Exploratory Data Analysis and Feature Engineering:

Initial exploration and veracity evaluation:

Presented with both an objective and a dataset, the natural first step was to explore the data and understand more about the variables available. One of the first things done was note they type of each variable. Our group found that the dataset included numeric variables that were not categorical (excluding INDEX and including AGE, WEEKDAY_ARR, HOUR_ARR, MONTH_ARR, WEEKDAY_DEP, HOUR_DEP, MONTH_DEP, DIAG_DETAILS, and CHARGES), and had a larger amount of categorical variables.

Next, our group examined the veracity of the dataset – we noted missing values in a variety of columns: ADMIT_RESULT, RISK, and SEVERITY (all categorical variables) each had frequently missing values, while columns like GENDER had far less but still a few (GENDER had 2 missing values). Interestingly, the values for ADMIT_RESULT, RISK, and SEVERITY do not seem to be missing at random - if an observation was missing a value for one of these variables, it usually also was missing values for the other two variables in this triad. Likewise, an observation had a value for one of these variables, it usually also had a value for the other two variables. This led the group to think that these values may not be missing at random, and led us to conduct further analyses on these three variables.

RETURN has some rows with missing values. To address this, either imputation could be used, or the associated observations could be omitted entirely. To avoid mislabeling the RETURN value, and the task of having to impute the very thing our group is concerned with predicting with a model as whole, omission was decided as the better option.

Addressing missing values:

Given this pattern that applied to both the absence and presence of variables in this triad, our group initially wondered if this could have been related to a specific hospital that was more particular or rigorous in their record-keeping. We then checked if there was any pattern of correlation between which hospital a patient was seen at, and the proportion that had values in this triad of variable. However, we did not note any significant pattern.

However, we still had to address that there are consistently missing values in ADMIT_RESULT, RISK, and SEVERITY for many observations. Given how common observations missing these values were, we did not want to just omit any observation that had missing values. Additionally, we also thought there may still be predictive value in these variables, so we did not want to omit the variables from analyses on the entire dataset. To reconcile these two issues, we considered possibly imputing values with the MICE package in R, but decided against this - because these variables are all three categorical, we decided to just create a new category for missing values - “Missing”. This solves the two issues mentioned above, and lets us incorporate these variables into our further analysis.

Other comments on dataset:

There are more than 5500 blank rows in the .csv file. We selected the first 38221 rows, which contain useful data. The most common class for RETURN is “No”, with this representing approximately 75% of the data for RETURN. Upon noting this, we immediately noted that baseline accuracy for any sort of prediction will be correspondingly high, and creating a model that has a higher predictive accuracy may be challenging. However, there are other metrics for predictive value, including sensitivity and specificity. These metrics are likely important here becauses false negatives (i.e., predicting a patient will

not return in the next 30 days when they actually do) are impactful in a negative way (since too many false negatives could result in a hospital underestimating their anticipated patient volume in the coming month, and lead to issues with proper staffing and preparation). However, since no predictive models had been built at this stage of exploratory analysis, more technical and detailed approaches to this issue were temporarily shelved until initial models were explored and evaluated, while we still kept this issue in mind.

Our training and testing data has also already been partitioned into two separate .csv files. While we will keep them partitioned for training and testing, we do need to transform them identically, and then partition them back into the same splits before we begin to train any models off the data.

Variable exploration:

We also were interested in exploring relationships between variables to potentially analyze any possible interactions between them. Furthermore, we considered the possibility that some variables may not be useful in our later modeling, and set out to determine if there were any variables we should omit altogether. Additionally, we considered transforming and creating new variables from others in order to gain further information.

We also used correlation coefficients to identify which variables were highly correlated. The data was unpaired (arose from separate individuals) but we also were interested in identifying if any variables had any potential correlation with others. MONTH_ARR and MONTH_DEP were found to be extremely correlated, and this suggested to us that we should probably omit one of the pair from our dataset.

Similarly, the other two arrival and departure pairings, HOUR_ARR and HOUR_DEP, and WEEKDAY_ARR and WEEKDAY_DEP, were found to be highly correlated with each other (to the point it looks like duplication), likely for similar reasons as the two variables for month were found to be correlated. Related to arrival and departure is the variable SAME_DAY. However, we noted that there was a patient who had the same values for all arrival variables as for their departure variables, yet their value for SAME_DAY was 0. This further made us suspicious of duplication with the arrival and departure variables. For this reason, we will omit all departure variables.

