Architectural Decisions Document: Predicting No-Shows at Health Clinic

Based on nature of project, the order of the ADD was altered slightly

1) Data Source
   a) Main dataset (containing patient and appointment data) obtained from Kaggle (https://www.kaggle.com/joniarroba/noshowappointments/kernels)
   b) Supplementary data:
      i) Weather data for relevant days obtained from https://www.timeanddate.com/weather/brazil/vitoria/historic
      ii) Location data from Google Maps.
      iii) Supplementary data sources were chosen for their ease of access.
2) Enterprise data:
   a) Not relevant to this project.
3) Streaming analytics
   a) Streaming analytics not applicable for this project, as there is no continuously updating dataset.
4) Data repository:
   a) All data stored locally and on Kaggle.
   b) Because the data are already stored on Kaggle, there is no need for an additional Cloud solution.
5) Data Integration:
   a) Data Transformation:
      i) Transformed no-show and gender data into binary format for visualization and model ingestion
      ii) Schedule and appointment dates were converted to datetime from which meaningful features could be extracted for both appointment and schedule days:
         (1) Month of year
         (2) Day of year
         (3) Day of week
         (4) Hour of day (only available for schedule date)
      iii) Extracted patient and location frequency information (how many times each patient came to the clinic and how may appointments from each neighbourhood were made)
      iv) Incorporated latitude/longitude and climate data (temperature high/lows and condition, containing 4 categorical values)
   b) Data Cleansing:
      i) No NaN values found in data
      ii) Removed outliers where:
         (1) age>105 years
         (2) location frequency < 2
         (3) day of week = 6
   c) To visualize and model dataset, features have to be numerical. Extracting features out of datetime objects converts them to integer values that can easily be analyzed and fed

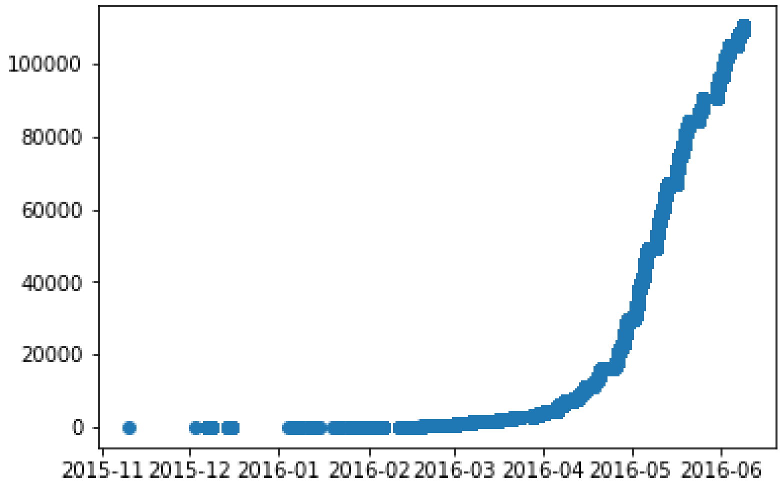
into a model. Lat/long coordinates create numerical representations of neighbourhoods and weather data may help contextualize no-shows. In regard to outliers, a few clients were listed as 115 years old; after research, there are no records of anyone from that age group living in Brazil; these values were therefore removed. If only one client visited the clinic from a specific neighbourhood, their data was also removed to avoid skewing the rest of the data. Saturday appointments were also removed: all recorded Saturday appointments occurred on the same day and are likely to indicate a special occasion or treatment and are not indicative of the general population.

6) Discovery and exploration

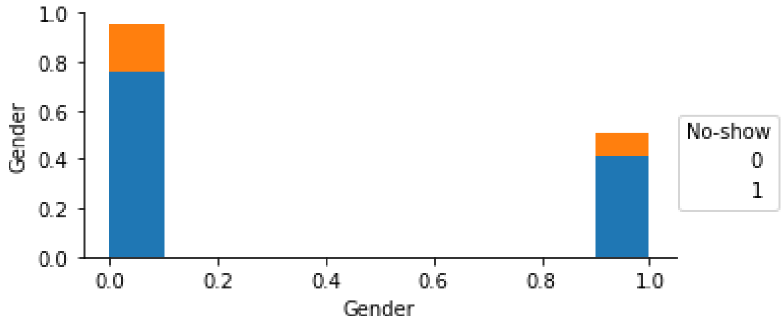
a) Final dataset includes the following columns:

i) Gender (binary)
ii) Age (ordinal numerical)
iii) Scholarship (binary)
iv) Hypertension (binary)
v) Diabetes (binary)
vi) Alcoholism (binary)
vii) Handicap (ordinal numerical)
viii) SMS received (ordinal numerical)
ix) Scheduled hour (ordinal numerical)
x) Scheduled and appointment day of week (ordinal numerical)
xi) Scheduled and appointment day of year (ordinal numerical)
xii) Schedule and appointment month (ordinal numerical)
xiii) Days elapsed between schedule and appointment date (ordinal numerical)
xiv) Patient frequency: number of times a patient has visited clinic (ordinal numerical)
xv) Area frequency: number of times someone from a specific neighbourhood visited clinic
xvi) Neighbourhood latitude and longitude (numerical)
xvii) Appointment day temperature highs and lows
xviii) Dummy variables for appointment day weather condition: fog, overcast, rain, sunny (binary)
xix) Whether patient showed up to appointment (binary)

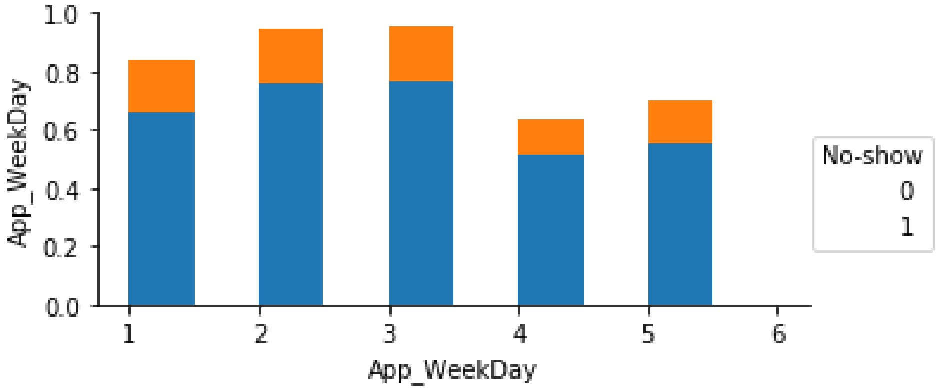
b) Schedule Date



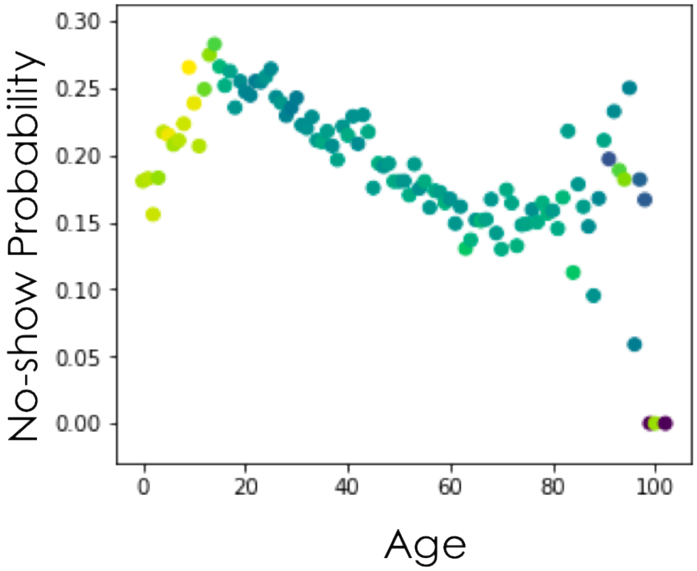
c) Gender distribution



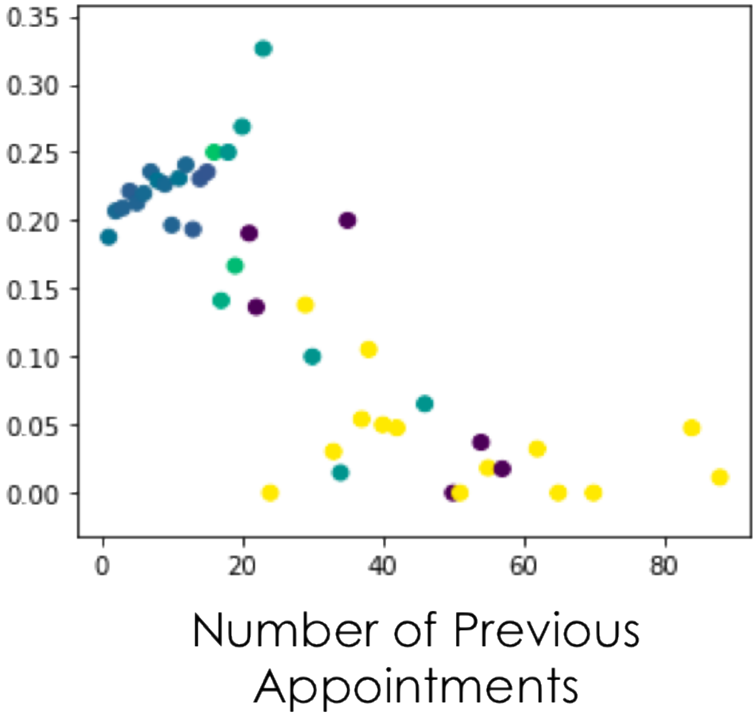
d) Appointment day of week distribution



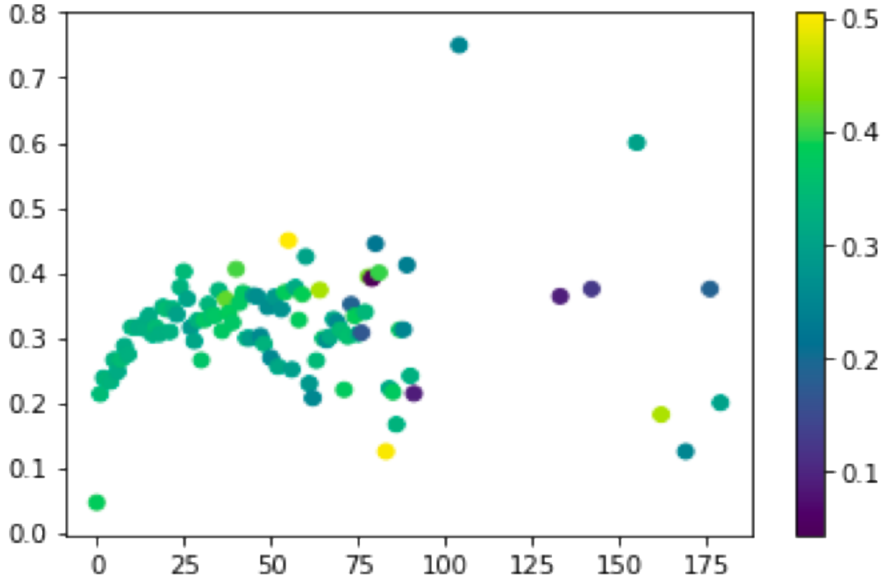
e) Age vs. no-show probability:



f) Number of prior appointments vs no show:



g) Days elapsed vs no show:



h) All selected data had at least a weak relation to no-show probability and kept in the final dataset.

7) Actionable insights: create and evaluate ml model

a) Model creation:

i) 20% of data set aside of independent testing

ii) XGBoost performed the best when selecting between ANN, XGBoost, LightBGM and Random Forest

iii) Evaluation metrics: logloss, which is good for classification problems, and area under precision-recall curve (AUPRC) which is sensitive to both false positives and false negatives and tries to balance the two; because our dataset is imbalanced, this is a good metric.

iv) Due to class imbalance between no-show vs show, oversampled no-show data to train model

v) Apply a 3 Stratified Kfold; stratified to ensure that the classes are balanced in training, and kfold to prevent overfitting

vi) Learning rate set to 0.1 to avoid overfitting

b) Model does not perform well despite many iterations and alterations. I conclude that these performance issues are likely due to poor data: we more patient information (i.e. occupation, number of family members in the same household, what kind of appointment it was, etc.) for model to perform better as there is simply not enough of a distinction between people that did and did not skip their appointments.

8) Application/data product

a) Will not set model up for deployment until there is better quality data for it to perform better.

9) Security/information

a) The dataset is publicly accessible through Kaggle and therefore is accessible to anyone who complies with Kaggle's policies.