes [2]. As a whole, in the United States, the healthcare industry is losing around $150 billion a year to no-shows alone [3]. On average, no-shows cost providers $156,000.80 [3].

Given these current challenges, finding ways to improve the productivity of scheduling and increasing the likelihood of patients showing up to their appointments could prove to be beneficial to clinics worldwide.

The ability to predict a patient’s show-up tendencies could be particularly useful. This data analysis hopes to build multiple classification models to assist in identifying these tendencies. Additionally, the paper explores the trends in the no-show data based on multiple factors and identifies the key factors influencing patient no-show tendencies.

1. **MATERIALS AND METHODS**

*Dataset and Subjects*

Data used in the analysis was selected from the website Kaggle. Data was collected from a Rio De Janeiro province clinic in Brazil. It contains 110527 patient rows and 14 associated characteristics, some of which include gender, scholarship, hypertension, Diabetes, Alcoholism, Handicap, whether or not an SMS was received, and the outcome variable: no-show. Patient age was nearly distributed equally.

*Variable Preparation*

In addition to the existing variables in the Kaggle dataset, we create the following variable to consider in our model:

Days Between (Lead Time): which is the time difference in days between the date of visit and date of scheduling.

Consequently, the following columns were dropped from our model:

1. Scheduled Date
2. Appointment Date

Additionally, we drop the following patient unique identifiers as they would be conflicting with our model (each id is affiliated with one row and as a result would over-fit our model):

1. PatientId
2. AppointmentID

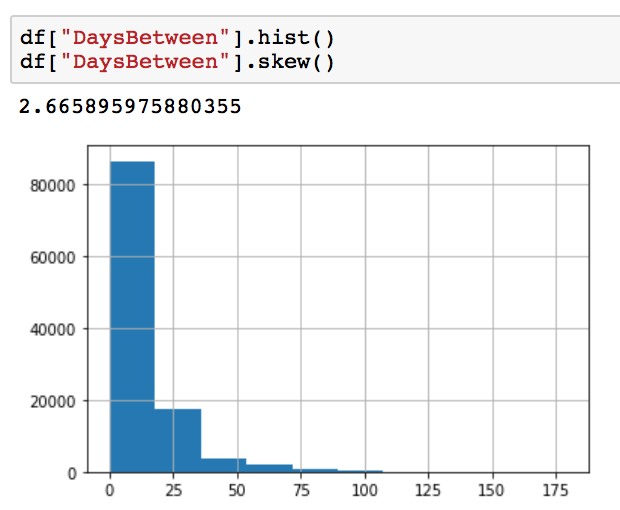
The “Scholarship” variable is a concept that exists in Brazil and is essentially a government welfare system that either partially or fully covers health costs. For clarity reasons, we renamed this variable “Welfare”.

*Data Pre-processing*

Prior to analysis, we checked the data for null values: none existed. However, several negative data points did exist, particularly with the “Days Between” variable. This is likely attributed to a mistyping error upon input into the EHR. We decided to drop all negative values as they logically do not make sense in the context of this dataset. Additionally, the insignificant quantity of negative values was bound to not have an overall impact on our analysis.

Our data contains only two quantitative variables, of which only one appeared to be skewed: Days Between. This likely due to the observed overwhelming number of same-day scheduled appointments. Of the typically used transformation methods, we found the log transformation the most practical method for effectively transforming the data to a normal distribution.

**Figure 1**



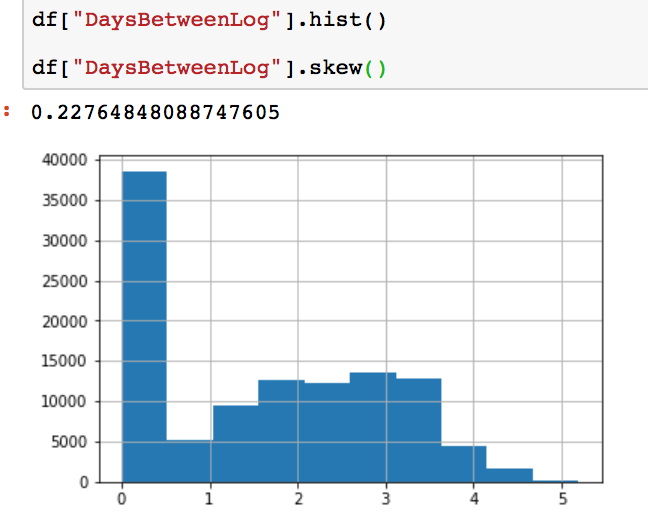


Figure 1 (top) shows the distribution prior to transformation (skewness score = 2.67). Figure 1 (bottom) shows the distribution after log transformation (skewness score = 0.23). The skewness score prior to transformation was indicative of highly skewed data, whereas the newly transformed variable showed low enough skewness scores to be considered normally distributed.

