Machine Learning Project

Final Report

Merrimack College

DSE6211GA Machine Learning

Taryn Popplewell

Data Science Program

Executive Summary - Hospital Recidivism in Diabetes Patients

Introduction

According to the Center for Disease Control's 2020 report on diabetes, approximately 34.2 million Americans have diabetes, and 88 million have pre-diabetes. That's roughly 1 in 10 and 1 in 3 Americans having some form of diabetes respectively. Diabetes is a disease that can be managed at home with the proper supplies and dietary habits, but left unchecked can lead to serious health problems and death. According to a paper published by the American Diabetes Association, diabetes is a major contributor to hospital recidivism, or readmission. Repeat hospital admissions are generally not desirable from the patient or hospital perspective, and the purpose of this study is to evaluate patient data in hopes of identifying factors to reduce diabetic hospital recidivism.

This document outlines the final results of a study attempting to identify factors contributing to the hospital recidivism of diabetes patients in the United States. The data used represents 10 years of clinical care at 130 US hospitals and integrated delivery networks. The original data set included more than 50 variables detailing patients' time in hospital and whether they had to be readmitted following their discharge.

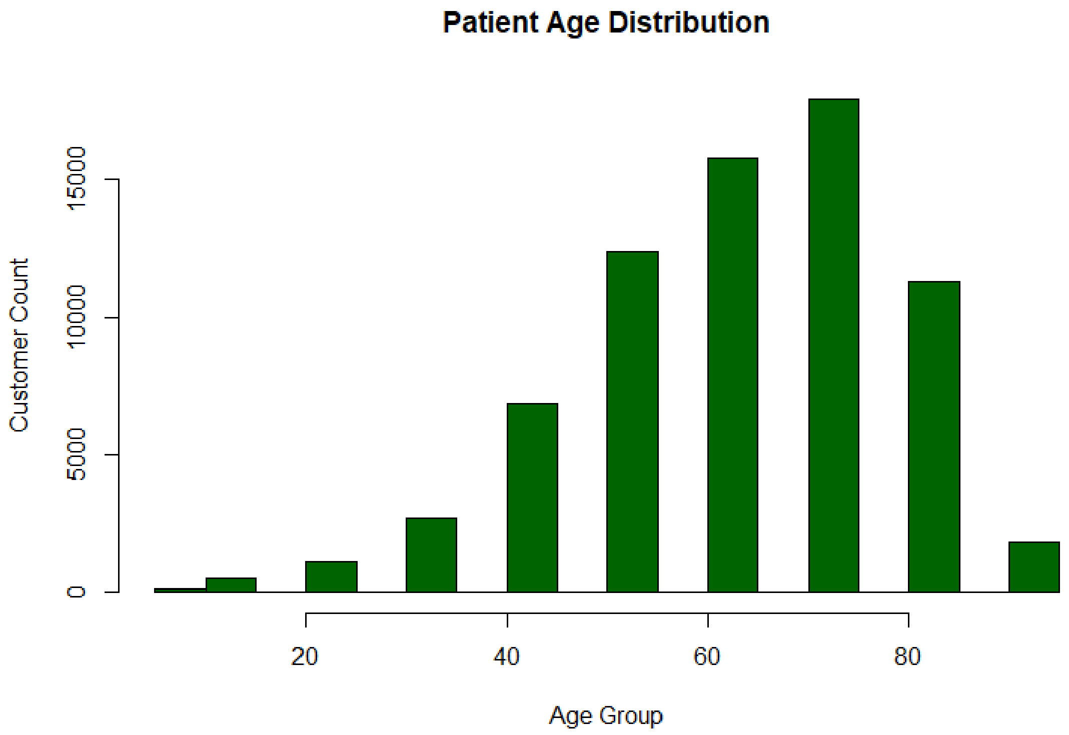


Figure 1. Patient Age Group Distribution in the data set.

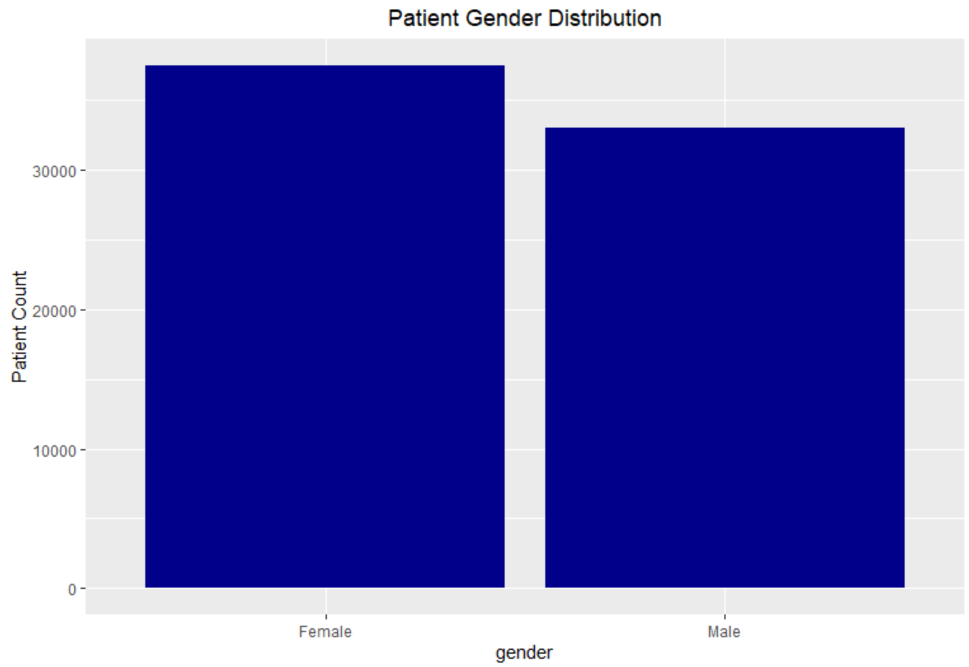


Figure 2. Patient Gender Distribution in the data set.

General patient ages fall roughly between 45 and 85. The data set is nearly evenly divided between males and females, and the predominant race represented is Caucasian. There was no additional information regarding where hospitals providing data were located, and this information could be useful in providing more context to scant demographic data included in the data set. From what is available, the patient base is older, white, and just over half had some kind of health insurance payer.

