Cristina ESPOSITO (260744222)
Mohamad KHALILI (260746712)

HEALTHCARE ANALYTICS – GROUP ASSIGNMENT

McGill University – Desautels Faculty of Management

DESAUTELS
Faculty of Management
Faculté de gestion

Data Cleaning

Before jumping into building our predictive model, the most important first step we must take is data cleaning. One of the first things to do is to check for NA values. If any NA values are present, they must be dealt with. If possible, the values are imputed using the complete() function from the mice package. Once the NAs have been dealt with, some more specific cleaning to the dataset can be applied. For example, some fields in the Chief Complain System were labeled as “ENVIRONMENTAL” and “Environmental”. Relabeling was applied as “Environmental”.

Feature Engineering

The given dataset provides a timestamp field. To use that timestamp as a predictor for the final prediction, a bit of feature engineering had to be done. First, we extracted the hours and regrouped them into bins of 3 hours. For example, we’d have midnight to 3AM, 3AM to 6AM, and so on. A new field called TOD (Time of Day) has been created to contain this newly created feature. Next, we used the given date to extract the month of the timestamp. Given that the test dataset might contain months that do not exist in the training dataset, we’ve grouped the months into seasons. For example, if the month is June, the new field “Season” contains the value “Summer”. A field called “Day of the week” has also been created to contain the day of the week at which the patient has been admitted.

A Random Forest model was developed to assess feature importance. From the results of the Random Forest, we found that *Arrival.Model*, *waitingcensus*, *Triage.Acute.Score*, *dayoftheweek* and *TOD* were the most important features, with *Arrival.Model* having a %IncMSE close to 600%, and *TOD* having a %IncMSE hovering around 20%.

Model Development

Various models have been tested with the aforementioned features. Linear models, multivariate regressions, as well as tree-based models (random forests & gradient boosted trees) have been tested. The gradient boosted trees provided the best RMSE out of all the other techniques. The hyperparameters such as the interaction depth and the number of trees have been tuned in order to find the combination that provides the best performance. With a test set of 58 values, using a Gradient Boosting Method with 1100 trees and an interaction depth parameter set to 8, we obtain an **RMSE of 5.64 minutes.** When computing the mean of the time to see an MD, we obtain a mean time of 39.2 minutes. **This means that our model predicts the wait time for a patient to see an MD with an accuracy of 86%.**