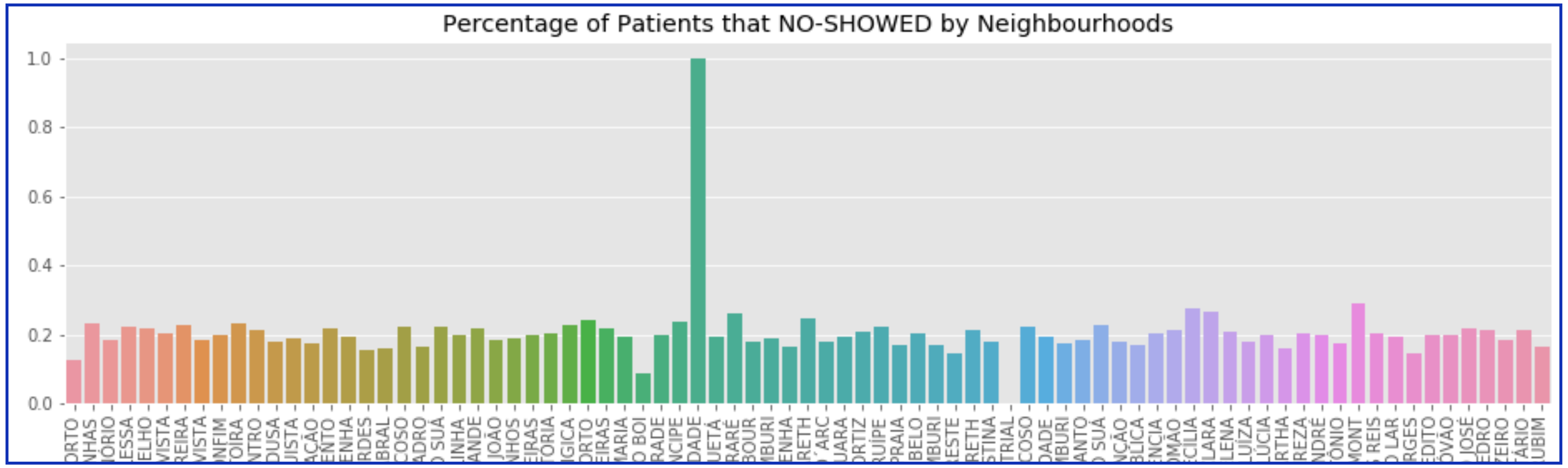
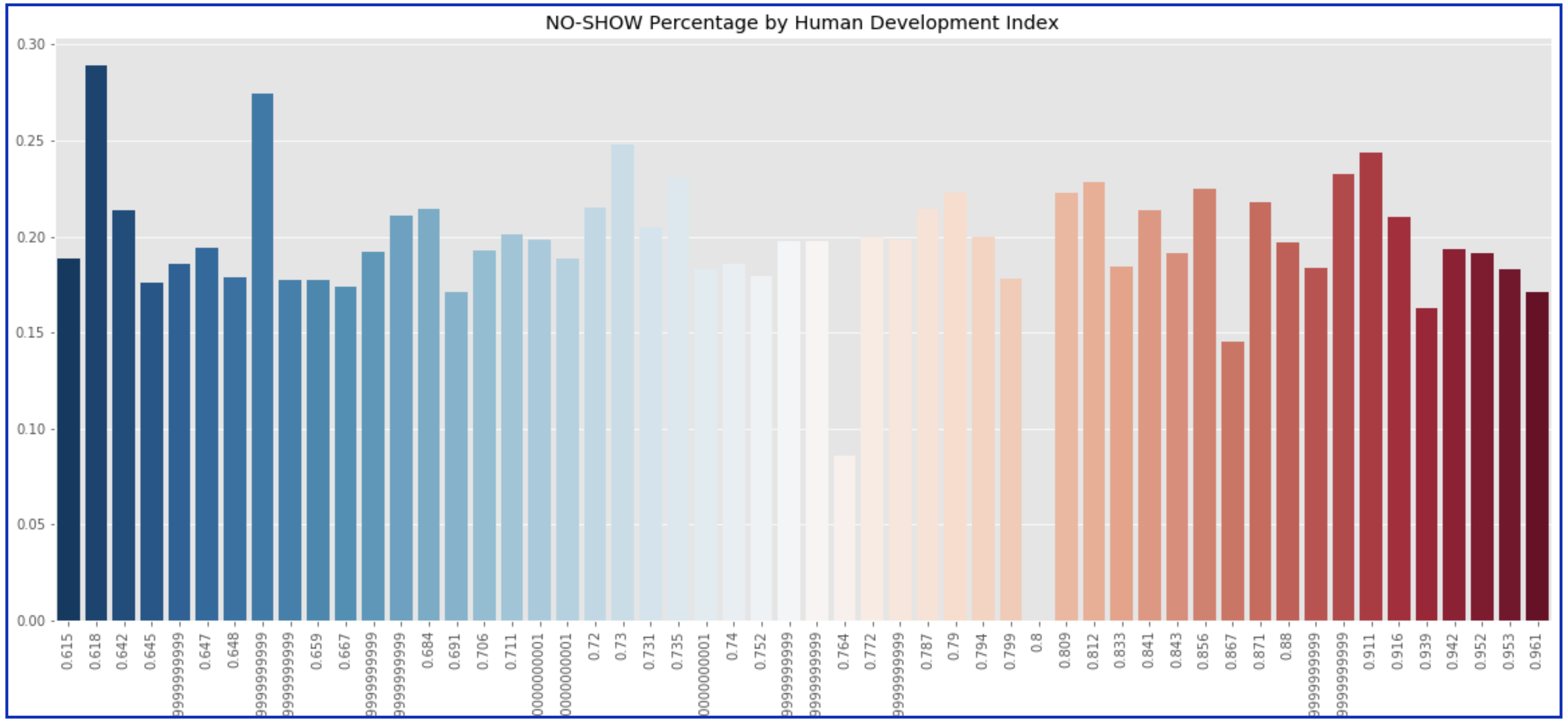
Predicting Appointments