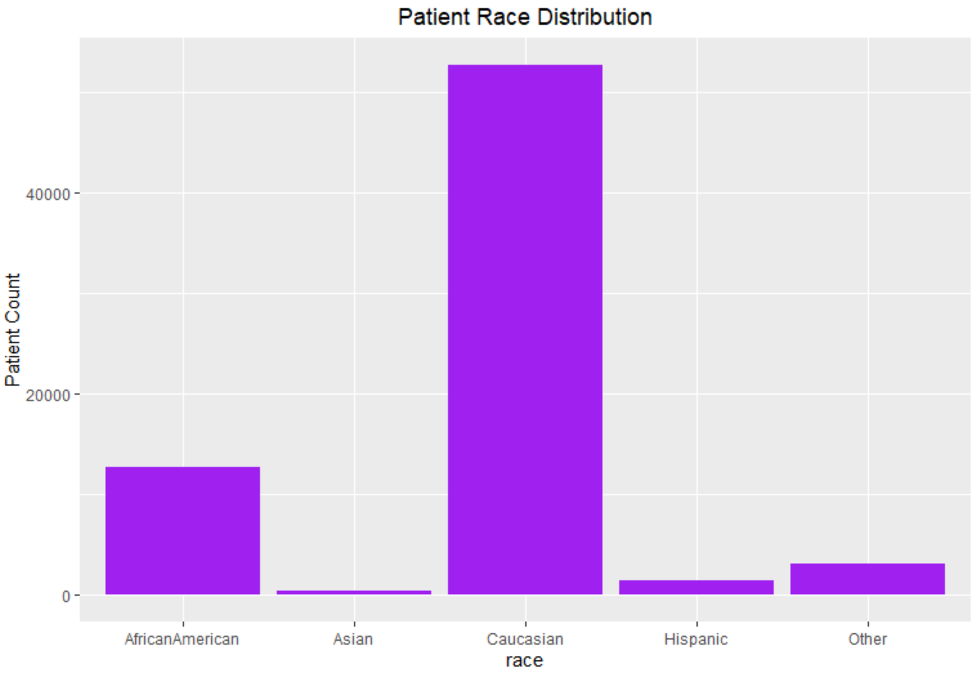


Figure 3. Patient Race Distribution in the data set.

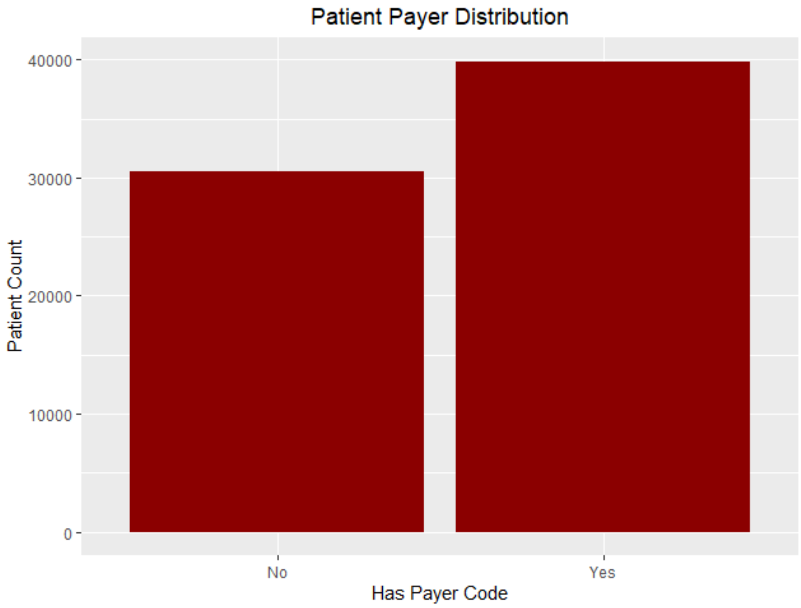


Figure 4. Patient Payer Distribution in the data set.

Key Findings

A number of different trials were run to search for the most predictive combination of features and algorithms. Ultimately, the data was run through a K-means cluster analysis to produce a cluster feature to be used in a random forest classification model.

Run without the K-means cluster assignments, the random forest model produced an Area Under the Curve (AUC) score of .66 which is good but not great. When the model was run a second time with the K-means cluster assignment included as a target encoded feature the AUC shot up to .939, which is significantly closer to the 'perfect score' of 1.

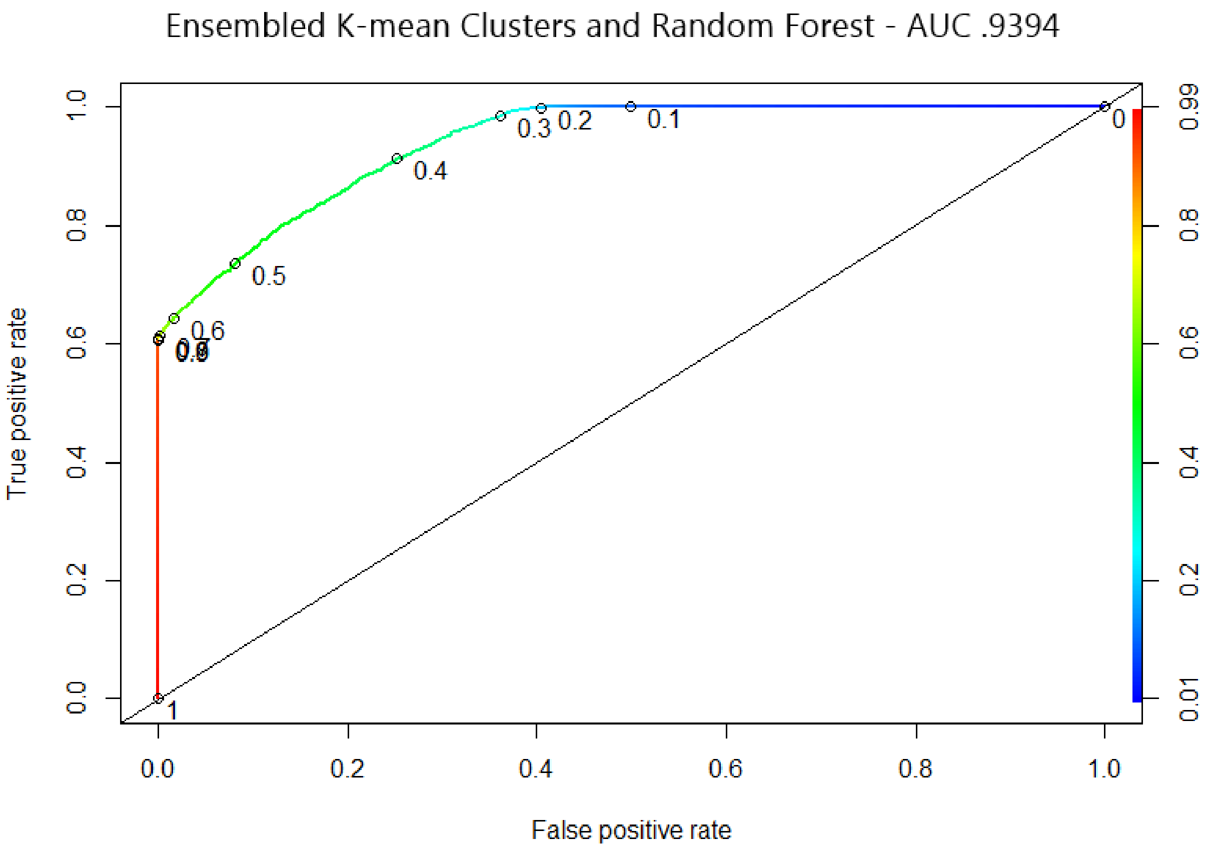


Figure 5. Ensembled K-means Clusters and Random Forest ROC Curve.

