project_report

November 4, 2019

1 Improving Hospital Inpatient Admissions Predictions Leveraging Temporal Relationships

1.0.1 Phil Johnson, Nick Tyler

1.1 Goals and Outcomes

Can we prove that temporal relationships in a medical patients history improves the ability to predict future hospital admissions? In this project we analyze a health insurance dataset on patient outcomes to try to predict future hospital admissions. We will be looking specifically into the temporal nature of some of the features we pull into the dataset, and specifically at if we can improve model outcomes when incorporating more temporal features.

1.2 Summary

Overall we found reasonable evidence that adding temporal features to a previously non-temporal data set can help improve modeling outcomes of readmission predictions. We looked at several time frames for features within the training data, as well as several timeframes for the response variable of "readmitted to hospital within x months". The timeframes we used were 12, 6, 3, and 1 months. We then compared this to the non-temporal version of the model dataset, containing only one of the flags for the feature variables, being the 12 month timeframe feature variables. We found that especially for the prediction of outcomes over a shorter timespan (1 or 3 months), there is strong evidence to suggest that adding in multiple features to create a dataset with temporal characteristics resulted in improved outcomes for the model over a range of measurements. We found model accuracy and AUC scores to be improved in models across all timeframes. We also found these improved differences to be statistically significant at the .05 level when predicting readmission within 1 or 3 months. Moving forward with the project, for the second half of the semester, we have a number of outlets to continue exploring. These include: maximizing model accuracy/other scores, exploring the temporal nature of other features not yet analyzed, predicting specifically for preventable hospital readmissions, and modeling specifically for the highest risk patients as this is the most pressing business case.

1.3 Data Processing and Transformation

1.3.1 Data Source

We were able to secure historical data from health insurance companies that provides information tracked by the insurance company to track payments for services rendered to patients, medication

prescriptions, and other relevant information pertaining to patients and the insurance company. This data is all cleansed and staged with the purpose of building out predictive modeling to support population health initiatives and improving patient care.

Example data elements include the following:

Patient Demographic Data

This is data collected on members of the insurance company when you enroll with a plan - age - race - gender - language

Enrollment Data

Historical information of what members were enrolled on the plan when tracks the enrollment segments of members on specific plan with their specific benefit packages - duration on plan type - insurance type - duration with primary care provider - eR visits - inpatient admissions - readmissions - PCP visits

Patient Medications

Historical prescription fills for each members - NDC (medication type) - dosage - supply - recent prescription fill count - drug type - patient diseases

Disease/Condition list for each member

Examples: - Asthma Indicator - Diabetes Indicator

1.3.2 Data Processing

In order to perform analysis and modeling on our dataset, our goal was to flatten all of the above data sources into a patient record. The reason behind this is so that we have a representation of the current and past state of each patient contained within one record. In order to do this, we created a record with features representing the different aspects of each patient we want to model on, and ran aggregation across the different data sources to fill in the features for each patient.

Demographic Data

This data was already in the form of one line per patient so no transformation was needed

Event Data

This data source was structured as a one line per visit dataset, meaning that a single patient may have multiple lines containing data relevant to our analysisn. This data was collapsed down into indicator features which were one hot encoded. This included features such as ER visits within the last 12, 6, 3, and 1 months. We performed similar tranformation on PCP visits and inpatient admissions.