For more additional data analysis, we chose to bucketize DaysBetween (lead time) into 7 bins. These size of these bins were based on both logic and histogram evaluation.

*Data Exploration and Statistical Evaluation*

Our dataset included 88208 patients who showed up to their appointment and 22314 no show patients. This indicated that about 20% of patients do not show up to their appointments, which is around the same rate that people in the U.S do not show up to their appointments [3]. Figure 2 (below) shows the data in detail:

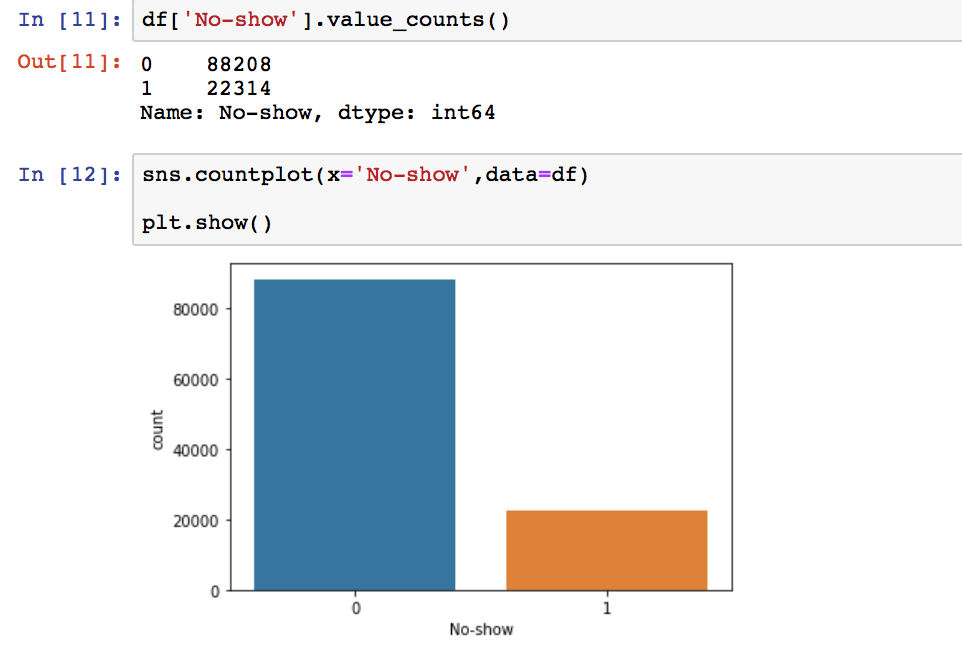
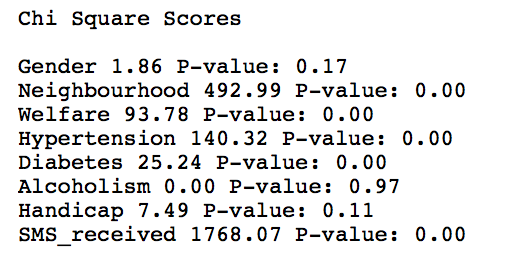
**Figure 2**

Figure 2 displays the number of no-shows marked by the binary value ‘1’ plotted alongside the number of patients that did show up indicated by the binary values ‘0’.

We individually tested variables for independence with our dependent variable of ‘No-Show’ using a chi-square test for all categorical variables. We chose to enter variables with a *P* value less than 0.2 into our model. Figure 3 (below) shows the results of the Chi-Square test.

