Hospital Readmission in Diabetic Patients

Braden Anderson, Hien Lam

September 19, 2022

I Introduction

Hospital readmission prevention remains a critical issue that affects patients, healthcare providers/insurers, and policymakers with its primary goal to simultaneously reduce cost while improving quality of care. The obvious downside of increased readmission rates for patients is the implication that they received poor outpatient care as well as paying out of pocket more often (if not the patient, then insurance agencies). Hospital providers are financially penalized if they maintain high readmission rates, enacted under the Affordable Care Act as a strong statement that readmission is a national priority. In particular, diabetic patients have higher risk of hospital readmission. We conducted multivariable logistic regression grid search with regularization, feature interactions, and polynomial features on relevant patient attributes. The target response encompassed three classes: no readmittance (`No`), readmittance within 30 days or less (`<30`), and readmittance after 30 days (`>30`). The feature importance was discerned and discussed below. We noted that with this type of data, personal identifiable information is an area of concern. Patient demographics contain sensitive data and could be used nefariously. For example, a patient that isn't diabetic but retained certain diabetic-like attributes could incur higher health insurance premiums. Even seemingly innocuous features such as patient number or encounter id could be mapped to the patient's name if this data were sold or stolen. On the other hand, a hospital or primary care physician with this information could classify at risk patients that could curb the diabetes diagnosis before they develop it.

II Methods

Data Preprocessing

The `diabetes` dataset consisted of 101,763 observations of diabetic-encounter, hospital admissions which span a diverse set of patient profiles and documentation practices. The `id_map` dataset contained the numerical indicator and its respective description for three separate data: admission type, admission source, and discharge type. For example, a patient being admitted into the emergency department would have an indicator value of 1. The `diabetes` dataset contained the indicator columns, and its description was mapped from `id_map`. There was a total of 37 categorical and 13 numeric data types.

The following preprocessing operations were performed to transform the original clinical records into a more consistent and useful set of model inputs:

1. Removed observations designated as "unknown": `gender` (there were three observations)
2. Removed column with 97% missing values: `weight`
3. Removed constant columns: `examide, citoglipton`
4. Binned categorical columns: `age, diag_1`

For `age`, years 0-30 are "young", 31-60 are middle, 61+ are old.
For `diag_1`, the top 10 most frequent classes were kept, and the other 706 classes reclassified as "other"

5. Imputed missing observations: `medical_specialty, payer_code, race, diag_1`

   All features here are categorical and thus imputed with the most frequent class.

6. `diag_2, diag_3` columns were dropped. Both had scant missing values however we did not feel comfortable imputing with the most frequent class given the 700+ categories. These two features represent the non-primary diagnosis that the patient came in with and we used our intuition that it likely would yield fruitful predictive power, especially more so than `diag_1`, the primary diagnosis.

Exploratory Data Analysis

The demographic analysis for this particular use case is a bit sensitive and we must acknowledge that the data will not be truly representative of the population for a myriad of reasons. Our intuition tells us that younger adults likely will avoid hospital visits (for readmission or otherwise), due to finances, time constraint, or personal beliefs, as is also the case for certain races. Older Caucasian was the most prevalent demographic (equivalent proportion for males and females). When accounting for age (**Figure 1**), we see that Asian and Hispanic exhibit the highest non-readmittance rate. We theorized that there are other factors at play, notably cultural norms and monetary behavior. Most of the numerical features (time in the hospital, number of procedures, number of medications taken) were right skewed and were addressed appropriately in the next sections. Of the features that relate to laboratory blood results, we noticed that the majority of patients did not have diabetic-related medication in their system.