Patient Medication Data

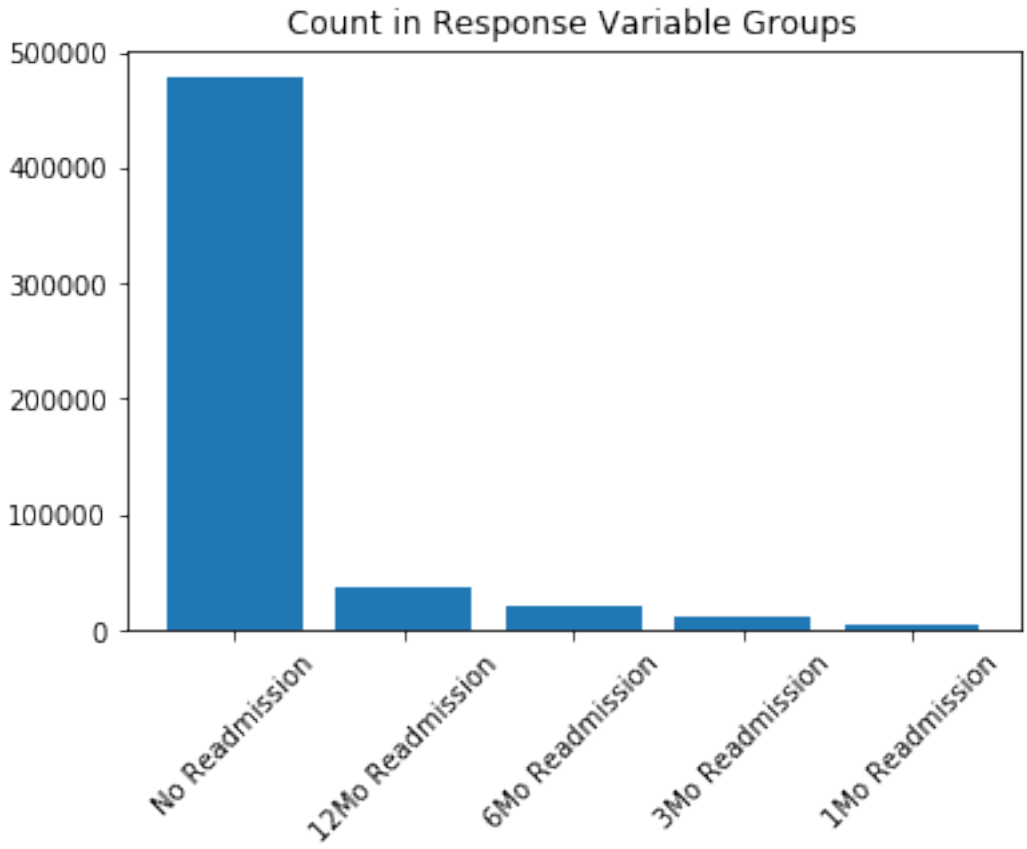
This data source was structured on a one line per medication per fill basis. This means that each patient would have a record for every time they fill a medication, for each medication that they are currently perscribed. This data was collapsed down into indicators features and into count features. The count features represented how many times a patient had filled any perscription within the past 6 and 12 months. The indicator features represent one hot features indicating the class of medication that a patient might have filled, such as "high risk", "opioid", and "potentially harmful".