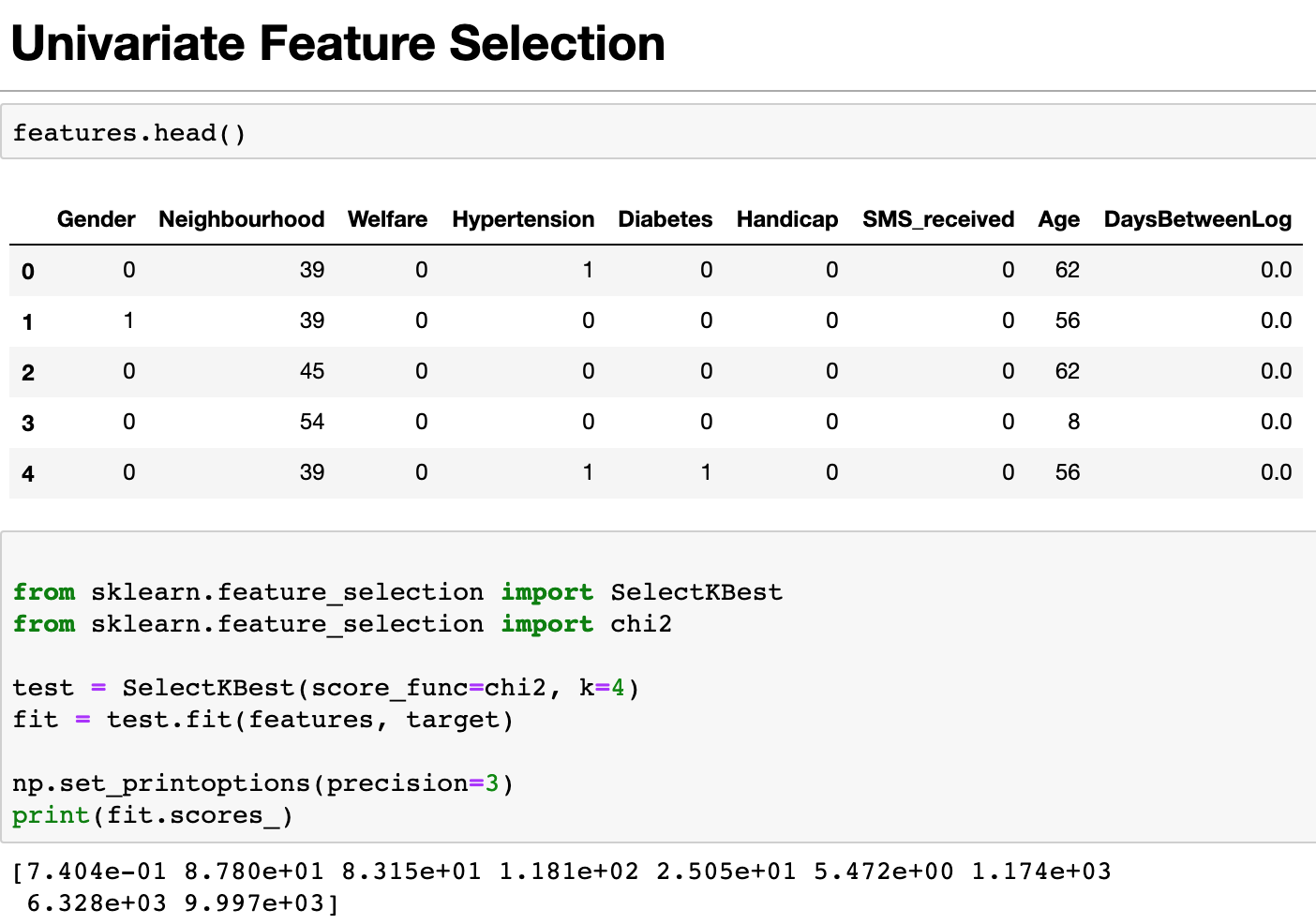
**Figure 3**



Consequently, the column ‘Alcoholism’ was not entered into our model.

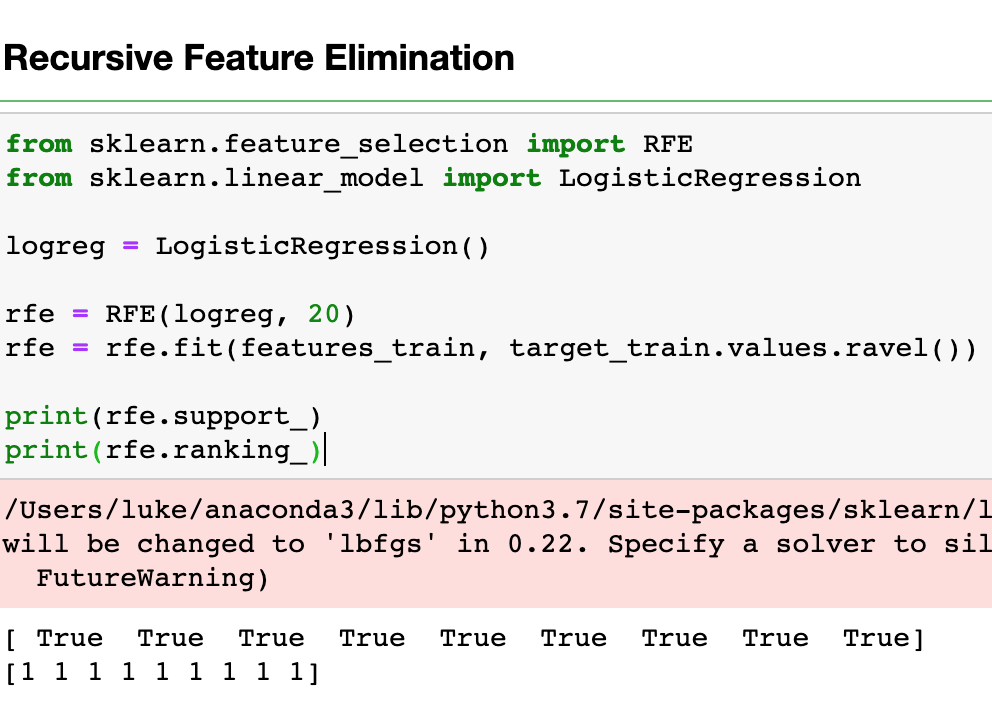
For the remaining of our variables, we used two feature selection methods. The first is outlined below in Figure 4.