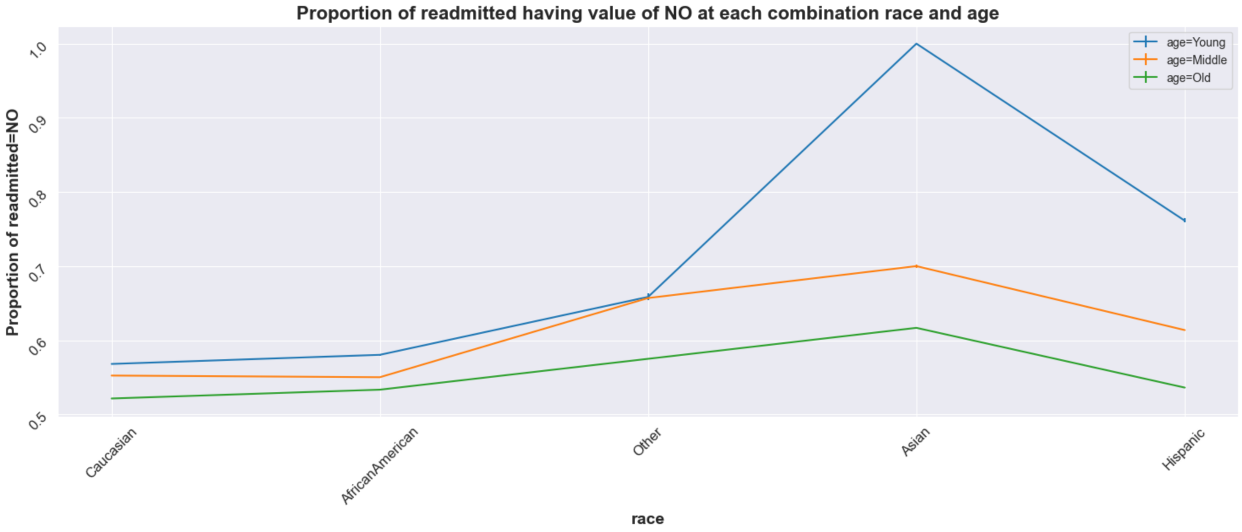

Figure 1: Proportion of Race "No Readmittance" After Accounting for Gender

Feature Scaling

In a multiclass classification for logistic regression, the loss function used is multinomial log loss. We felt confident SKL's Power Transformer Scaler was the most conducive transformer for the data based on its robustness to heteroscedasticity and the desired normality assumption for logistic regression. The Yeo-Johnson method was implemented within this scaler class. We also experimented with SKL's Standard Scaler.

Establishing a Baseline

Prior to model building the 101,763 observations in the preprocessed dataset were split into cross validation and final test sets with 96674 (95%) and 5089 (5%) observations respectively. The baseline models and all hyperparameter tuning activities are performed using the cross-validation set, with the test set being reserved to obtain an unbiased estimate of the final selected models generalization error. To ensure the error estimate obtained from the final test set remains unbiased, no model selection decisions will be made on the basis of final test set performance.

Model Specification **(Strategy, preprocessing, new features)**	**Train Accuracy** **(5-fold CV average)**	**Validation Accuracy** **(5-fold CV average)**
Multinomial, Yeo-Johnson	59.58%	59.36%
Multinomial, Standard Scaler	59.51%	59.31%
One vs Rest, Yeo-Johnson	59.53%	59.34%
One vs One, Yeo-Johnson	59.60%	59.38%
Multinomial, Yeo-Johnson, Polynomial	59.74%	59.51%

Table 1: Baseline Model Performance

From table 1 we see that the logistic regression model with default hyperparameters has remarkably consistent average cross-validation performance across a variety of feature transformations and multi-class strategies. The impact of hyperparameter values on cross-validation performance will be explored in the next section.

Modeling

To improve upon the baseline model performance shown in **Table 1**, a hyperparameter search was performed to tune the logistic regression penalty type, regularization strength, and class weighting strategy. The search evaluated all three penalty types (L1, L2, and elastic net) at 15 different regularization strengths ranging from 0.05 to 10. For the elastic net penalty, three values of the elastic net mixing parameter (l1 ratio) were explored at 0.25, 0.5 and 0.75. In addition to the default (uniform) class weighting, the search included a balanced strategy which applies a weight to each observation that is inversely proportional to the frequency of the target class label.

The hyperparameter search outlined above was performed for three of the baseline models shown in **Table 1**, the results of which are displayed in **Table 2**. From the model performance summary in **Table 2**, we see that the best performing hyperparameters for each model are very close to the default values with the largest deviation being a small increase in regularization strength to C=0.7 (default 1.0). We also note that the cross-validation performance of the best models identified by the search are nearly identical to those of the corresponding baseline model, which is an expected result given the similarity between the selected and default hyperparameter values.