Such an improvement in the random forest model warrants a deeper dive into the results of the k-means cluster analysis results. A quick look at counts of readmittance within the four clusters produced shows clusters 2 and 4 comprised almost completely of opposing readmittances. Cluster 2 has a few readmittances but is by and large comprised of patients who were not readmitted following their discharge. Clusters 1 and 3 have a more balanced mix of readmittances.

This cluster analysis process used about half of the available variables, none of which were patient demographic variables. This could be an important point to keep in mind since it suggests that elements of the patient’s stay such as duration, number of diagnoses, or number of lab procedures could provide an early indication of whether a patient is likely to require readmission, and not necessarily their background or other outside factors.

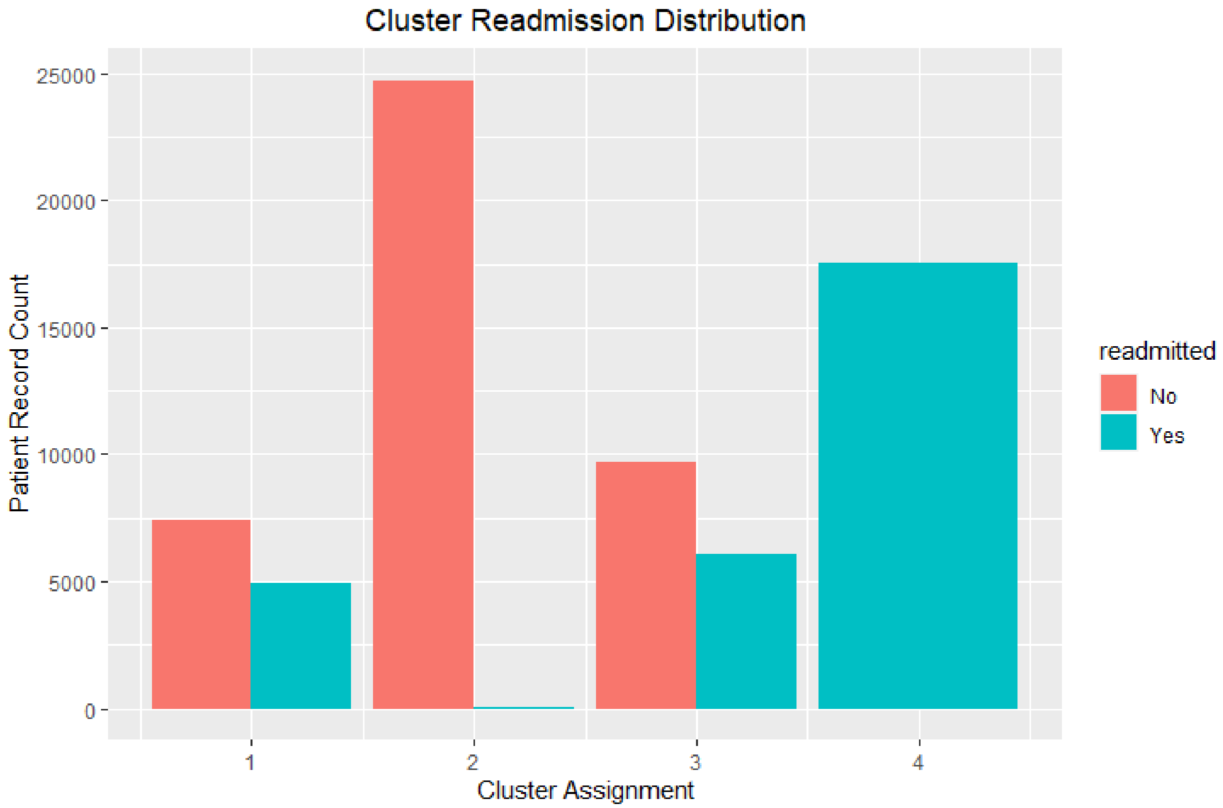


Figure 6. Readmission Distribution by K-Means Cluster Assignment

Comparing variable importance plots from the initial random forest run against the ensembled run shows the cluster feature dominating all others. There is vertical movement of other features between runs, though a few features remain consistently at the bottom of the plot. This suggests these features are do not contribute enough to the overall power of the model to be worth maintaining.

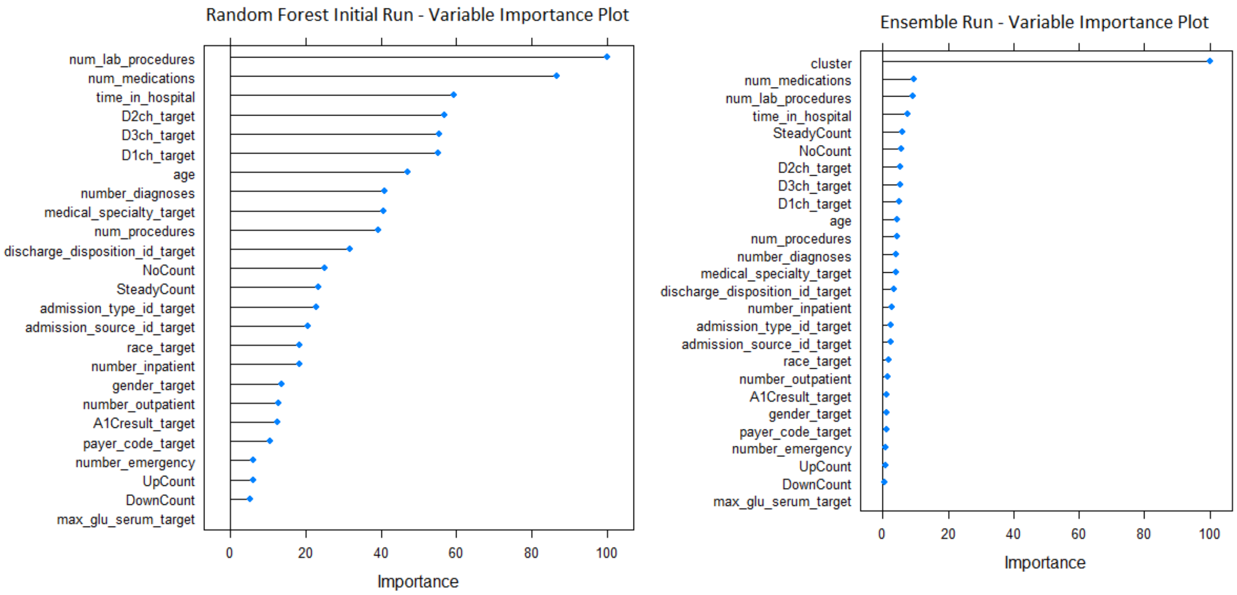


Figure 7. Random Forest Initial Run vs Ensemble Run Variable Importance Plots.