Another variable we explored was AGE. Running a tree split on AGE > 81.5 years, the majority class for RETURN was ‘No’ by a large margin. Considering that median life expectancy is also around this value at 84.5 years, this led us to consider another interesting point. If a elderly patient who visits has a very serious condition and dies either then or in the near future, there is no possibility for them to return, even though they may appear similar (in terms of having similar values for predictor variables) to other elderly patients who do survive to return; this may “confuse” our predictive models, as elderly patients may be very sick and otherwise would have been flagged as very likely to return if they hadn’t passed away first. From a modeling standpoint, patients who pass away shortly after their visit are essentially observations “leaving” the dataset while other observations are still tracked and have their return status monitored, whether they do or do not return in the next 30 days. This is an interesting conundrum, although more related to data collection and inclusion methods more than data analysis, since we have no way of telling if a patient expired soon after their particular visit that was noted in this dataset. While there is the DC_STATUS of “Expired”, this only tells us which patients died literally during the recorded visit, not those who died after being discharged alive. Another interesting fact found related to age: the maximum value for AGE was 127 - however, the oldest ever documented human was purportedly only 122 years old, so this is an oddity of this dataset. In similar medical databases, an age of 127 is entered to imply that the age is unknown, so this might have been true of this dataset, as well.

There were three variables related to physician consults, and they seemed correlated from an intuitive standpoint, in how the variables were defined. CONSULT_ORDER, CONSULT_CHARGE, and CONSULT_IN_ED were all related to consultations: whether a consultation was ordered, whether a consultation was charged for, and whether a consultation was ordered while still in the emergency department.

CONSULT_ORDER and CONSULT_CHARGE seemed to be very similar in that they related to whether a consultation was ordered and whether it was paid for. When calculating correlation coefficients, CONSULT_ORDER and CONSULT_CHARGE were also found to be highly correlated. Interestingly, CONSULT_CHARGE had an interesting relationship with RETURN - recall that the normal return rate is approximately 25%. However, when patients had a value of 1 for CONSULT_CHARGE (implying that a physician consultation was charged for), only 12% of these patients returned in the next 30 days, compared to the baseline return rate of approximately 25%. This suggests that if a physician consultation is charged for, the probability of that patient returning in the next 30 days is less than double that of the baseline return rate. This is interesting, and from an intuitive standpoint, it does seem possible that a hospital visit for a condition that being charged for a physician consult may result in care that adequately treats the patient's condition to the point where they have no need to return in the next 30 days.

Likewise, CONSULT_ORDER and CONSULT_ED exhibited an interesting correlation. When CONSULT_ORDER takes the value of 0 for an observation, the corresponding value for CONSULT_ED is 0. CONSULT_ED does not add much new or unique information to the dataset, and is a candidate for omission as well.

DC_STATUS is a categorical variable that is text-based, with many different discharge statuses (over 25). This is likely to create issues in data analysis in R, since it will treat them as categories. However, some discharge statuses are far more common than others. There were approximately 38000 rows, and 29906 of these had a discharge status of being discharged to home or self care. On the converse, there were some discharge statuses, like discharge to substance abuse rehabilitation facility, or expired, that only had 100s or even tens of corresponding observations.

RACE also had similar issues with text-based categories - there are only 6 observations for "Declined to Answer" for example, 8 for Hispanic, 12 for Two or More Races, and 6 for Native Hawaiian or Other Pacific Islander.

While we have now omitted INDEX, and are still on the process of omission, we also recall that we have identified 4 pairs of highly correlated variables: CONSULT_ORDER with CONSULT_CHARGE, ACUITY_ARR with ED_RESULT, MONTH_ARR with MONTH_DEP, WEEKDAY_ARR with WEEKDAY_DEP, and HOUR_ARR with HOUR_DEP. We also believe that CONSULT_ED may not add any valuable information to the dataset, and it is a candidate for omission as well.

For this issue of categorical variables with very sparsely populated categories, we decided to make "catch-all" aggregate categories - i.e., a race category of "other races" that includes all races that make up less than 3% of the overall testing set, and a discharge status of "other discharge" that includes all discharges that make up less than 3% of the overall testing set. Because many variables had so many categorical levels to them, we found that accommodating this many levels was very computationally expensive when it came to model-building. One possible response to this issue is to reduce the number of categories within each categorical variable, but the question of how to do so remained unanswered. Manual aggregation of similar categories for each variable, influenced by domain knowledge, is perhaps

one of the most ideal approaches to this question of how, but given our outsider status relative to this type of domain knowledge, we instead decided to encode with respects to the target variable. Our approach to this issue was to aggregate these rare categories into new "catch-all" categories with respect to the mode of the target variable.

The mode of return rate associated with each variable was determined, and then used to reduce the amount of categories within that variable: if any category's respective return rate was less than this mode for the categorical variable to which that category belongs, then it was binned into the aggregate category "Lower than Mode". This process is like a form of target encoding but with the mode rather than the mean, and was repeated for two more aggregate categories: "Mode", and "Higher than Mode". This reduces the amount of categories in categorical variables by aggregating them into 3 categories in a meaningful way, by correlation with modal return rate.