Model Specification **(Strategy, preprocessing, new features)**	**Selected Hyperparameters**	**Train Accuracy** **(5-fold CV average)**	**Validation Accuracy** **(5-fold CV average)**

Multinomial, Yeo-Johnson	C = 0.7
	Penalty= L2
	Class weights = 1
		59.58%	59.36%

Multinomial, Standard Scaler	C = 0.7
	Penalty= L2
	Class weights = 1
		59.50%	59.32%

Multinomial, Yeo-Johnson, Polynomial	C = 0.9
	Penalty= L2
	Class weights = 1
		59.74%	59.51%

Table 2: Model Performance Summary

Feature
num_procedures
num_medications
number_inpatient
time_in_hospital
number_diagnoses
number_outpatient
number_emergency

Table 3: Interaction Term Features

The best performing model in **Table 2** utilized the multinomial loss function, Yeo-Johnson power transformations of all numeric inputs and additional features created as interactions between the attributes listed in **Table 3**. This model achieved an average 5-fold cross validation accuracy of 59.51%, a modest improvement over the null model which predicts the "No" target class for every observation (53.91%). Due to the relatively small improvements made by the best logistic regression over the null baseline, an error analysis was performed to quantify the model's performance on each of the three target classes, the results of which are described below.

Of the 96674 validation set observations, 52118 (53.91%) have a true readmittance status of "No", which was predicted by the model 68562 (70.92%) times. Of the 52118 cases where the patient was not readmitted to the hospital, 43241 (82.96%) were correctly classified, 8714 (16.71%) were incorrectly classified as ">30", and 163 (0.31%) were incorrectly classified as "<30".

The ">30" readmittance status is the true label for 33767 (34.92%) observations, and was predicted 27559 (28.50%) times. Of the 33767 patients which were readmitted to the hospital more than 30 days after their visit, 14306 (42.36%) were correctly classified, with the incorrect classifications being assigned to “<30” and “No” 184 (0.5%) and 19277 (57.08%) times respectively.

The “<30” readmittance status was the true label for 10789 (11.16%) observations, but was only predicted by the model 553 (0.527%) times. The model correctly identified only 206 instances (1.9%) of hospital readmittance within 30 days, with the “No” and “>30” target categories being incorrectly predicted for 6044 (56.02%) and 4539 (42.07%) observations respectively.

The error analysis outlined above indicates that model is incapable of reliably identifying cases of hospital readmittance within 30 days, predicting this class for roughly 20 times fewer observations than the categories true frequency. Further, of the 553 observations which were predicted to be cases of short term readmittance, less than half were correct classifications (206).

Further review of the associated grid search results shows that the best model in terms of F1-score (F1-score=0.5533) is also ranked 6th in terms of accuracy, with very similar cross validation accuracy (59.51%) to the model discussed in the error analysis. An interesting observation is that this similar performance was achieved by a model with vastly different hyperparameter values, utilizing the elastic net regularization type, a regularization strength of C=10, and an l1-ratio of 0.75. Analysis of the model with the best observed F1-score reveals it also severely under predicts short term hospital readmission, predicting this label for only 577 observations, with 214 correct classifications. Although a detailed analysis of each case will not be discussed here, we also note that this same phenomenon of being unable to predict short term readmission was observed for logistic regression models with a variety of regularization strategies as well as a completely unregularized model.

III Results

Final Test Set Performance

The logistic regression model with the best observed cross-validation F1-score (penalty=elastic net, C=10, l1-ratio=0.75) was evaluated on the 5089 observations in the final test set. The average training, average cross-validation, and final test set performance of this model are summarized below:

Model Specification	**Average Train Accuracy**	**Average Validation Accuracy**	**Final Test Accuracy**	**Average Train F1-score**	**Average Validation F1-score**	**Final Test F1-score**
Multinomial, Yeo-Johnson, Polynomial	59.75%	59.51%	59.70%	0.5544	0.5512	0.5531

Table 4: Final Model Performance

The performance summary in **Table 4** shows no indications of overfitting to either the training data or cross-validation process during hyperparameter tuning. These metrics provide evidence that the short-comings of this model stem from error due to bias, which could be caused by either an inadequate set of features, or the model having insufficient flexibility to find useful relationships within the existing feature space. From the discussion on error analysis in the modeling section, we believe the majority of this bias is related to the detection of short-term hospital readmission, and therefore the best path forward to improving model performance would be to find new features or relationships which are predictive of the less than 30-day hospital readmission class.