Show vs. No Show across hospitals in Brazil

July 2020 | Chibz + Tom © ph[0]ton

Contents:





Objective:

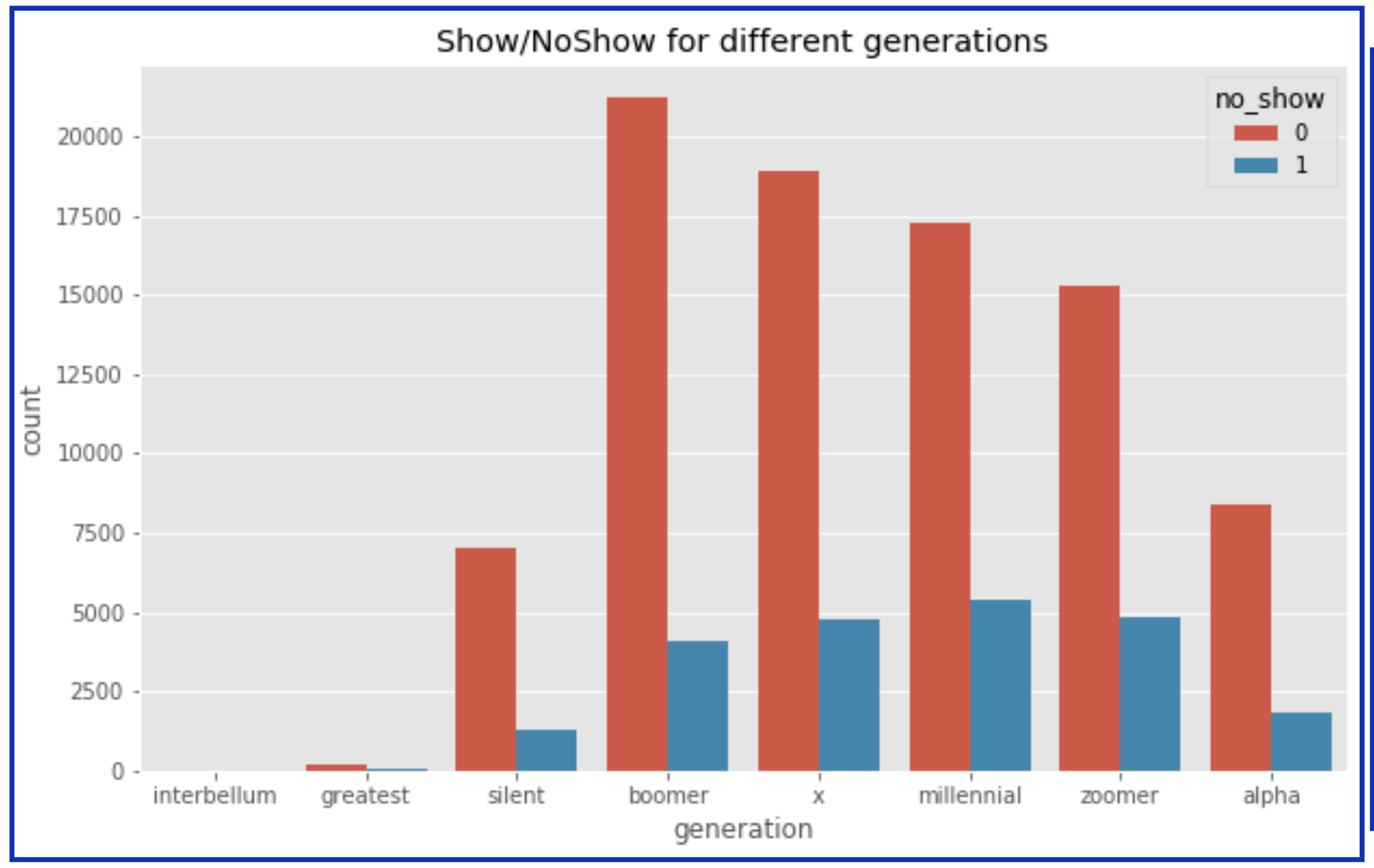
- 20% of hospital appointments across 80 neighbourhoods in Brazil are a No Show
- The average of Show vs No Show is consistent across all 80 neighbourhoods
- We included the HDI from a separate dataset and used this to confirm that regardless of standard of living, the averages remain the same.
- Can we make a model that will predict if a patient will not show up for their appointment?