There were some thoughts of overfitting that may result from this procedure, but because this does not directly replace the categories with an aggregation statistic of the target variable, and rather just aggregates into discrete categories, we thought we may gain some advantage from this process: "local" testing accuracies (accuracies on our labelled testing data set rather than the withheld testing set) improved in some models using this encoding.

Modeling and Model Evaluation

After variable exploration and transformation, our group's next step was to partition all the data available to us. Our group decided to partition this data into a testing set (20%) and a "rest of data" set (80%). Then, the rest of the data was partitioned again: 80% of the rest of the data was partitioned into a training set and the other 20% was partitioned into a validation set.

Before running any preliminary models, we needed to make sure our data was clean enough. We had already replaced all missing values, and now need to create our catch-all aggregate categories for RACE and DC_STATUS. We also needed to make the decision about which variables to omit. We ended up deciding to omit INDEX, CONSULT_CHARGE, CONSULT_ORDER, WEEKDAY_DEP, HOUR_DEP, and MONTH_DEP, for the reasons mentioned in the Data Exploration sections above.

At this point, we were again conscious that the baseline accuracy of this dataset is already approximately 75%. Knowing that our initial models would likely not significantly exceed this accuracy, we still began to explore classification models in order to gain insight on which variables may be more useful than others, and to verify the usability of our processed dataset. Accuracy was however the metric we were interested in using for initial model evaluation, as this was the provided target metric that competition was centered around. We did consider the importance of sensitivity and specificity in this scenario as well since there is a 75% - 25% split of the two classes, a fairly uneven distribution, but decided to target accuracy since this was still the requested measure to improve - accuracy was how models were evaluated in terms of scoring. Balancing both these concerns, the prevailing mindset was that we would aim to improve our true positive rate (TPR), since this reflects the portion of the more rare class ("Yes" for RETURN) that we correctly classified as such, but also ensure that we were beating the baseline in terms of accuracy.

One of the first models our group developed was a classification tree. We were not aiming to use this for any real prediction, but just as another form of early exploration. The first split it made was on GENDER, suggesting this may be a useful and important variable. Indeed, there is a 0.1849 return rate for

females (3431 ‘yes’, 15047 ‘no’), compared to a 0.3026 return rate for males (5952 ‘yes’, 13648 ‘no’ for return).

Another early exploration for modeling was the use of Principal Component Analysis (PCA), to attempt to linearly combine some variables into higher variance principal components. The rationale behind implementing PCA was accommodating for potential correlation we may have missed either due to insufficient exploration or lack of domain knowledge, while still retaining valuable information about our data through linear combinations (the principal components). We first (temporarily) excluded AGE and CHARGE because they are continuous numerical variables that would not scale well, while other categorical variables can be represented as either 0 or 1 and were kept for PCA. When running PCA on this dataset, we found that 50 principal components would have to be used to achieve approximately 95% cumulative variance, and the first three principal components captured a cumulative proportion of variance around 45% and showed little increase in proportion for each component thereafter. We then used the first 50 principal components and merged them back with AGE and CHARGE respectively, in order to retain the information contained within these features, even though they were not part of the original PCA process.

Using this dataset with 52 variables (50 principal components and AGE and CHARGE), we developed many models via a variety of approaches (LASSO, Ridge, k-NN, 50 bagged logistic regressions, 50 bagged trees, and regular Logistic regression) on our training set, and all except logistic regression beat baseline accuracy. The model that beat the baseline and had the highest true positive rate was produced by bagging 50 trees, and had a 17.25% TPR and 77.5% testing accuracy.

We wanted to then explore the changes in TPR that would occur with adjustments to cutoff values, so we began by iteratively decreasing the threshold from 0.5 to 0.35 (at 0.35, all models’ accuracies were below baseline) and again evaluating all above models’ TPR and accuracy. The highest TPR of a model with an adjusted threshold that still had above baseline accuracy was the model produced by 50 bagged logistic regressions; it had a TPR of 31.05%, near double that of the previous best TPR, and the accuracy was 76.03%.

We considered that PCA may have obscured some valuable features, so we also were curious to attempt a similar procedure without PCA (on the same variables with the same omissions as mentioned earlier). We initially developed a standard logistic regression and classification tree from our training data and evaluated the models on our testing data. The logistic regression model failed to beat the baseline accuracy, and the classification tree’s accuracy was quite close to the baseline accuracy. Since we aimed for our models to at least beat the baseline, we then considered using ensemble methods to improve our prediction accuracy We also tried LASSO and Ridge models, since these two methods are capable of reducing variance (in exchange for an increase in bias). k-NN was also considered, since it is data-driven and feasible to use (after converting categorical variables into dummy variables). Using the above specified models, the model with the highest TPR was produced by a random forest of 300 trees that considered 5 variables at each split; its TPR was 37.20% and its accuracy was 75.26%. We noted that using the original data instead of principal components led to a different model being selected, with a significantly higher TPR (compared to 31.05%) and a slightly lower testing accuracy (compared to 76.03%).