Feature Importance

The top 5 coefficients by absolute value for the "No Readmittance", "Short Term Readmittance", and "Long Term Readmittance" sub-models are displayed in **Figure 2**, **Figure 3** and **Figure 4** respectively. By comparing these charts, we observe that the binary indicator column for a discharge disposition of "Expired" is the number one most important feature when generating each of the three probability estimates. Considering that a discharge disposition of expired implies that the patient is now deceased, it is not a surprising result that all 1642 cases of this disposition are associated with patients who were never readmitted to the hospital. We believe there is minimal impact to the above analysis associated with including these observations because all models were biased in the same manner and a sample of 1642 observations from a dataset of this size could at most swing the prediction accuracy by 1.61%. That said, excluding these observations would be recommended for any future analysis.

The impact on model parameter estimates can be observed by comparing **Figures 2** through **4**, to **Figure 5**, **Figure 6**, and **Figure 7** which display the top 5 largest coefficients after refitting the model on data with these 1642 deceased patients removed. From **Figure 5**, we see the largest magnitude coefficient associated with estimating the probability of "No Readmittance" has a negative relationship with the number of inpatient visits that patient had in the preceding year, coefficient = -1.911. This result matches intuition that patients which have recently had a larger number of visits to the medical clinic may have more severe illness and therefore a lower probability of not be readmitted to the hospital in the future. **Figure 6** shows that most influential coefficient for predicting short term readmittance is associated with the binary indicator column for the admitting physician having a specialty in gynecology. In particular, the coefficient of -2.21 indicates that a physician of this specialty will be associated with a reduced probability estimate for short term readmission. Similarly, **Figure 7** indicates that an admitting physician with a specialty of pediatrics will be associated with a reduced probability of long-term readmission (coefficient=-2.93).

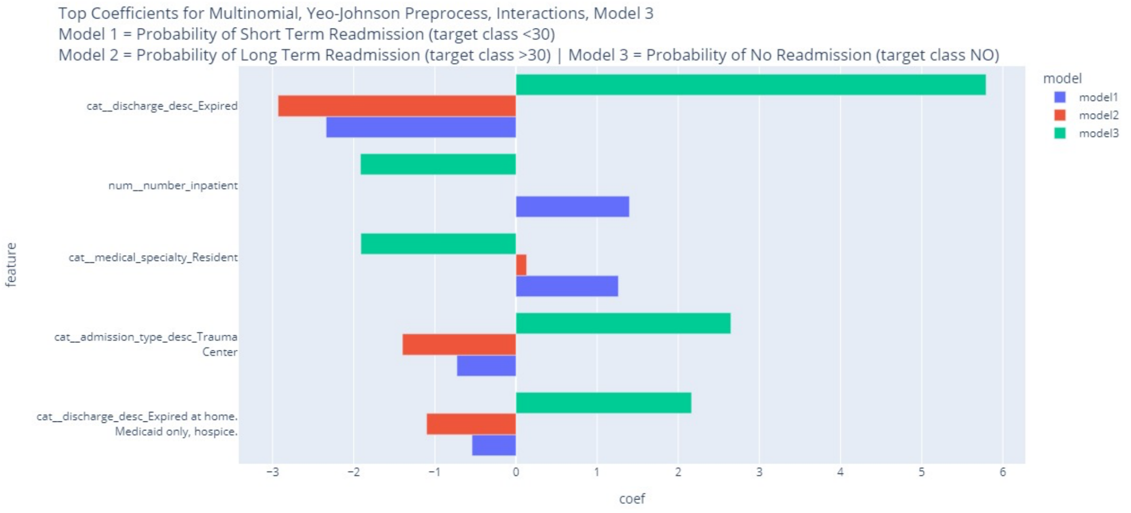

Figure 2: Top 5 coefficients (Absolute Value) for Predicting "No Readmittance"