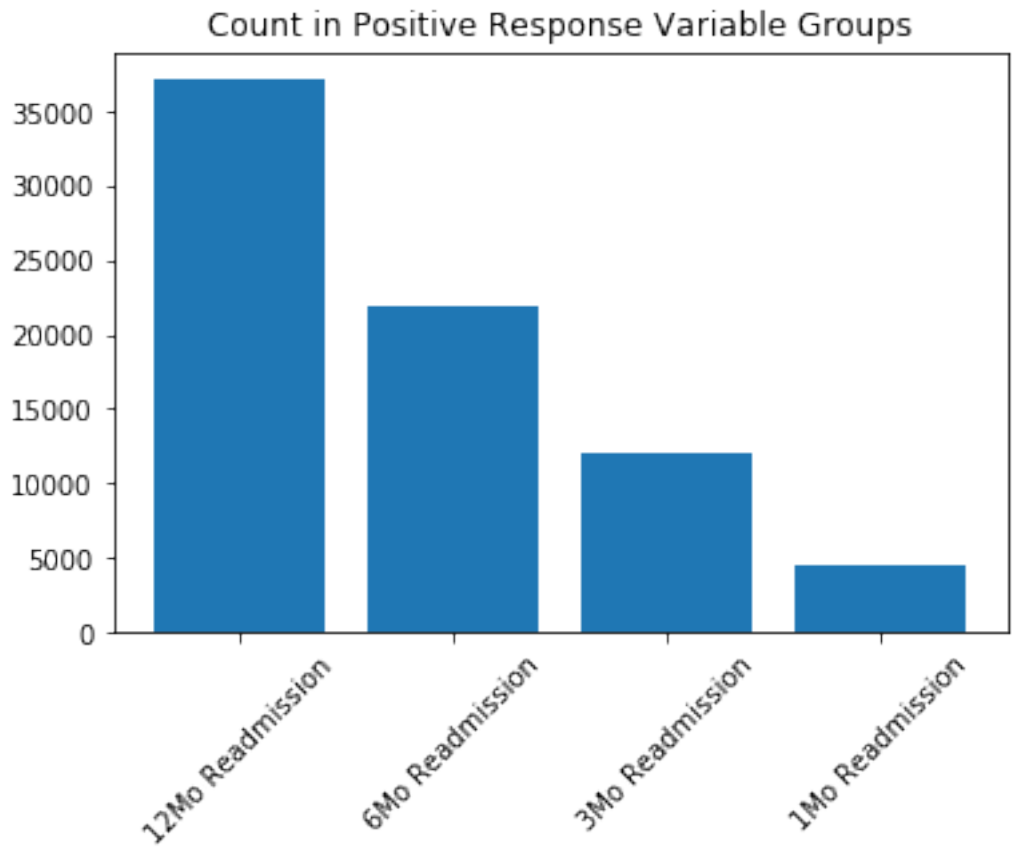
Disease Data

This data source was structured on a per disease per patient basis. This data was collapsed down into a high number of one hot features indicating whether or not a patient had been diagnosed with a given disease. For the majority of patients this resulted in a sparse data set with only one or two features that were relevent to them.

1.4 Data Exploration

Our final reduced observation count of observations we used to build our datasets was 515k records. Of these records, 37k were readmitted to a hospital within 12 months of the "as-of" date. 21k were readmitted within 6 months, 12k were readmitted within 3 months, and 4.5k were readmitted within one month. Summary graphs of the data are included below.





1.4.1 Datasets for Analysis

For the analysis portion of the project, we reduced the data as described above into lists of observations of patient records. However, our intention is to analyze the effects of adding temporal features to the observations rather than using simple indicator variables. Because of this, we have created two datasets of these observations described above.

Non-Temporal

The non-temporal dataset contains all of the indicator variables for diseases, counts of medication, and medication indicators. For the event data described above, we have included only indicator variables for 12 month periods. This means that we are only tracking whether or not a certain event or visit occured for a patient at one time range rather than over a spectrum of time.

Temporal

The temporal dataset contains all of the same features as the non-temporal dataset, but also includes indicator variables for medical events over 6, 3, and 1 month periods. This means that medical events are encoded in 4 variables per event type rather than just 1 indicator variable.

1.5 Feature Importance Analysis

As an initial look at the various time based features and their effects on any modeling that is done, we ran an initial feature importance test on the temporal features within the temporal dataset. We

sorted the variables in descending order of importance with the most important features being first and the least important being last. The importance scores that we used to rank the features were pulled from a binary tree classification model. Scores were given for all features within the dataset, but we ranked only the temporal features, any only looked at features which ranked inside the top 6 temporal features. A table summarizing the results is below.

Response Variable	12mo	6mo	3mo	1mo
12 Month Readmission	3	1	1	1
6 Month Readmission	2	2	1	1
3 Month Readmission	1	2	2	1
1 Month Readmission	2	1	2	1

The above table shows that there is a case for the importance of the additional temporal features over only the 12 month features. With the exception of the 12 month readmission response variable, at least one of the 12mo temporal features is not included within the top 6 for that response variable. This means that some of the other time periods for these models are more important predictors, which suggests that excluding them may produce worse results.

One additional observation from the above table is that the shorter term features seem to become more important as the time range of the response variable becomes shorter. The 3 and 1 month features are more important for predicting 3 and 1 month readmission, based on the model used to generate this table. This suggests that the temporal nature of these features may be more beneficial as the prediction period shortens.