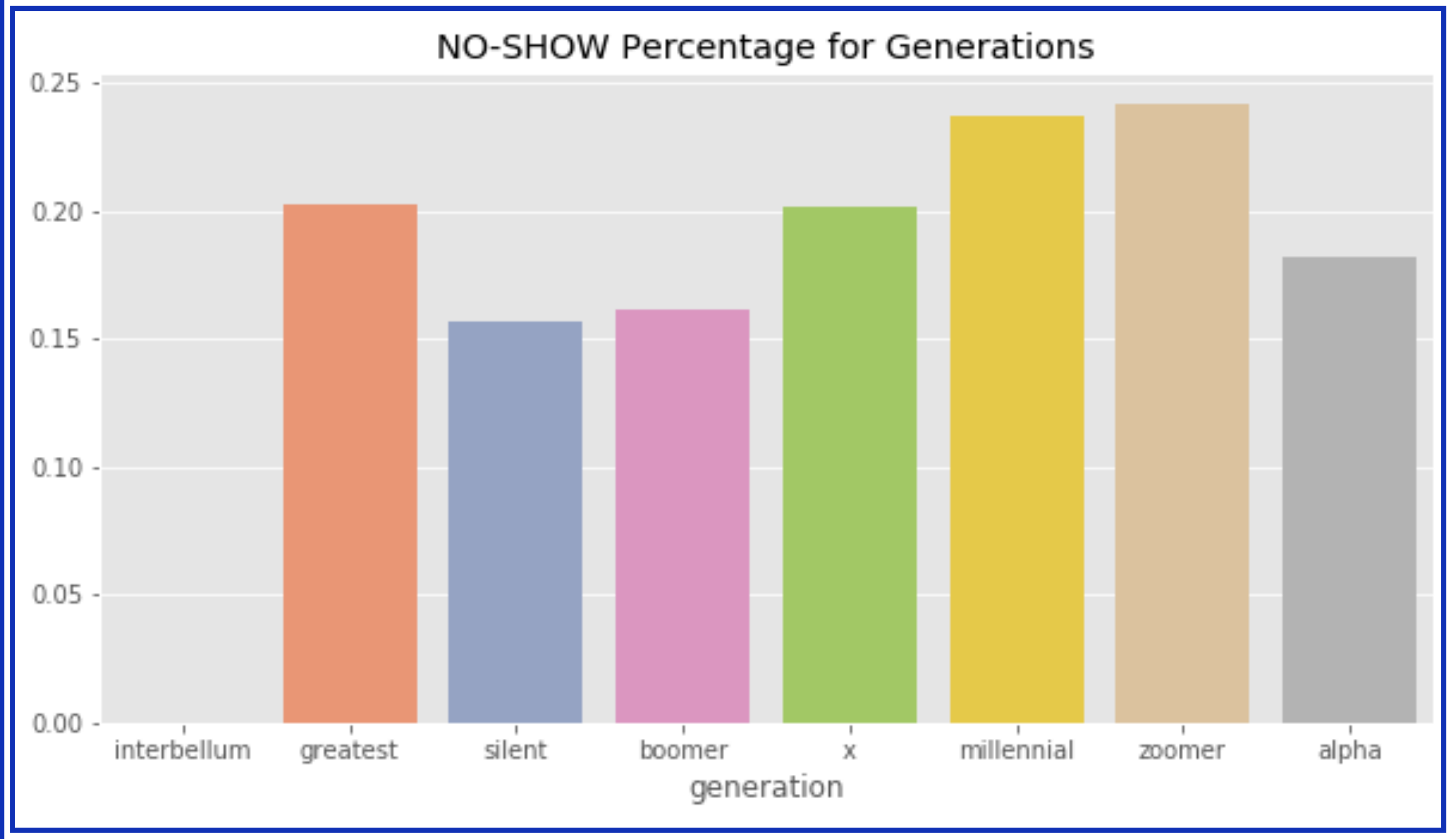
Data:

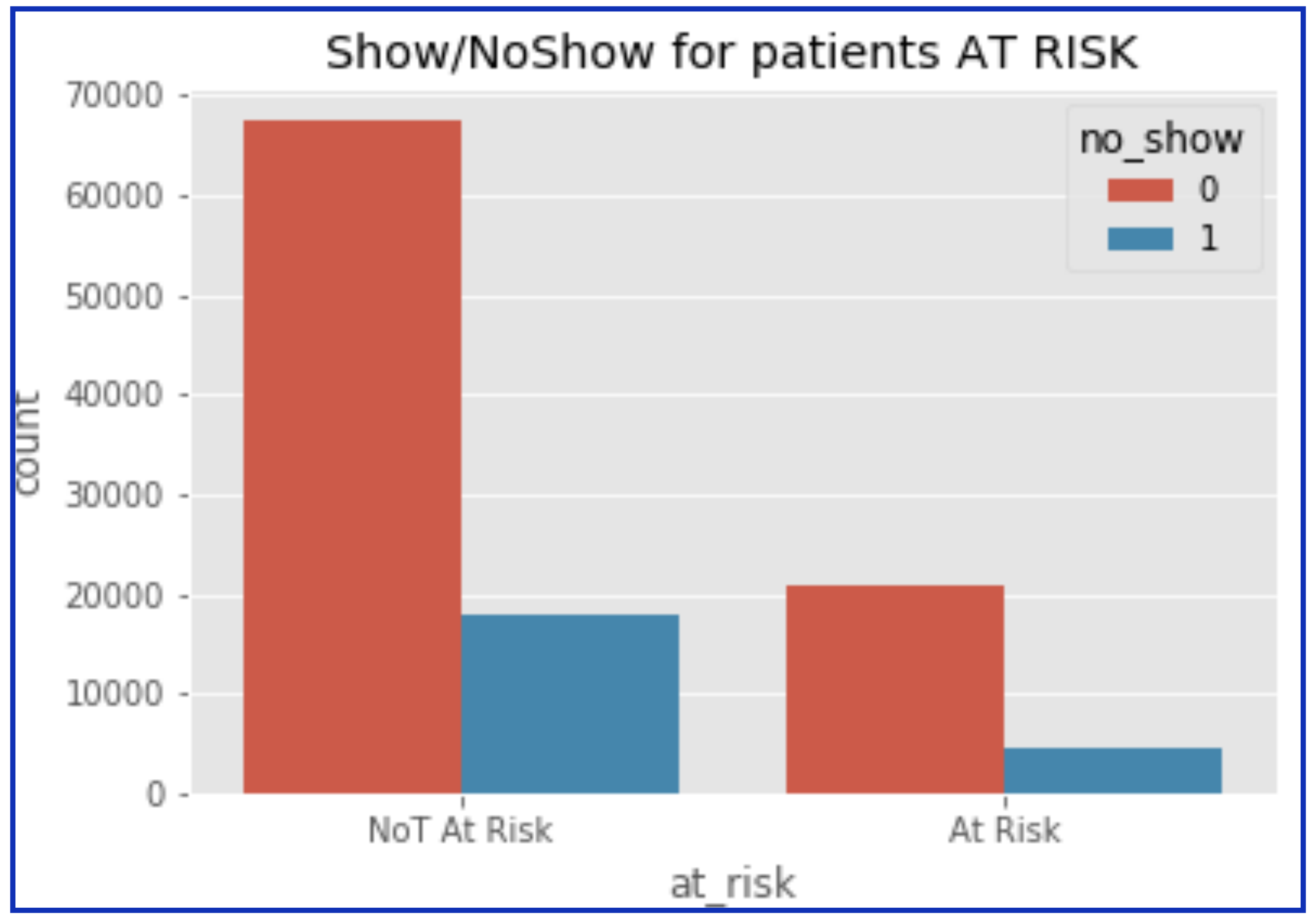
- Datasets obtained from Kaggle*
- Range of data: Apr 2016-Jun 2016
- Number of Observations: 110,527
- 14 Features
- Rows consisted of patients with appointments
- Features consisted of patient information such as age, illnesses, scholarship, scheduled and appointment days