**Figure 4**

****

We performed a Univariate Feature Selection to make sure all remaining features had a strong relationship with the outcome we were trying to predict (No-Shows). The scores for each variable in the feature selection can be shown in figure 4. We found all the features to be significant to our model.

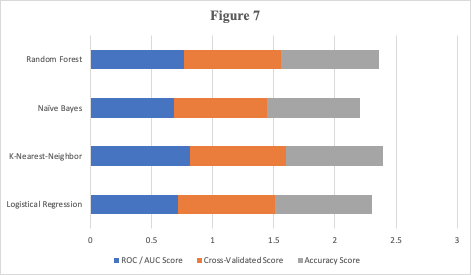
After making sure our variables all had a strong relationship to the outcome, we used Recursive Feature Elimination to see if any of them were too weak of predictors to be a good feature. All of our features passed, which is shown in Figure 5 the recursive elimination so we moved forward with this set as our predictors.

**Figure 5**

*Model Development*

We randomly split 80% of the data into a training set and 20% of the data into a testing set. Additionally, we performed 10-fold cross-validated as a method of reducing selection bias.

Furthermore, we fit our data into the following models:

1. Logistical Regression: A logistic function that models both the binary outcome of the dependent variable and the probability of that outcome.
2. KNN: A classification model that is flexible and can handle multi-class cases.
3. Naive Bayes: An easy and fast to predict the class of test data set. It performs well with categorical variables compared to numerical. This classification assumes independent predictors.
4. Random Forest: A classification model consisting of many decision trees, which uses feature randomness to create a forest of uncorrelated trees which is more accurate than any one tree alone.

*Further Analysis*

We used an ensemble learning algorithm, very similar to Random Forest, that randomizes subsets of data. The ExtraTreesClassifier() allowed us to evaluate the top features in terms of importance.

1. **RESULTS**

*Model Results*

After fitting the four models with the testing data, validating on the training data set, and running it through our the 10-fold cross-validated method, we obtained the following scores:

**Figure 6**

|  |  |  |  |
| --- | --- | --- | --- |
| Classifier | ROC / AUC Score | Cross-Validated Score | Accuracy Score |
| Logistical Regression | 0.714726025 | 0.79614321 | 0.795298 |
| K-Nearest-Neighbor | 0.815631007 | 0.788431628 | 0.791325887 |
| Naïve Bayes | 0.685698092 | 0.758669402 | 0.760340235 |
| Random Forest | 0.763599592 | 0.797106008 | 0.797121092 |

As a quick overview, we’ve included Figure 6. This table shows the ROC / AUC Score, Cross-Validated Score, and Accuracy Score each of the 4 models we trained with this data.

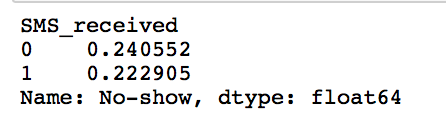
As shown in Figure 7, no single model takes the lead in total by a large margin, but KNN seems to be the most well-rounded model we fitted the data to. It’s ROC / AUC is certainly the highest at nearly 81.6%.

*Additional Data Analysis*

Our initial hypothesis was the SMS notifications would be an important factor in determining if a patient would should up to their appointment or not. Diving deeper into the data, we found that patients who received an SMS notification, as a group, were less likely to show up to their appointments. This seemed odd, but the underlying reason is that only patients who had longer lead times were getting notified more frequently. Those who scheduled a same-day appointment were not notified by text messages. This was skewing our perceived impact of text messages as method of increasing patient show-up rate.

