Clinical Data Analysis of Hospital Appointment No-Show Rate for Outpatients

Aarthi Anbalagan (aanbalag), Katie Hsia (khsia), Zhe Zheng (zzheng15)

BROWN

Motivation

- 30% of the patients DO NOT show-up in hospital appointments.
- Negative impact and under-utilization of health care system.
  - Why patients do not show up? Deep diving and analyzing.
  - Evidence-based predictive model for patient no-shows.
  - Calculated overbooking approach to avoid the under-utilization cost being passed onto other patients, increasing the healthcare costs.

Datasets

(Initially earmarked data from data.gov and healthdata.gov are all removed from federal websites.)

1. Raw data from Kaggle on medical appointment no-shows. 300,000 entries and 15 columns including age, gender, appointment date, does the patient carry a certain disease and etc.
2. Twitter data for patient sentimental analysis on certain hash tags: #hospital, #appointment, and #hospital appointment.

Methodology

1. Scrapping twitter data with hashtags: hospital, appointment and hospital appointment.
2. Cleaning the patient no-show dataset and twitter dataset.
3. Correlation and exploratory data analysis on both datasets.
4. Twitter dataset: sentimental analysis.
5. No-show dataset: bar charts, line graphs, scattered points, linear regression, K-clustering, time series analysis(lag-plot) and etc.
6. Classifiers for predicting whether patient will show-up.
7. Answer when can the hospitals overbook.

Technology Used

Classifiers Tried/Used

Bayesian Classifier (Gaussian and multinomial), Logistic Regression, SVM (linear), GradientBoostingClassifier, MLPClassifier, Random forest classifier, Decision tree classifier, K nearest neighbor classifier, Bagging classifier, One vs. Rest Classifier

Neural Network Used

Keras binary classifier

.csv Keras Id

python D3.JS - Data Driven Documents +tableau TensorFlow scikit learn

Visualizations

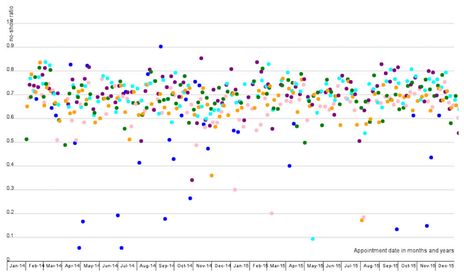

Figure 1. Interactive visualization for analyzing patient no-show ratio based on days of the week.

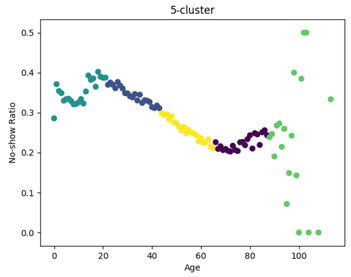

Figure 2. Analyzing correlation between patient's age vs. no-show ratio using K-clustering algorithm.

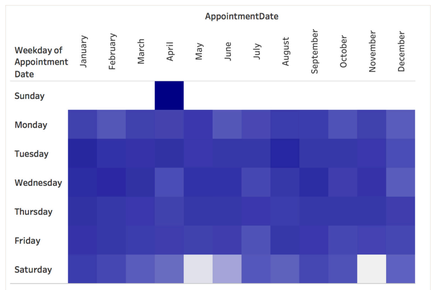

Figure 3. Analyzing patient no-show probability with respect to multiple data such as date of appointment, day of week, month.

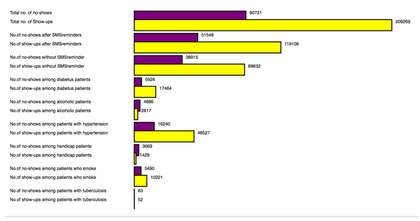

Figure 4. Part of the feature-label correlation analysis.

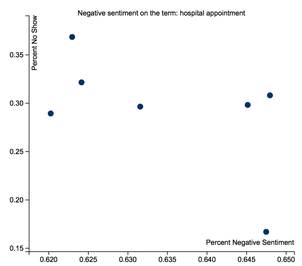

Figure 5. Combine Twitter sentiment analysis and no-show data.

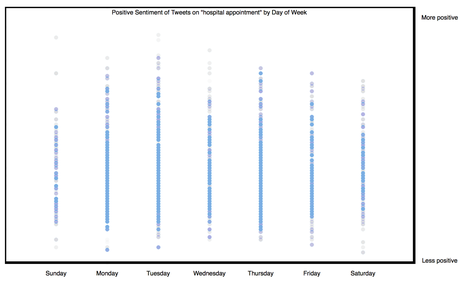

Figure 6. Plots positive or negative sentiment about Twitter keywords based on day of the week.

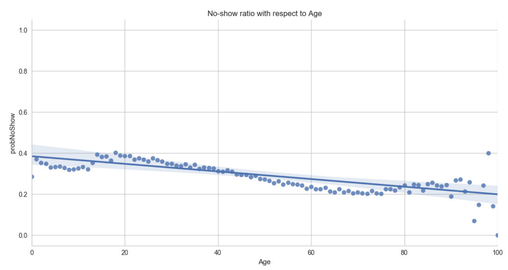

Figure 7. Correlation analysis between label and age feature with linear regression.

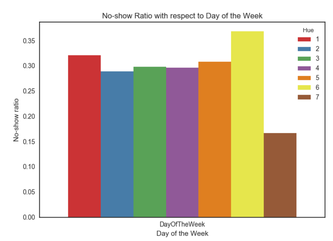

Figure 8. Correlation analysis between label and day of week (appointment) feature.

Machine Learning

- Linear regression and K-clustering: analyze the correlation between a certain feature and the no-show ratio.
- Feature Engineering: As per the initial data correlation analysis using box-plots, bar graphs, and exploratory data analysis using linear regression and k-clustering, we decided to drop few features which has least weightage in our prediction accuracy like 'AppointmentDate', 'AppointmentRegistration','GapNoOfDays', 'Tuberculosis' and 'AwaitingTime'.
- Classification: We are dealing with classification (whether patient will show up or not) rather than regression here.
- Parameter Tuning: SVM (kernel, number of estimator), Neural network (number of neurons on each layer, number of layer, batch size and etc.), Gradient Boosting Classifier(number of estimators, learning rate, max depth).
- Results: Top Five Classifiers
  - Accuracy for Keras Classifier: 0.7120
  - Accuracy for MultinomialNB: 0.7022
  - Accuracy for Linear SVC: 0.7019
  - Accuracy for Logistic Regression: 0.7019
  - Accuracy for GradientBoostingClassifier: 0.6979

Conclusion

- We were able to predict whether patients would show up with a maximum accuracy of 71%.
- Limited to the information and features from the dataset, we need more data in order to achieve higher accuracy. Only two features have strong correlation to the label as shown in Figure 7 and Figure 8.
- In terms of recommending overbooking approach, we had pretty interesting insights from all the visualizations as to when and how can the hospitals can actually overbook.

Acknowledgements

- Background information
  - Lacy, Naomi L. et al. "Why We Don't Come: Patient Perceptions on No-Shows." Annals of Family Medicine 2.6 (2004): 541–545. PMC. Web. 1 May 2017.
  - Huang, Y., and D.A. Hanauer. "Patient No-Show Predictive Model Development Using Multiple Data Sources for an Effective Overbooking Approach." Applied Clinical Informatics 5.3 (2014): 836–860. PMC. Web. 1 May 2017.
- Data
  - Twitter, Kaggle, text-processing.com
- Technical Support
  - bl.ocks.org, stackoverflow.com
- Support
  - The Computer Science Deparmtent, the Professors, and the TA staff