Conclusion

The ensembled approach of k-means cluster analysis and random forest produced a strong predictive model. Further investigation should be conducted to understand the patients grouped into the four clusters produced by the k-means cluster analysis. Understanding what differentiates these groups could be the key to understanding what feature, or groups of features could be used to flag patients with a high likelihood of recidivism before they are even discharged from the hospital.

Other recommendations include broadening or altering the type of hospital metrics collected for potential feature use. Some of the collected data points contributed very little to the predictive power of the model. Other patient-stay metrics such as blood pressure, temperature, diet, etc. may have a strengthening effect on the predictive power of the model.

Detailed Procedures and Findings

Data Due Diligence and Feature Engineering

The provided data set contains roughly 71,500 rows and 50 variables. Of these variables about half appear to be categorical blood sugar test results in response to the administration of specific diabetes drugs. The remaining variables are an assortment of patient demographic data points, hospital metrics, lab results, and diagnostic outcomes and metrics.

The blood sugar management drug test results have four categorical results options: Up, Steady, Down and No. Not all drugs administered have all four results present in the data set. In the case of the drugs examide, citoglipton, and glimepiride.pioglitazone all results were No, so these variables will be excluded from analysis. Other drugs have a very high instance of No, meaning they were not administered or are perhaps rarely administered and will also be excluded from analysis.

Variable	No Count	No %
acetohexamide	71517	100.00%
tolbutamide	71499	99.97%
acarbose	71316	99.72%
troglitazone	71515	100.00%
tolazamide	71488	99.96%
examide	71518	100.00%
citoglipton	71518	100.00%
glipizide.metformin	71511	99.99%
glimepiride.pioglitazone	71518	100.00%
metformin.rosiglitazone	71516	100.00%
metformin.pioglitazone	71517	100.00%

Table 1. Drug Tests Removed for Low Administration.

Since the general desired outcome of blood sugar medications is to stabilize blood sugar, medication test result variables will be pre-processed into a set of numerical features that count how many Steady, Up, Down, or No responses a patient received. This was the most major variable grouping performed on the raw data.

There are three rows with ‘Invalid’ Gender entries which will be excluded from the final data set. The Race variable has about 1,950 rows with NA entered. Given that the Race variable also contains an “Other” entry, all NAs in Race will be converted to “Other” before being one-hot encoded. Some patient records had a discharge type that indicated death while in hospital. These records were removed. A weight variable was available but highly suspicious due to the number of clearly false entries and high instance of NAs. As such, the weight variable was excluded from the final data set.

There were three variables devoted to diagnosis codes, which were converted to character representations of the chapter from which the diagnosis code originated. This method helped to further reduce feature levels by grouping very specific diagnosis codes into their over-arching medical emphasis. The field of ‘medical specialty’ also contained many unique entries which were regrouped similarly to the diagnosis codes variables. Many of the more specific medical specialties were categorized as ‘Misc’ while others such as those relating to the care of infants and children were categorized simply as ‘Pediatrics’.

Patient age was represented as being within a 10-year range and was converted to the midpoint of the represented range before being converted into a numerical feature. The other numerical variables present in the data represent counts of certain types of events, i.e. count of lab procedures, or count of mediations. Since there is some variability across these types of counts, all the numerical variables will be scaled to values between 0 and 1 as features.

These considerations reduced the data set from 50 to 26 variables to be converted to features. A 27th feature was added using the cluster assignment from k-means cluster analysis, which is discussed in further detail later in this document.

Variable Name	**Variable Category**
race	Demographic
gender	Demographic
age	Demographic
admission_type_id	Hospital Metric
discharge_disposition_id	Hospital Metric
admission_source_id	Hospital Metric
time_in_hospital	Hospital Metric
medical_specialty	Hospital Metric
num_lab_procedures	Hospital Metric
num_procedures	Hospital Metric
num_medications	Hospital Metric
number_outpatient	Hospital Metric
number_emergency	Hospital Metric
number_inpatient	Hospital Metric
diag_1	Diagnosis Metric
diag_2	Diagnosis Metric
diag_3	Diagnosis Metric
number_diagnoses	Diagnosis Metric
max_glu_serum	Lab Test
A1Cresult	Lab Test
change	Lab Test
Steady Count	Medication Metric
Up Count	Medication Metric
Down Count	Medication Metric
No Count	Medication Metric
Change Count	Medication Metric
readmitted	TARGET

Table 2. Final Feature Selection.

Finally, the ‘readmitted’ variable had three result options, <30, >30, and No. For the purposes of this study, the focus was on any readmittance, so the target variable itself needed to be recoded to 1 and 0 where 1 includes any readmittance (<30 or >30) and 0 for No entries.

All features were target encoded for the final runs of this study. Preliminary runs used a mix of one-hot encoding and target encoding, but target encoded features ran faster and lessened constraints felt by hardware and processing deficiencies.

Methods

This study focused on two supervised machine learning algorithms and one unsupervised method. The unsupervised method of K-Means Cluster Analysis was used primarily to create a cluster variable to be used as a feature in one or both of the supervised learning algorithm (thereby ensembling the two super/unsupervised methods). K-means clustering also offered an exploratory, untargeted perspective on the data. The supervised approaches included Random Forests and Neural Networks.

These two algorithms respectively used out-of-bag error and k-folds training methods. Both supervised methods used the Held-Out method for evaluation and their training sets contained 70 percent of the original data.

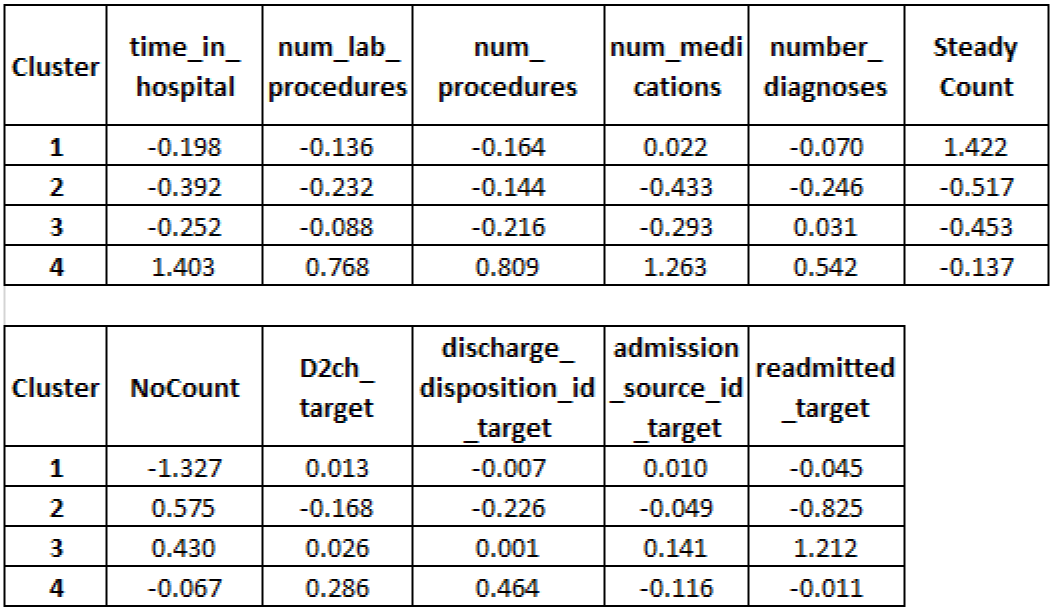
K-Means Cluster Analysis

This method used the same set of target-encoded features as the random forest and neural network models and included a target-encoded version of the target variable, readmitted. Since the goal of this study was focused on the target of readmittance, minimal time and effort was spent conducting and evaluating the clusters produced by k-mean analysis. The default Hartigan-Wong algorithm was used.