1.6 Comparison of Temporal and Non-Temporal Models

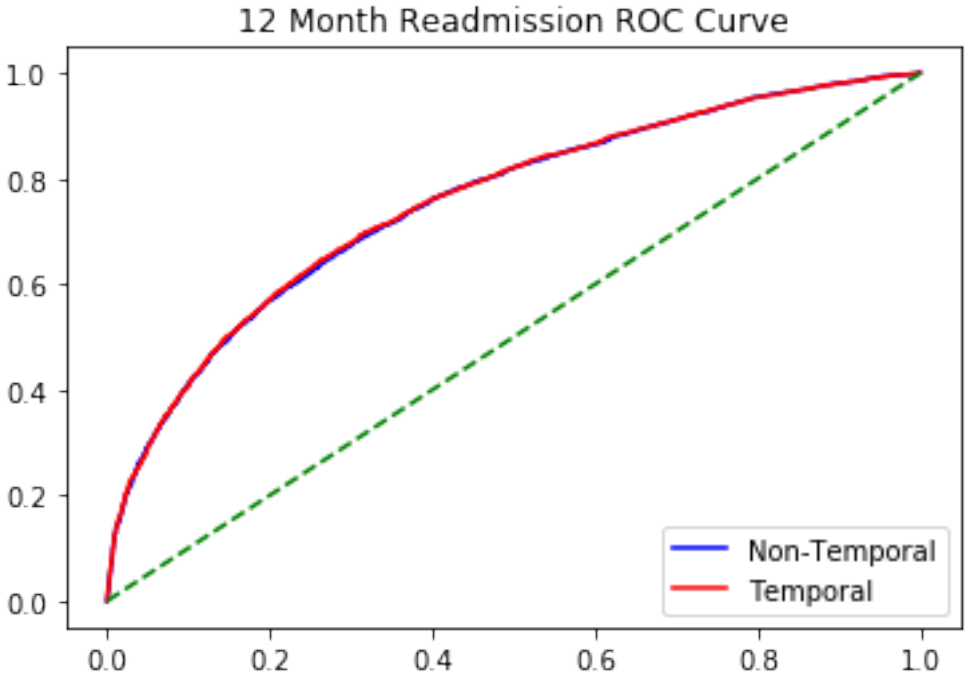
1.6.1 Overview

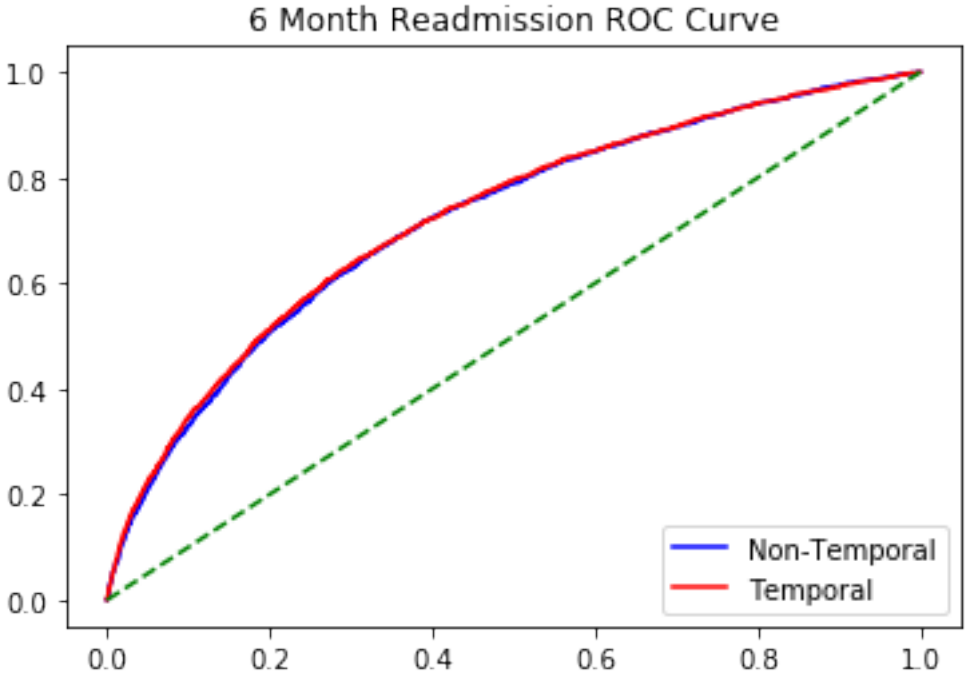
We ran two iterations of a model, one for the non-temporal dataset, and one for the temporal dataset. In these models, we made predictions against the range of response variables, readmission within 1, 3, 6, and 12 months. In both of the models, we used a logistic classification model.