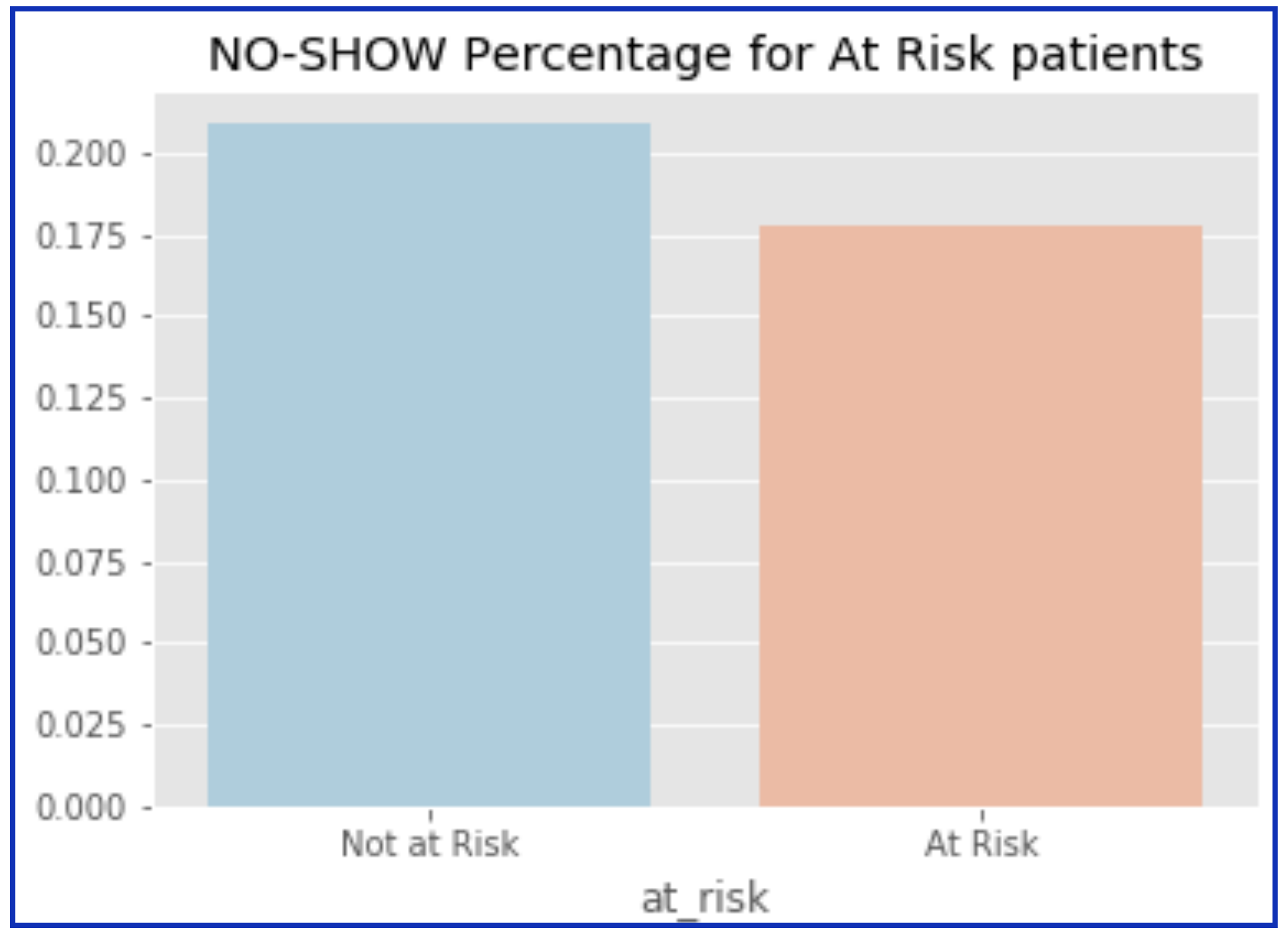
Methodology:

- Used Numpy and Pandas for data manipulation
- Used matplotlib and Seaborn for data visualizations
- Used SciPy, statsmodels, and scikit learn for statistical tests and base models
- Used XGBoost for the final model