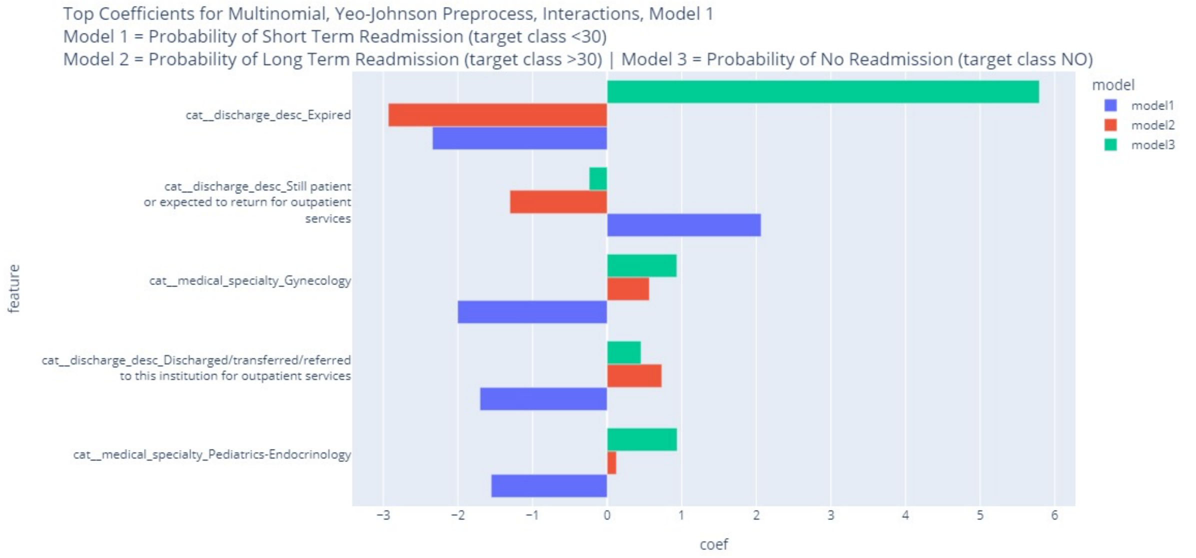

Figure 3: Top 5 coefficients (Absolute Value) for Predicting Short Term Readmittance "<30"

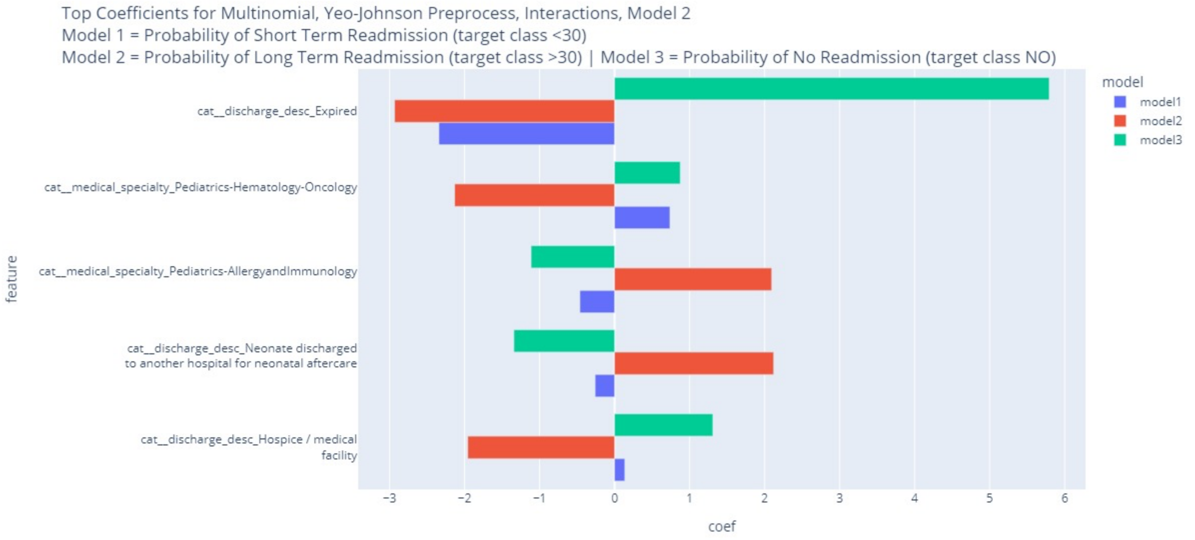

Figure 4: Top 5 coefficients (Absolute Value) for Predicting Long Term Readmittance ">30"