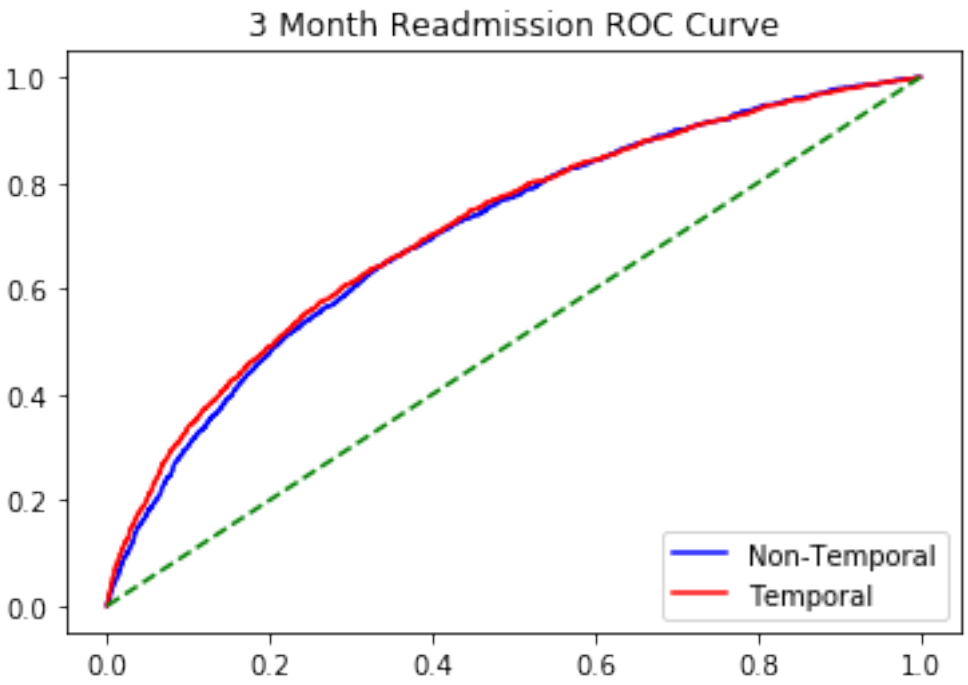
1.6.2 Sampling

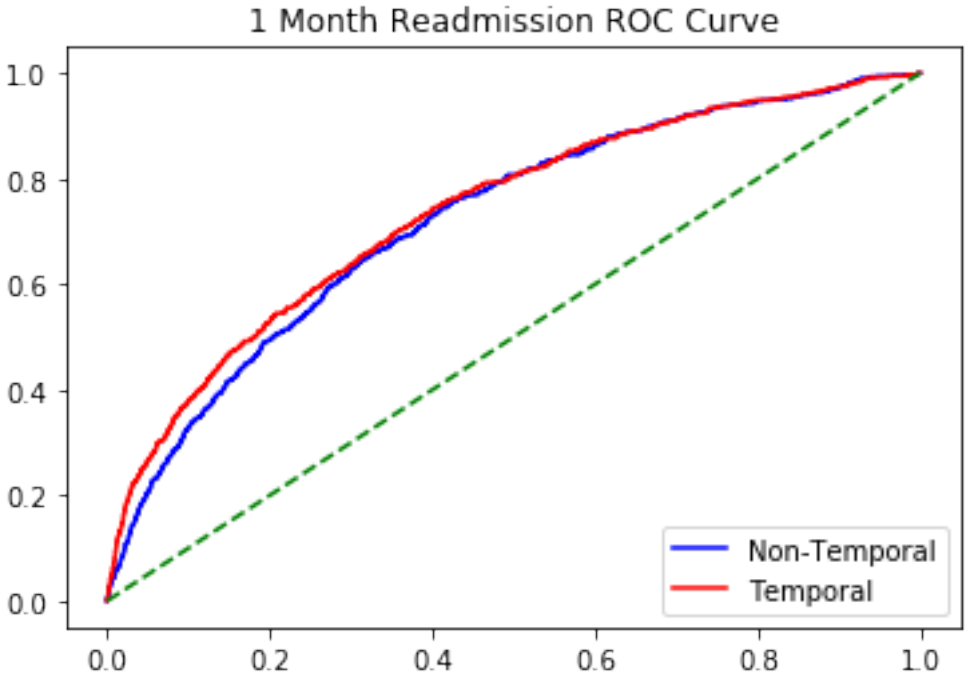
For each dataset, we originally used the complete dataset of 500k rows to build our models. However, due to the unbalanced nature of our response variables, our models performed quite poorly. We received high accuracy rates but had too high of a false positive rate relative to our true negative rate, showing the model was simply just predicting the most frequent value for the response variables. After this, we downsampled the number of "negative" samples to get a more balanced subset of the data. We sampled 30k records with a positive 12 month readmission rate, and 30k with a negative, resulting in 60k samples for modeling. From these 60k samples, we used 80% for training data, and 20% for testing data.