The feature combination that produced the highest silhouette width (Si) ratio, well balanced cluster sizes, and well distributed variable means across clusters.

Cluster 1	**Cluster 2**	**Cluster 3**	**Cluster 4**
15,794	24,772	17,564	12,301

Table 3. Cluster Row Counts.



Cluster	time_in_ hospital	num_lab_ procedures	num_ procedures	num_medi cations	number_ diagnoses	Steady Count
1	-0.198	-0.136	-0.164	0.022	-0.070	1.422
2	-0.392	-0.232	-0.144	-0.433	-0.246	-0.517
3	-0.252	-0.088	-0.216	-0.293	0.031	-0.453
4	1.403	0.768	0.809	1.263	0.542	-0.137

Cluster	NoCount	D2ch_ target	discharge_ disposition_id _target	admission _source_id _target	readmitted _target
1	-1.327	0.013	-0.007	0.010	-0.045
2	0.575	-0.168	-0.226	-0.049	-0.825
3	0.430	0.026	0.001	0.141	1.212
4	-0.067	0.286	0.464	-0.116	-0.011

Table 4. K-Means Cluster Analysis Means.

The cluster plot shows relatively well-defined groups, with some overlap occurring. This overlap can also be seen in Figure 6 in the Executive Summary. Cluster assignments were joined back to the feature encoded data set to be used in the random forest model as a new feature.

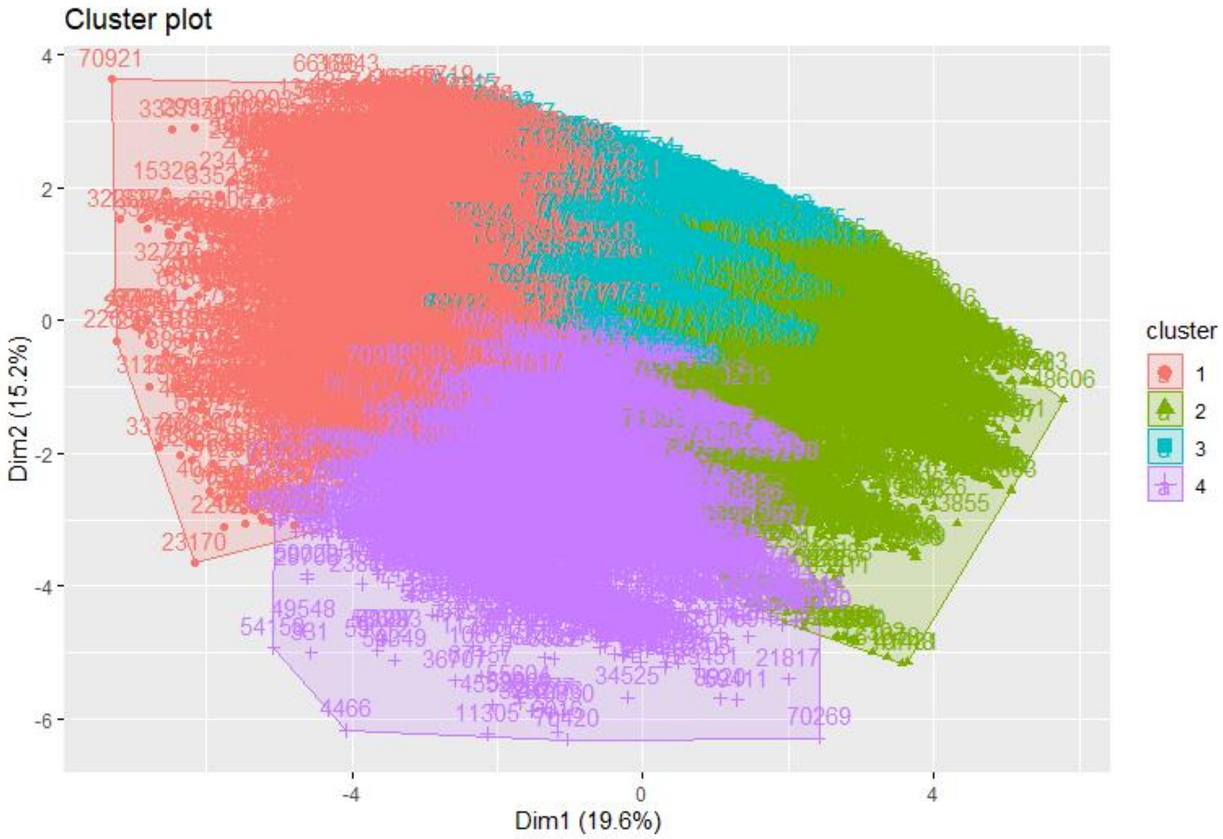


Figure 8. K-Means Cluster Plot.

Random Forest and Ensembling Results

The initial random forest run produced an Area Under the Curve (AUC) of .66 and a Receiver Operating Characteristic (ROC) curve that arches delicately over the ab line, which represents random guessing. When this model was re-run with the cluster assignments from the K-mean Cluster analysis included as a target encoded feature, the AUC increased to .939, a vast improvement. See Figure 5 in the Executive Summary for the ensembled method’s ROC curve.