Furthermore, we decided to investigate the highest accuracy that we could obtain with these procedures, with no consideration to the final model’s TPR. The highest testing accuracy obtained was a bagging procedure with 100 trees, which produced 77.97% testing accuracy, followed closely by random

forests. From this we noted that ensemble methods and regularization measures like LASSO and Ridge regressions also tended to produce higher testing accuracies relative to other methods with this dataset.

Of the models submitted for prediction of the withheld dataset, our best model was produced by 80 bagged trees with an mtry of 5. The accuracy from this model was 77.73% on the withheld data. True positive rate on this data set was not determinable due to the nature of the testing and submission process, but it is worth noting that the model that beat the baseline and had the highest true positive rate on our testing set was produced by a random forest of 300 trees, and had a 37.20% TPR and 75.26% testing accuracy (local testing baseline was 75.02%).

Conclusions:

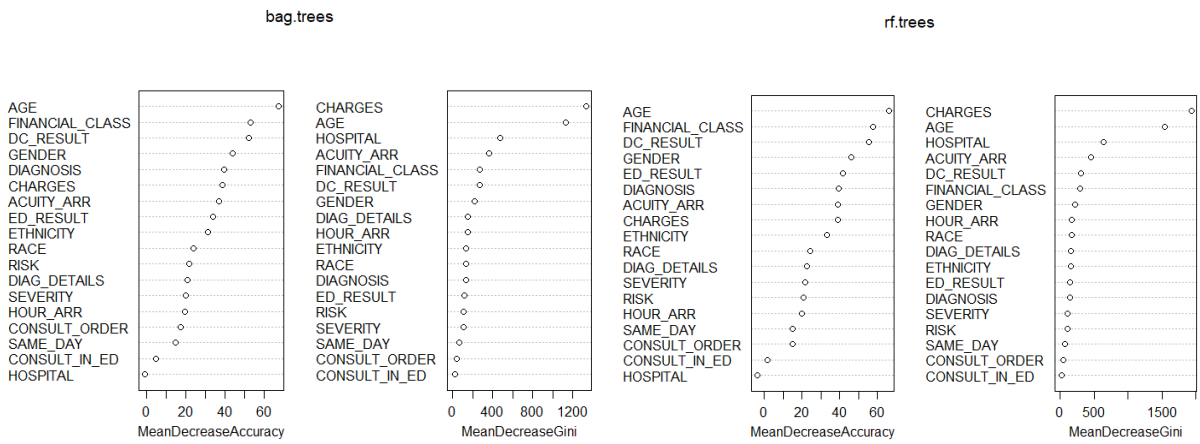
A general pattern of tradeoff between TPR and accuracy were clearly evident in our process, as expected and is common with many classification problems.

The model that we would implement in practice is the one with the highest TPR and above baseline accuracy on the withheld set. In our case, that would be our model produced from randomforest of 300 trees, mtry=5. It had an accuracy of 75.26% on the local testing data, compared to the baseline of 75.02%, but had the highest TPR of 37.20% (on our testing dataset).

While disappointing that our best accuracy was only 2.29% above baseline, it seems this set was difficult to significantly beat the baseline of and created challenges for many - out of 20 groups' best models, the highest testing accuracy was 1.30% higher than ours. Of these 20 groups, all beat the baseline, but none by more than 3.6% - this again suggests similar universal challenges inherent in the task. The factors behind a patient's return (or lack of) in 30 days are very likely to be complex and numerable, and given the results of the 20 groups, likely difficult to predict with significantly higher accuracy than a baseline of common class.

Further insights came from variable importance plots, which identified age, insurance status, discharge result, and gender to be the most important variables (see Figures 1 and 2, attached). While interpretation from a variable importance plot cannot be directly claimed, one would suspect that each of these variables have intuitive rationales for their relation to return rates: the elderly may be more likely to return for chronic age-related conditions, those without insurance are likely to not return due to financial hardship, discharge status tends to reveal if further follow-up care is needed, and a female may be more likely to return than a male since, for example, obstetric medicine only applies to women and often involves frequent check-ups.

Again, it is stressed that direct interpretation cannot be claimed from these plots, but perhaps this dataset could be explored from an inferential perspective in a future analysis to gain more insight into the domain knowledge involved with this goal of classification. This in turn could then be applied to feature engineering precluding another, and this time more-informed, foray into this predictive task. Hopefully then, a more meaningful and appreciable increase in predictive accuracy and TPR could be produced.



Figures 1 and 2. Variable importance plots for bagged tree and random forest models, respectively.