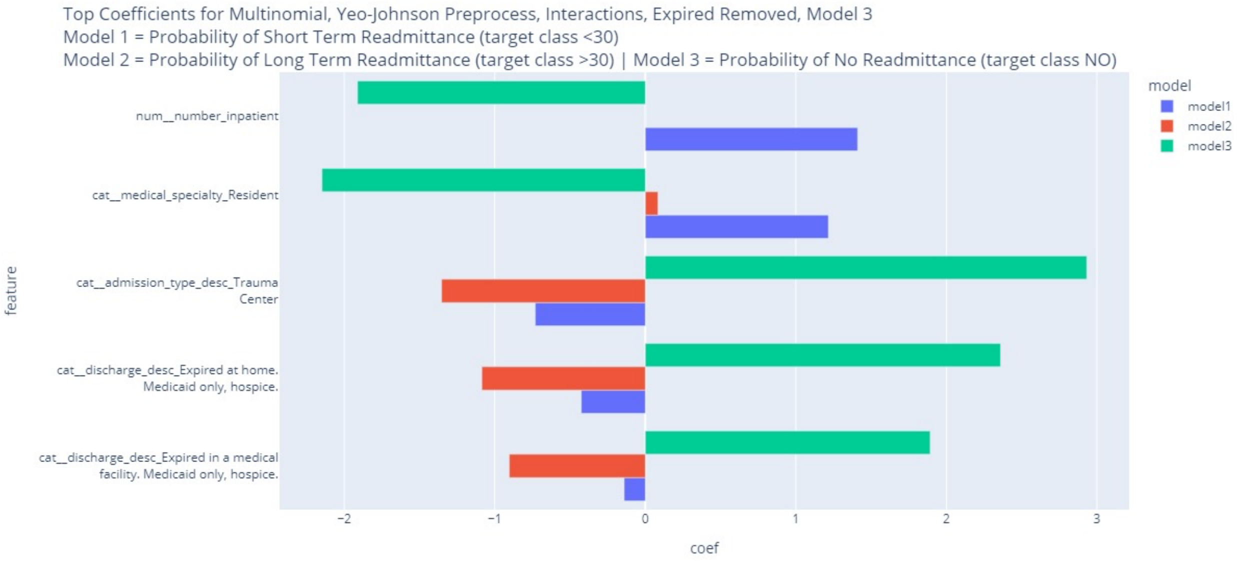

Figure 5: Top 5 coefficients (Absolute Value) Predicting "No Readmittance", No Expired