1.6.3 ROC Analysis



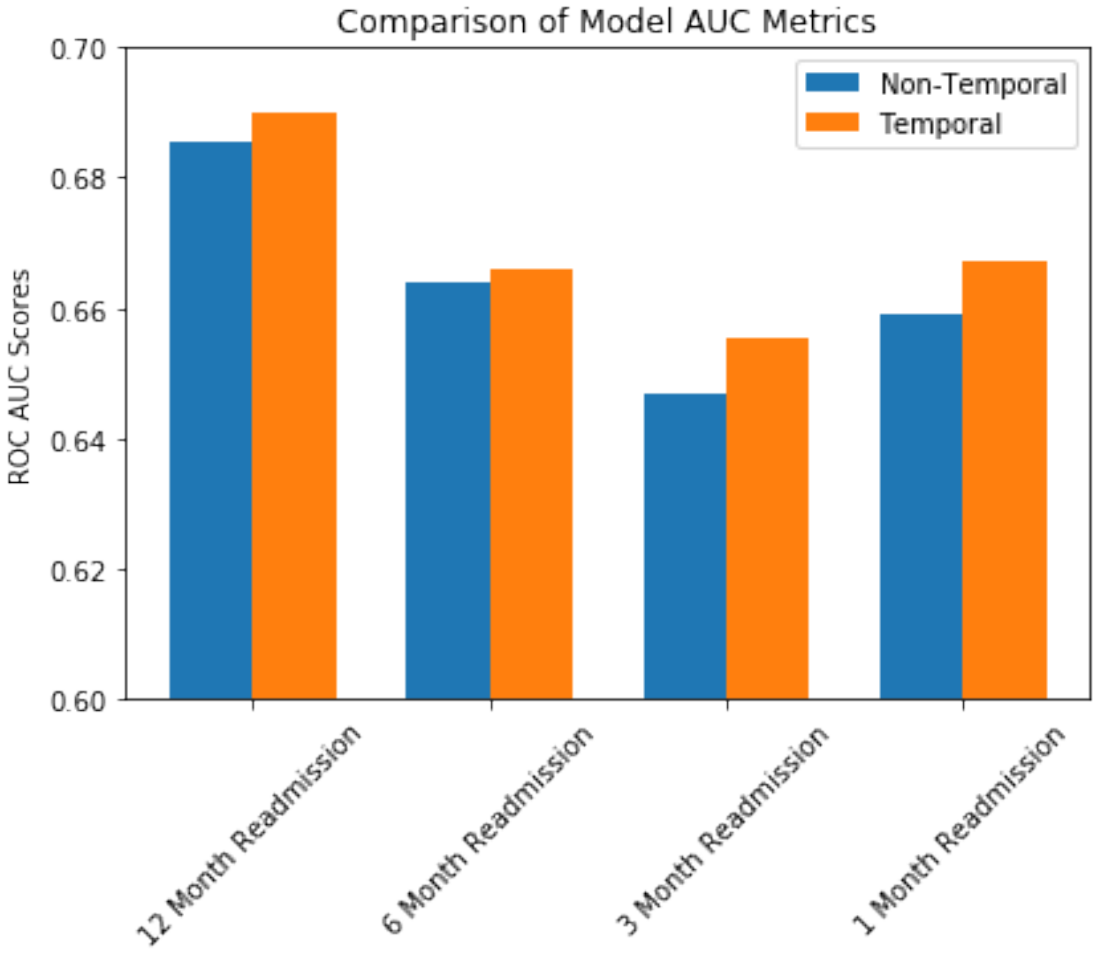






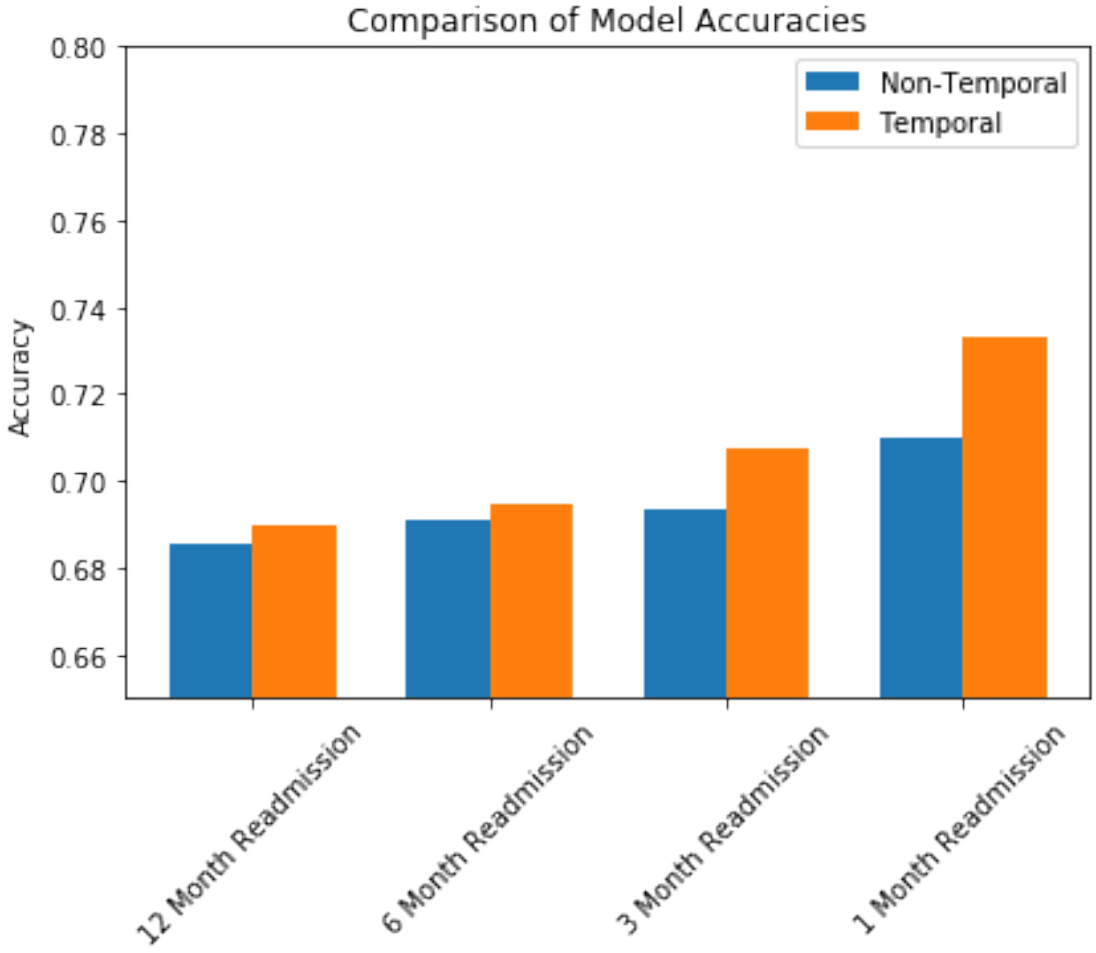
Above are ROC plots, one for each of the four response variable timeframes. While the performance of the temporal and non-temporal models are similar for the 12 month readmission

prediction, the models begin to diverge as the estimation period becomes shorter. In the 1 month readmission model, there is a clear seperation between the ROC curve of the two models. Below are the AUC scores for each model on each response variable.