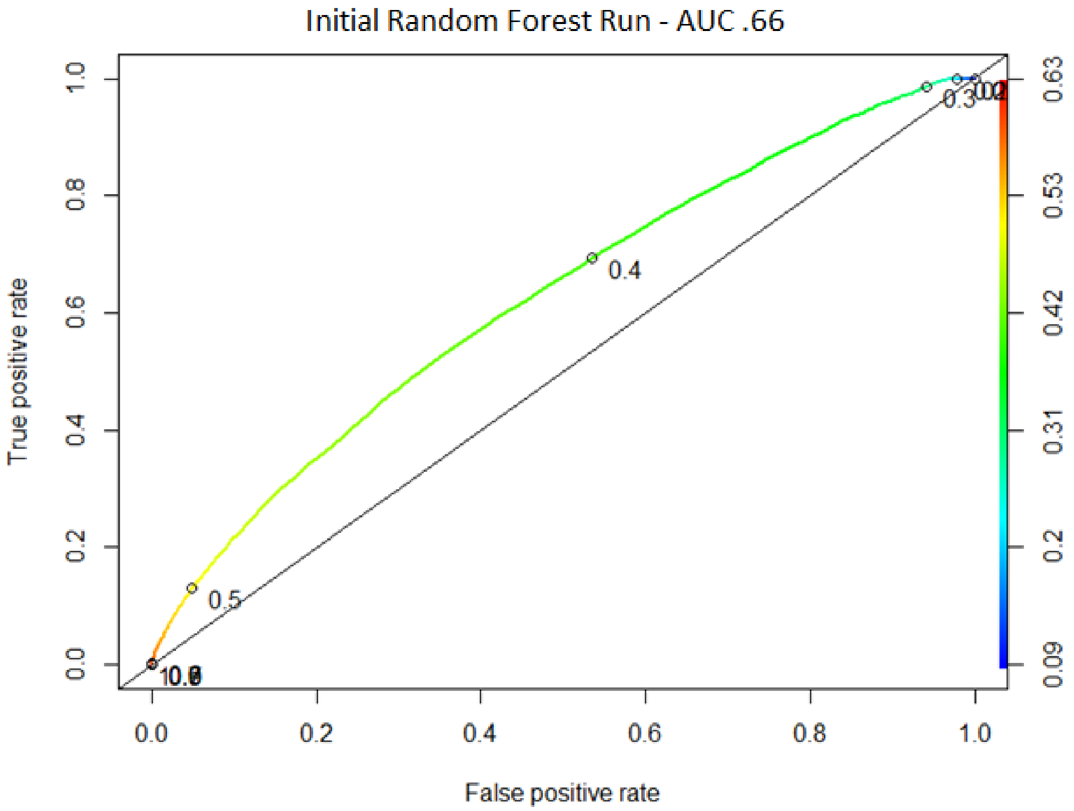


Figure 9. ROC Curve for Initial Random Forest Model Run

The calibration curve produced by the initial random forest model (represented by the blue line in Figure 9 below) show tempered confidences all the way up to the .95 bin, where it drops. This drop is due to a lack of predictions falling within that .95 confidence bin, which is acceptable. Platt Scaling, down sampling, and up sampling of this random forest model were conducted with no improvement to the AUC or calibration curve. Please note that SMOTE sampling was eliminated as an option everywhere as it took too long to run.

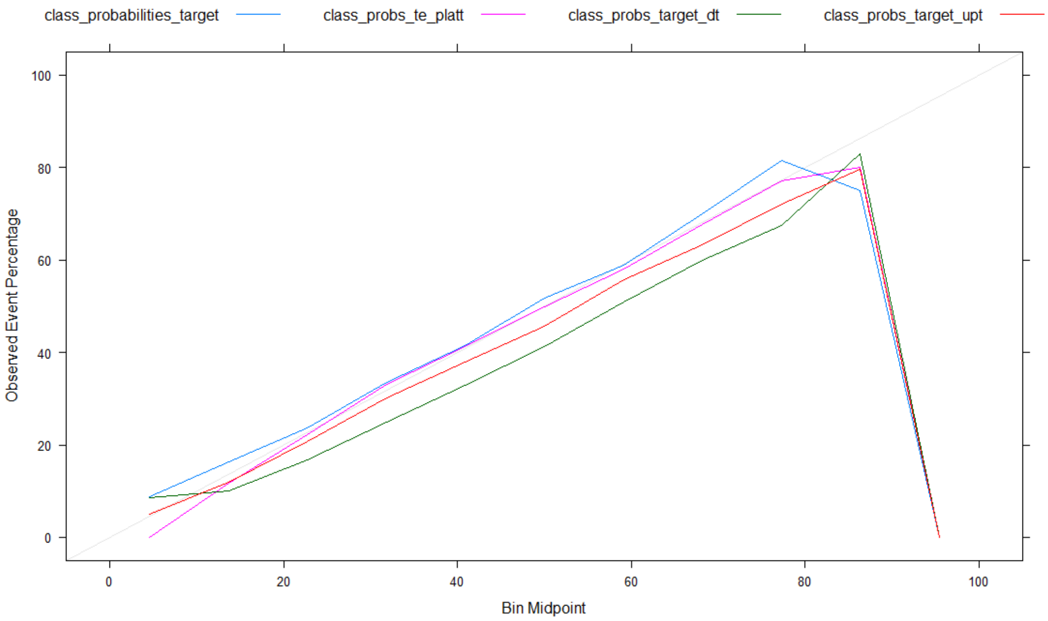


Figure 10. Random Forest Initial Run Calibration Curve with Corrective Sampling Runs

When the random forest model was re-run to include the encoded cluster feature, the previously mentioned marked improvement in AUC was seen, and the drop at the .95 mindpoint bin in the calibration curve was removed. Predictions for the ensembled run are varyingly over confident until around the .75 midpoint bin where predictions dip toward under confidence. Again, Platt Scaling was used to try to smooth this curve, with similar results. Well calibrated predictions confidences are not required for the context of this study but should be sought when possible.

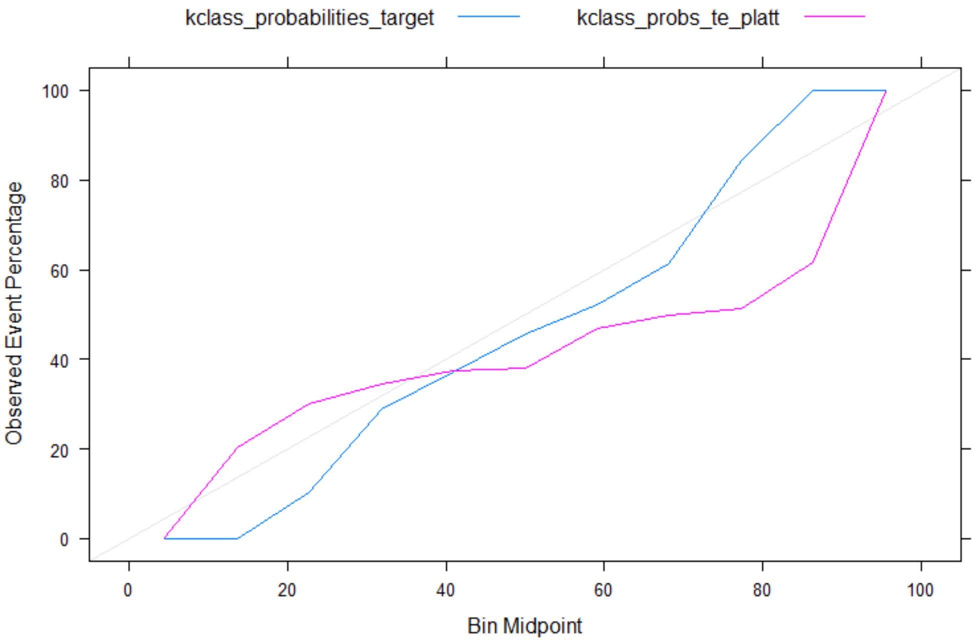


Figure 11. K-means Random Forest Run Calibration Curve with/out Platt Scaling

To view the differences between the initial random forest run and the ensembled run see Figure 7 in the Executive Summary.