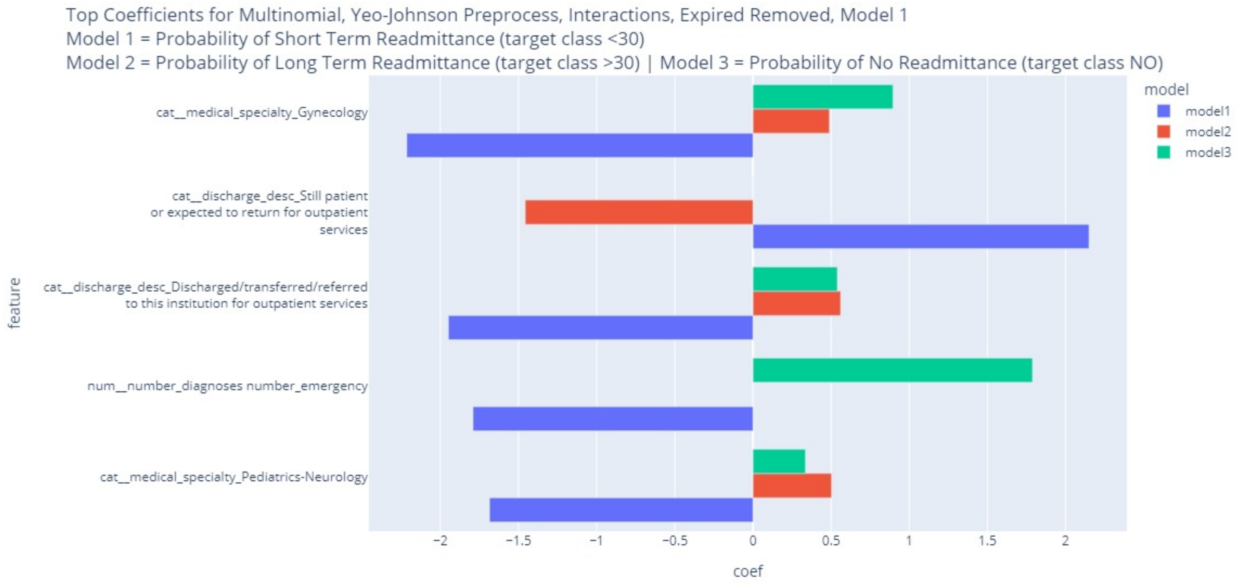

Figure 6: Top 5 coefficients (Absolute Value) Predicting Short Term Readmittance "<30", No Expired

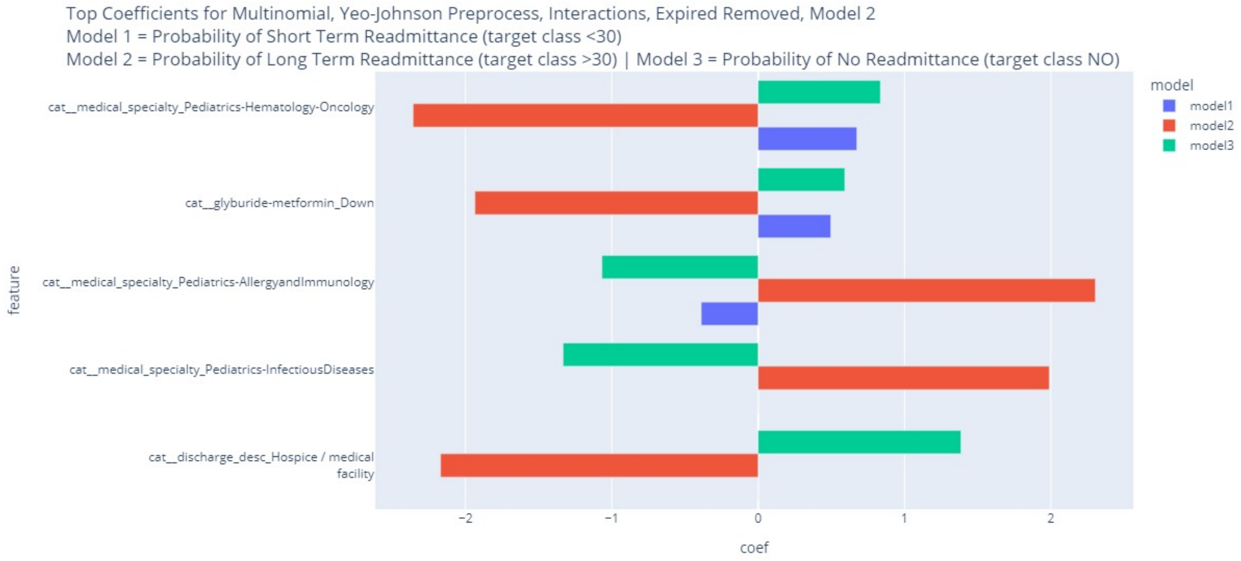

Figure 7: Top 5 coefficients (Absolute Value) for Predicting Long Term Readmittance ">30", No Expired

IV Conclusion

In this paper, we utilized multivariate logistic regression to predict a diabetic patient's hospital readmission status from attributes which described the individual and the medical care they received. The logistic regression model was selected because it provided a direct relationship between each attribute and its impact on readmission probabilities, which can inform strategies for reducing patient costs and improving quality of care. To this end, the model coefficients which are used to estimate the probability of each target class were inspected and discussed in section 3. Our analysis found that the logistic regression model can reasonably (about 83% accuracy) distinguish patients who would not be readmitted to the hospital from those who would but was much less reliable at discerning cases of short-term and long-term readmission. One factor which may be related to the reduced performance on the short-term and long-term classifications stems from the somewhat arbitrary usage of these terms, which in the context of these data is defined as less than or greater than 30 days. In future work, we plan to investigate the use of machine learning to directly map the input features to an estimate of time until readmission and generate any desired groupings as an additional post-processing step.

V Appendix

Code

Please refer to the attached python source code and python notebook.