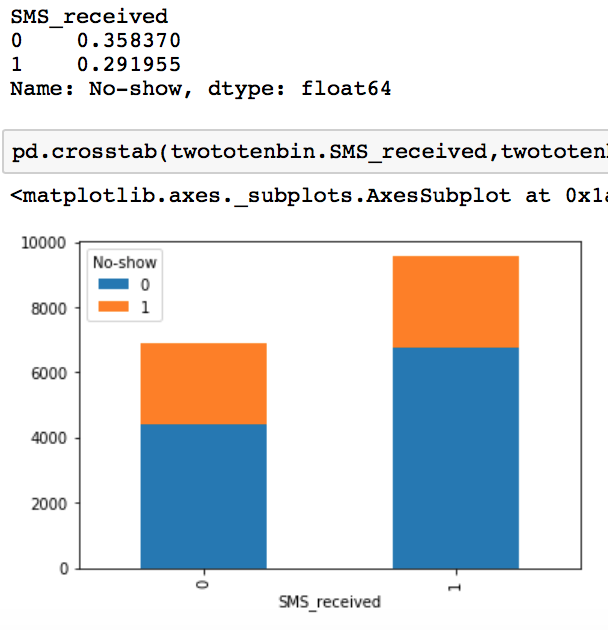
We controlled for this by isolating the data frame to only those who had lead times between two and five days. What we initially hypothesis was confirmed. Those who had received an SMS notification and waited more than 1 day in between, were, in fact, more likely to show up to an appointment than those who had waited more than 1 day and not received the SMS notification. The results of these findings are found below (Figure 8).

**Figure 8**



On average, there was a 2% difference in the show-up rate between those who received an SMS and those who did not. We also found that as the lead time increased the effectiveness of SMS\_received increased. Figure 9 (below) displays the 10 to 22 days bucket statistics.

**Figure 9**



A near 7% difference in show-up tendency is seen when isolating for a lead-time bucket. In conclusion, when controlling for lead time, SMS\_received encouraged patients to come to their appointments.

*Predictors*

Using the Extra Tree Classifier built-in feature importance method we were able to generate our top 3 predictors for no-show outcomes:

1. Neighborhood
2. Age
3. DaysBetweenLogged (log of the lead time)
4. **DISCUSSION**

Patient neighborhood was the most important predictor in our model, but it also exposes a quirk in our research. Since the data was taken at a Brazillian clinic, our data, and consequently our model, is localized to Brazil.

Several factors could influence why neighborhood was the most influential factor. The most obvious being the relative distance between where the patient lives and the clinic. The next, and often overlooked factor, is the socioeconomic status of each of these neighborhoods.

Age is our second most potent predictor in this model. We found that younger patients are more likely to miss an appointment then middle-aged ones. The older you get, up until a certain point, the more likely you are to make it to your appointment.

DaysBetweenLogged is a value we calculated from the scheduled day and appointment day. It’s the third most important predictor (to our outcome). The general rule of thumb is the farther away an appointment date is from the scheduled date, the less likely a patient is to show up.

Once expanded to be more universal, this model can be utilized by healthcare infrastructure to predict whether a patient is likely to miss an appointment or not. This is a useful capability for this industry to have as no-shows cost countless amounts of lost resources and capital in the medical field.

As an example of how this model can be used, at the end of the code, we’ve included a test run of our model based on random data to predict the likelihood of no-show rate. We find it interesting and potentially useful to include both the binary predicted outcome and the predicted probability outcome.

Moving forward, this research and model can be used to develop applications designed for use by medical offices to predict at-risk patients. Once identified, this system could send automated SMS or email notifications to patients above a certain likelihood threshold. This could potentially save a lot of time and resources for institutions that provide medical expertise. The less time and resources these firms spend on missed appointments the higher quality and quantity of healthcare they can provide overall.

1. **REFERENCES**

[1]

Hoppen, J. (2019). *Medical Appointment No*

*Shows*. [online] Kaggle.com. Available at:https://www.kaggle.com/joniarroba/noshowappointments/data [Accessed 8 Dec. 2019].

[2]

Nguyen, Douglas L, et al. “Missed Appointments

in Resident Continuity Clinic: Patient Characteristics and Health Care Outcomes.” *Journal of Graduate Medical Education*, The Accreditation Council for Graduate Medical Education, Sept. 2011, www.ncbi.nlm.nih.gov/pubmed/22942961.

[3]

Ekram, T. (2019). *How Much Are Cancelled*

*Appointments & No-Shows Costing You? - Luma Health*. [online] Luma Health. https://www.lumahealth.io/blog/how-much-are-cancelled-appointments-no-shows-costing-you/ [Accessed 10 Dec. 2019].

[4]

Mohammadi, I., Wu, H., Turkcan, A., Toscos, T.

and Doebbeling, B. (2019). *Data Analytics and Modeling for Appointment No-show in Community Health Centers*. NCBI.