Neural Network Results

The neural network model produced an AUC of .6478 and an ROC curve that also arches delicately over the ab line. The goal of the ROC curve is for it to bend as close to the upper left of the plot. As with the initial run of the random forest model, this result could be improved. However, because such a strong result was produced with the ensembled random forest model, there was little reason beyond simple curiosity to pursue improving this model.

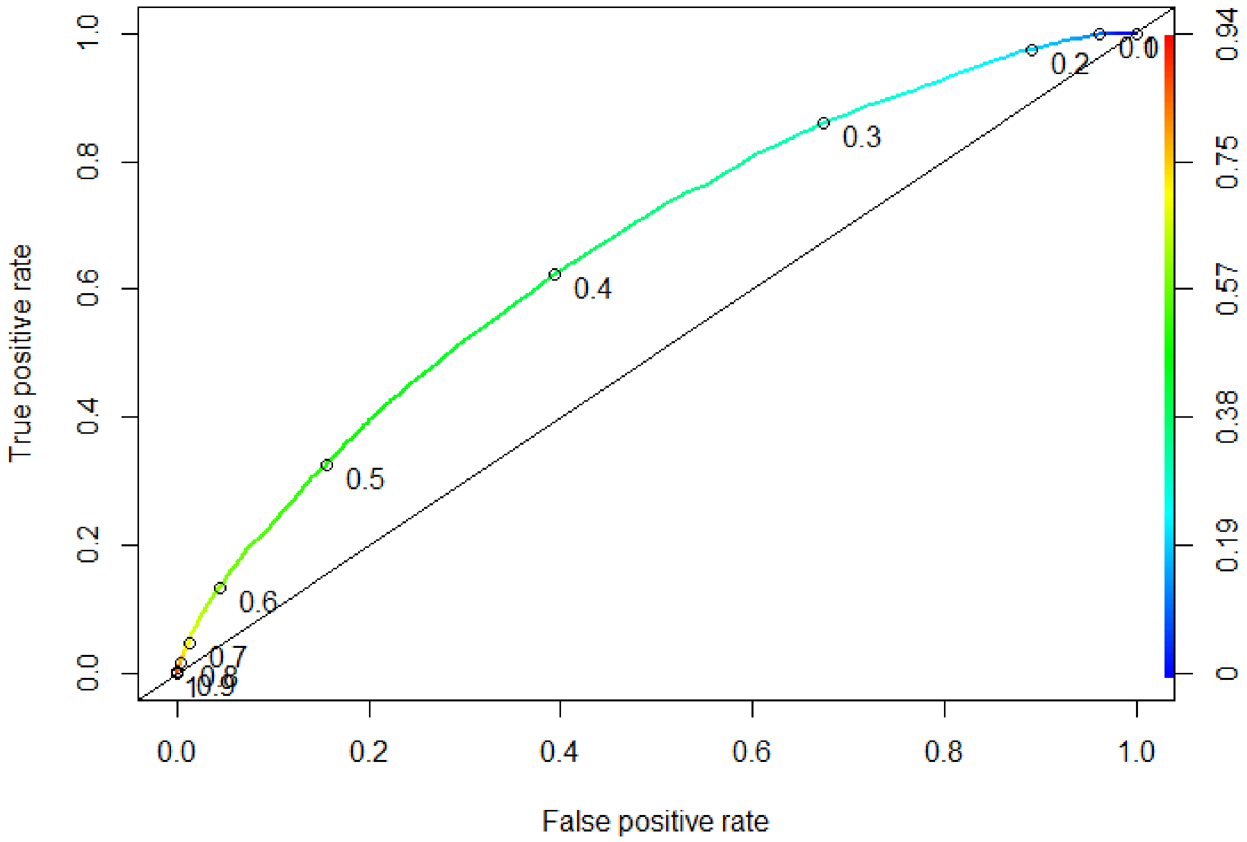


Figure 12. Neural Network Initial Run ROC Curve

Curiosity could not be resisted however, and an ensembled k-means and neural network run produced an AUC of .97 and an ROC very similar to the one produced by the ensembled random forest model (Figure 5).

The calibration curve produced by the initial run of the neural network model (not pictured) was similar to the initial run of the random forest's calibration curve, with the same drop at the .95 midpoint bin. The curve produced by the k-means run of the neural network is very different from that of the random forest. Platt scaling was run on the k-means neural network calibration curve and provided slight smoothing as well as a shift of some of its over confidences to under confidences in the lower midpoint bins.