Here we can see a clear advantage for the temporal model overl the non-temporal model. The temporal model performs better that the non-temporal model over every response variable. As is the case in the ROC curves, the temporal model has the largest advantage over the non-temporal one during the 3 and 1 month readmission predictions. also of note is the overall shape a score progression on the chart. Both models do relatively well during the 12 month prediction bucket, and worst in the 3 month prediction bucket. This is particularly interesting as this does not follow the trend of the accuracy scores below which increase steadily from 12 to 1 months. It is possible that due to the proportions of the 3 month dataset that the model is overfitting for one of the response variable groups.

1.6.4 Accuracy



Overall the temporal model once again outperforms the non-temporal model over all of the response variables. We see once again its advantage grows larger as we move from the 12 month to 1 month readmission predictions. It is also noteworthy that the accuracy of each model improves moving from 12 to 1 month predictions. One possible explaination for this is that it is generally easy to predict outcomes over smaller periods of time (to a point). However, it is also possible that this is due to an imbalance in the data. The 1 month readmission field class had the lowest number of positive observations compared to the other 3 timeframes. It is possible that the model is overpredicting the negative class during the 1 month predictions due to this imbalance. However, the AUC for 1 month predictions is similar to the 6 month predictions and better than the 3 month predictions, which would suggest that the 1 month predictions are not overpredicting more than those 2 timeframes.

1.6.5 Hypothesis Testing

Response Variable	T Statistic	P Value

12 Month Readmission	0.74	0.4604
6 Month Readmission	0.59	0.5568
3 Month Readmission	2.47	0.0137
1 Month Readmission	4.03	0.0001

Lastly, we ran T-statistic tests for the differences in accuracies for the non-temporal and temporal models. The results are listed in the table above. We found that the improvement in accuracy seen in the above graphs is not statistically significant at any reasonable level for the 6 and 12 month readmissions. However, it does indicate that the 3 month is significant at the .05 level and that the 1 month difference in accuracy is significant at better than the .001 level, showing strond statistical evidence that the differences in the 1 month are legitimate. These results are in line with the above graphs and our original expectations. These show stronger evidence for improvement as the response timeframe moves from 12 month to 1 month readmission, consistent with our figures from above.

1.7 Next Steps

Further improvement of the model Overall, our models had accuracies around 71%. While this is ok, especially for a dificult to predict variable, it is not ideal and not likely to be a sutible level for a business case. Moving forward, we will look to improve this accuracy rate without diminishing related statistics such as the f-score or AUC measurement.

Explore other possibly temporal variables

During our original analysis, we found the medical event based features to have the strongest predictive power and as a result, selected those for our temporal analysis. However, there are other features which were not included in this analysis which have a temporal nature to them, such as when a patient fills a drug. We may look to include some of these features in future models to improve outcomes.

Predict specifically for preventable readmissions

We have access to whether or not a readmission to the hospital was preventable or not. For the purposes of our original analysis, general readmission was sufficient for a POC and exploration. However, the business case for the model is for preventable readmissions, and we will look to predict on these in future models.

Predict for the highest risk patients

The business case for the model is strongest for higher risk patients. It is not very interesting to predict that many patients will not be readmitted to the hospital because they are generally healthy. The business case revolves around predicting outcomes for patients who are more likely to be readmitted than others. In future iterations of the model, we will look to model specifically for these high risk patients, and tune the model for the best outcome within this observational slice.

1.7.1 Statement of Work

Both members of the group contributed equally to the project thus far. Phil Johnson has acted as the point of contact between the group and the interested company, as well as contributed to the data modeling and transformation pieces. Nick Tyler has also contributed during the data transformation steps in addition to the modeling. Both members contributed equally to the report.