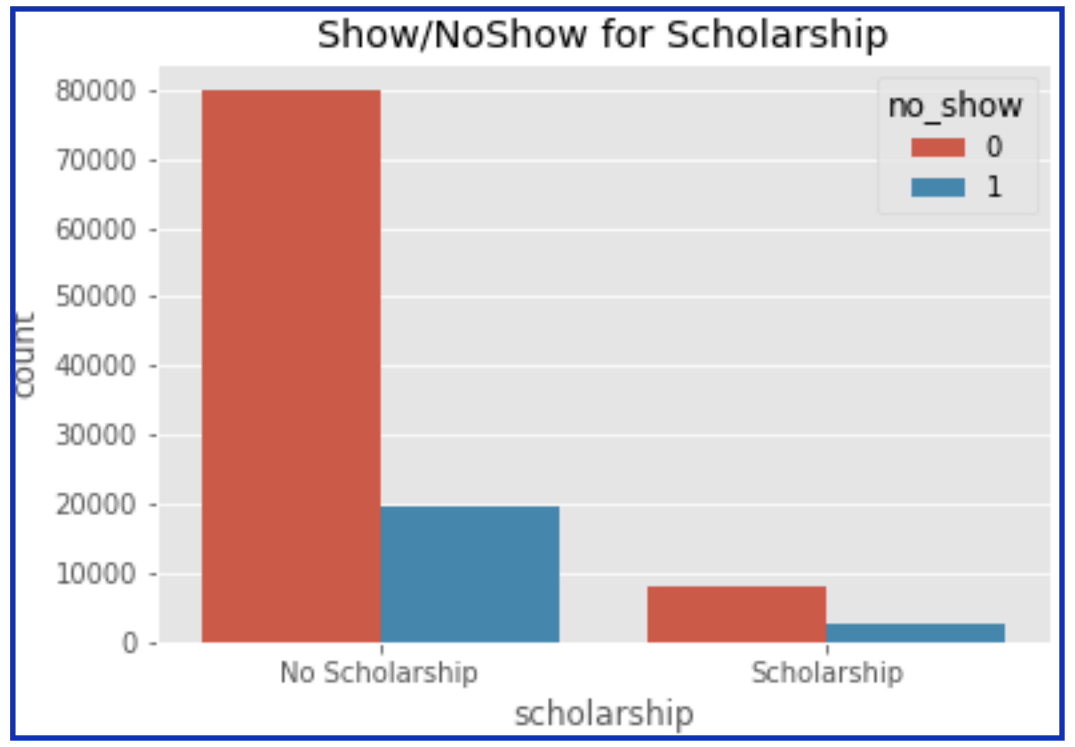


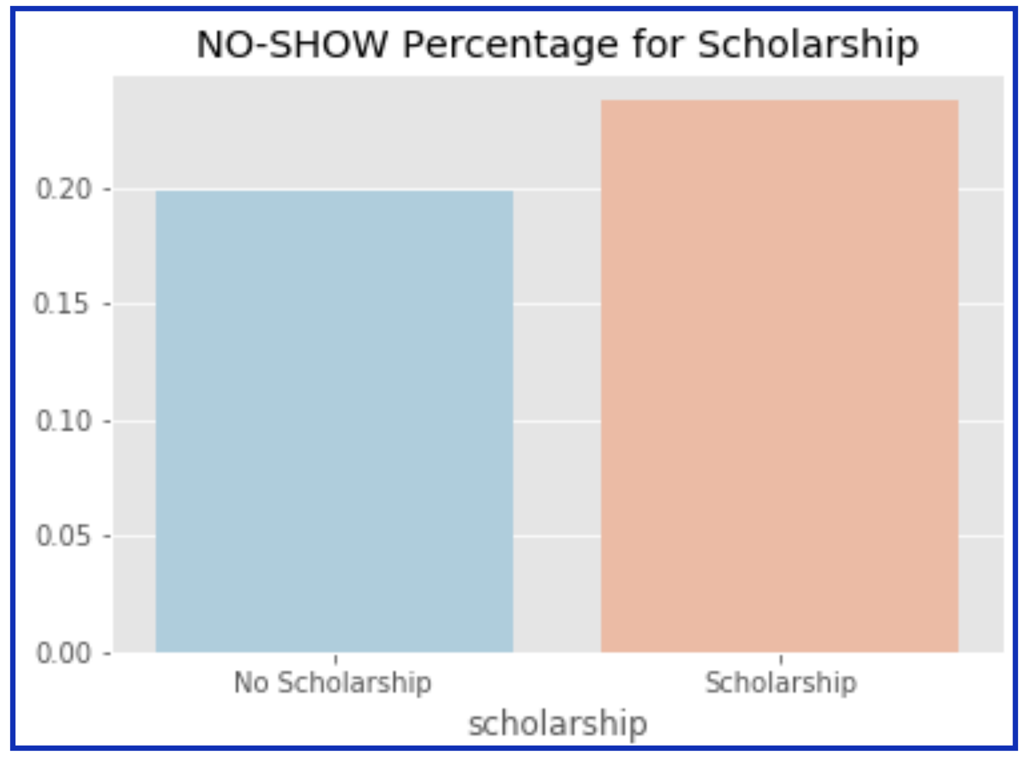
EDA:

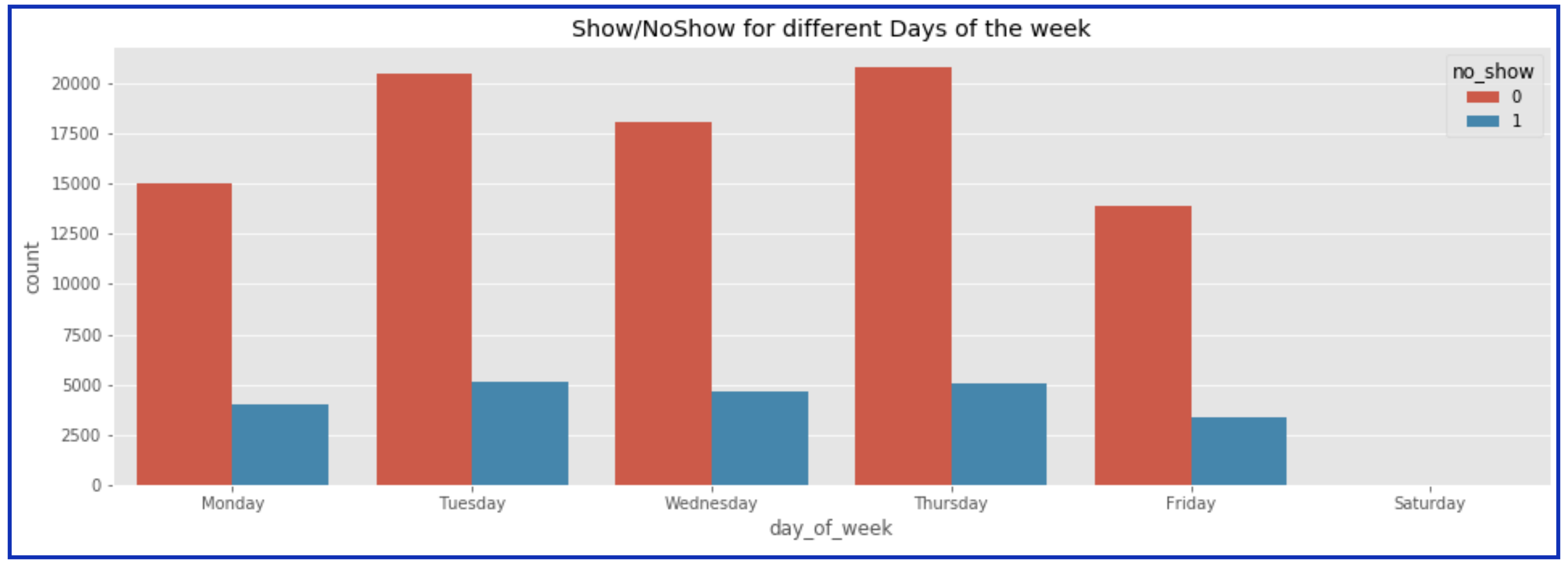
- In order to find the features that would contribute most to our predictive model we started with the following:
- Looking across generations, we found a significant difference in the populations as well as a cool “X” trend.
- Additionally, we looked at a feature we created called “At Risk” and found a significant difference there.
- “At Risk” is defined as a patient having more than one of the conditions present in the dataset: Hypertension, Alcoholism, and Diabetes