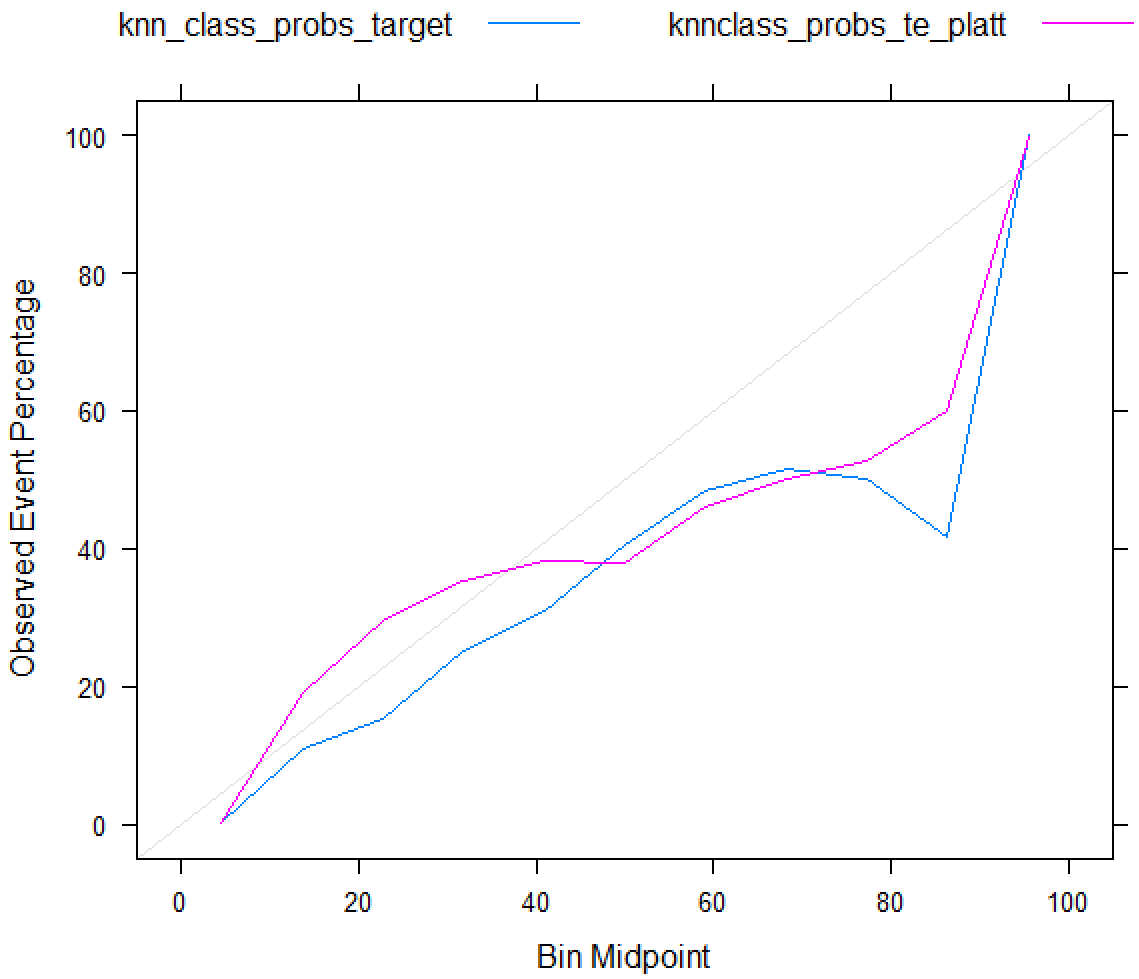


Figure 13. K-Means Neural Network Calibration Curve Before and After Platt Scaling

The variable importance plot produced by the k-means neural network model shows a greater shuffling of importance compared to the variable importance plots produced by the k-means random forest (Figure 7). This is most likely due to the obvious fact that different algorithms were used, though why random forest select the feature ranking it did, versus that of the neural network is unknown.

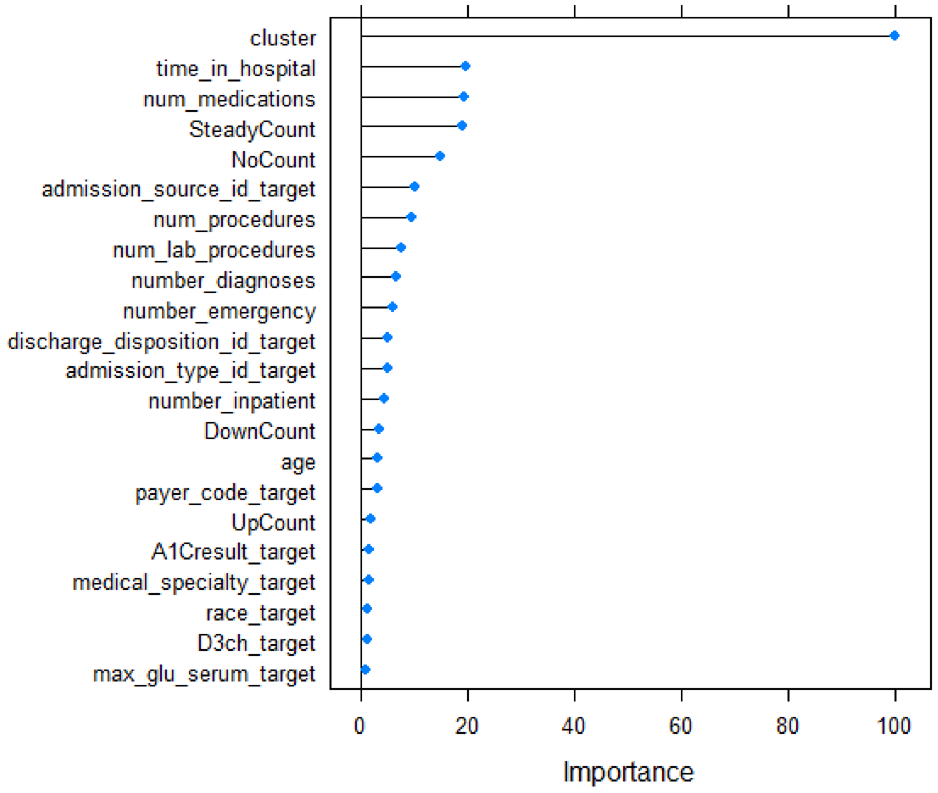


Figure 14. K-means Neural Network Feature Importance Plot

Conclusion and Recommendations

The two algorithms' initial runs produced similar results of AUCs around .66. The inclusion of the k-means cluster assignment as a feature in both of the supervised machine learning algorithms result in great predictive improvements and relatively smooth confidence calibration curves. The ensembled neural network produced a slightly higher AUC than the ensembled random forest, but took longer to run and produced a more erratic calibration curve. This is why the ensembled random forest model was chosen as the best model produced by this study.

Because both models used the same set of variables, it's not surprising that they produced relatively similar results with or without the inclusion of the k-means cluster feature. One of the primary differences between the two supervised learning methods was the order of features in the feature importance plots. As noted in the conclusion of the Executive Summary, many, if not all the variables in the upper portion of all feature importance plots were hospital-stay-related metrics. This suggests that the answer to a diabetes patient's likelihood of recidivism is hidden within the details of their hospital stay and not among their individual demographic information. That being said, this data set offered very little in the way of patient demographic information so it's hard to know how much it really contributes to the prediction power.

This study was limited by hardware and processing power that meant only algorithms that ran in under two hours were used. The training set data contained just under 50k rows, which is relatively small in the field of machine learning. Future iterations of this study will likely remain limited to training sets of a similar size and similar algorithms and class imbalance sampling methods unless a more power machine can be procured.

Recommendations for future iterations of this study include broadening the type of hospital stay data and patient demographic data collected. From an analysis standpoint, Principle Component Analysis might be considered as a method of identifying which variables contribute the most to understanding diabetes patient hospital recidivism.

Appendix A contains all code used in this study.

References

Centers for Disease Control and Prevention. (2020). National Diabetes Statistics Report, 2020. Retrieved September 10, 2020, from https://www.cdc.gov/diabetes/library/features/diabetes-stat-report.html

Swami, J., Donihi, A., Siminerio, L., French, E. K., Delisi, K., Hlasnik, S. D., Patel, N., Pinkhasova, D., Rubin, D., Korytkowski, M. (2018, July). *Readmission and Comprehension of Diabetes Education at Discharge (ReCoDED Study*. Retrieved September 10, 2020 from https://diabetes.diabetesjournals.org/content/67/Supplement_1/147-LB

Kho, J. (2018, October 19). *Why Random Forest is My Favorite Machine Learning Model.* Retrieved September 10, 2020 from https://towardsdatascience.com/why-random-forest-is-my-favorite-machine-learning-model-b97651fa3706