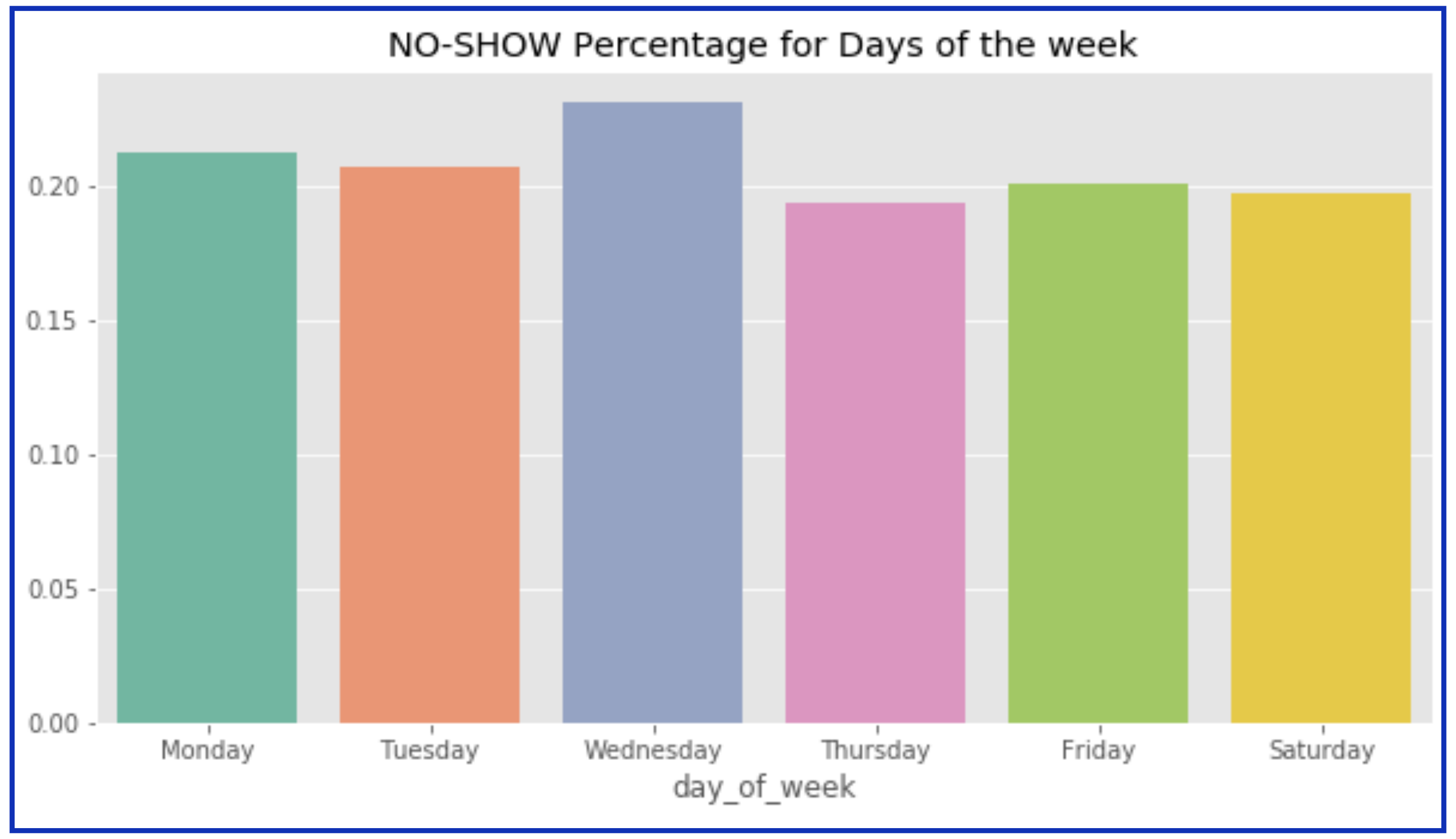
EDA (cont):

- When looking at scholarship, while there is a significant difference, it's opposite of what we assumed in that those WITHOUT scholarships would have a higher rate of no show
- Finally, we looked at the day of the week and found there is a significant difference of show vs no show
- Of note: SMS reminder had the opposite result than what we expected
- This is most likely due to the fact that we do not know how SMS reminders work in this dataset, so we could not use this feature









Model	ACC	F1
Clean Data		
KNN	0.78	0.12
Logistic Regression	0.6	0.35
Decision Tree	0.56	0.36
Random Forest	**0.63**	**0.35**
Feature Data		
KNN	**0.77**	**0.87**
Logistic Regression	0.57	0.35
Decision Tree	0.58	0.67
Random Forest	0.57	0.65
Grid Search		
Random Forest	**0.77**	**0.86**

Models:

- First we ran 4 models against our cleaned data
- Then we ran the model against our data after feature engineering
- Finally, we ran a grid search on the Random Forest model
- Ultimately, given that predictive power is more important than interpretability, we decided to go with a different model altogether for our final.
- Of note: the KNN feature model was run after our final model and as such was missed as something we should have explored further
- (That one is on Tom)

Results:

- We chose an XGBoost model for our final model
- We decided that a predictive model would be more important to a hospital as opposed to interpretability
- The Accuracy and F1 of the XGBoost model was the highest out of all of our models

```
In [214]: gs_preds = gsearch1.best_estimator_.predict(X_testGBCV)


          gs_test_f1 = f1_score(y_testGBCV, gs_preds)
          gs_test_acc = accuracy_score(y_testGBCV, gs_preds)

          print("Accuracy: %f" % (gs_test_acc))
          print("F1: %f" % (gs_test_f1))

          Accuracy: 0.799609
          F1: 0.888625
```

```
In [211]: gsearch1.best_params_

Out[211]: {'colsample_bytree': 0.3,
           'learning_rate': 0.01,
           'max_depth': 3,
           'min_child_weight': 2,
           'n_estimators': 1000}
```

ph[0]ton

Conclusion:

- Our model is better at predicting No Shows than just guessing the Dominant Class
- We did find a few limitations in our ability to predict, that if given more data, we feel we could be even more accurate.
- One of the limitations we discovered exploring the data is the time frame the data spans.
- In addition more context was needed for certain features to better explain, example; SMS received.
- Finally context to the reason for the appointment would have been also a helpful tool.


PISO 1
ASCENSOR
SANITARIO
BANCO DE SANGRE
TERAPIA NEONATAL
TERAPIA ADULTOS

thanks