Predicting readmission in diabetic patients

Problem Statement

What factors of a diabetic's health profile are most likely to predict that a patient will have at least 1 readmission for any acute care admission and can be researched in the next 2 years to explore alternative care and treatment options to prevent readmission?

After the Affordable Care Act, acute care readmissions have significantly impacted hospital financials. Diabetics in particular are prone to readmission as the disease affects so much of the body and complicates care. With healthcare moving to a pay-for-performance model, providers often do not get compensated for the readmission and total compensation is decreased when they occur. In addition to the obvious patient safety and quality of life concerns, preventing readmissions can also protect a hospital's revenue.

This analysis will be considered successful if it delivers at least 1 hypothesis for future research into what can be done during a diabetic's acute care stay to prevent a future readmission.

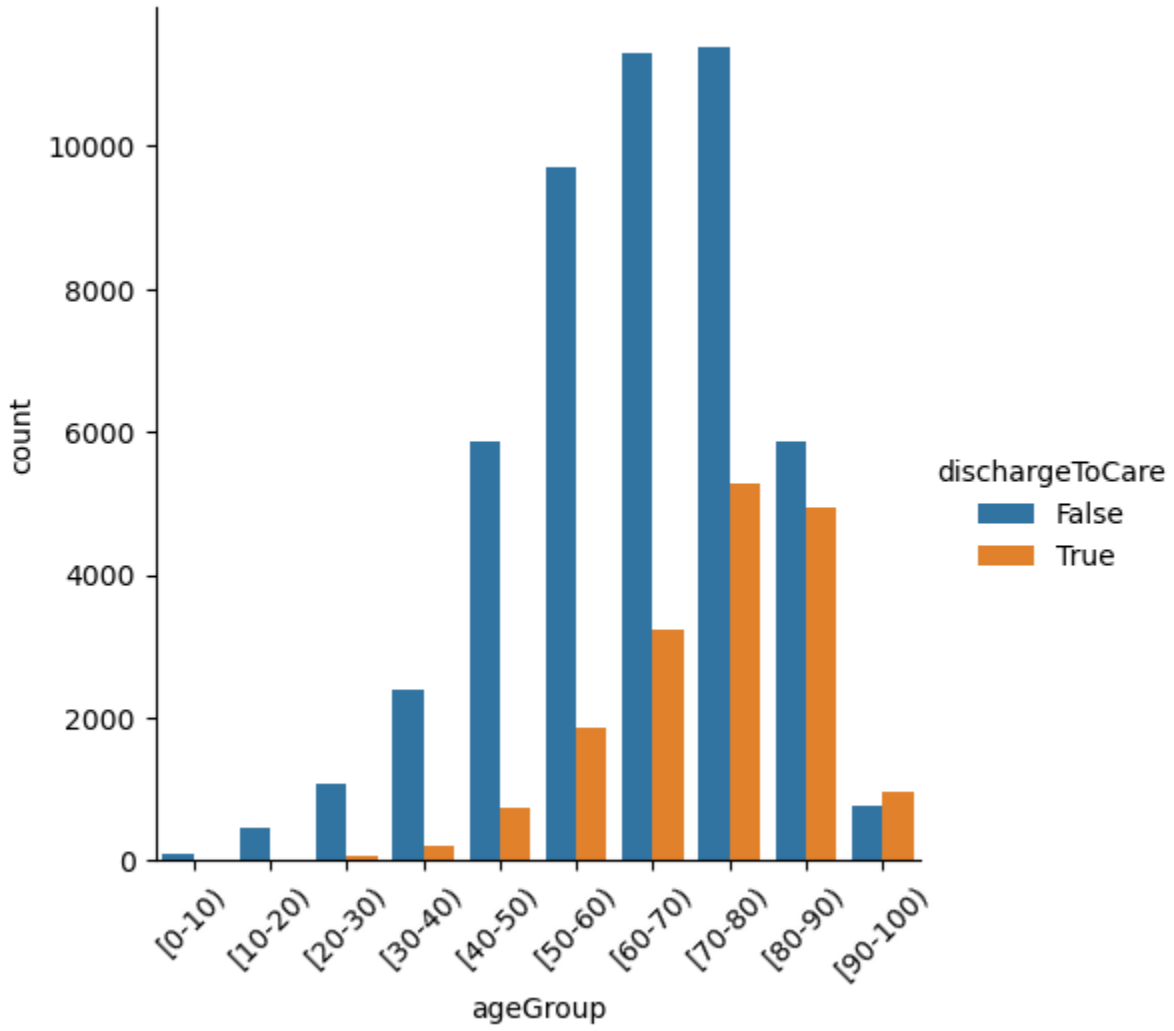
Data Wrangling

The data used for this analysis was provided by UCI's machine learning repository. It was donated by the authors of this paper. They had collected it from Cerner's HealthFacts database for this research project. The original shape of the data was 101,766 rows and 50 columns.

As part of data cleaning, several features were dropped including patient weight, payor, medical specialty of the admitting clinician and many features describing patient prescriptions. These had constant, near constant or missing values. Admission and discharge type and source codes were also combined into larger hierarchical categories. Outliers in other features were handled by deletion, or transformation. The final shape of the data set was 101,727 rows and 38 columns.

Exploratory data analysis

Analysis started with an examination and correction of racial bias. Races other than caucasian and african american weren't represented well and removed. Caucasian observations were removed at random to match the ADA's reported distribution. Age bias was the next item checked. It was not present, but this did lead to a new feature describing whether or not a patient was discharged to care and an insight that as patients age, they are more likely to be discharged to care and patients that are discharged to care are more likely to not be readmitted.



Analysis continued by targeting interactions with providers in the year prior to the first admission. Outpatient, ED and inpatient encounters were recorded. These started as numerical features, but were transformed to boolean as they were all heavily imbalanced in favor of “1”. Next were the numerical descriptions of the first admission. These were all numerical and had mostly good distributions. Some transformations were necessary to reduce outliers or correct data.

The top 3 diagnoses were recorded for each observation. These had high cardinality that needed to be reduced. Word2vec was applied to sentences of the diagnosis and Kmeans clustering applied to the resulting vectors. 9 diagnosis clusters were created representing common groupings.

Finally, prescriptions for each patient were addressed. These were highly dimensional and sparse. They were combined into 4 drug classes based on the intended effect of the drug.

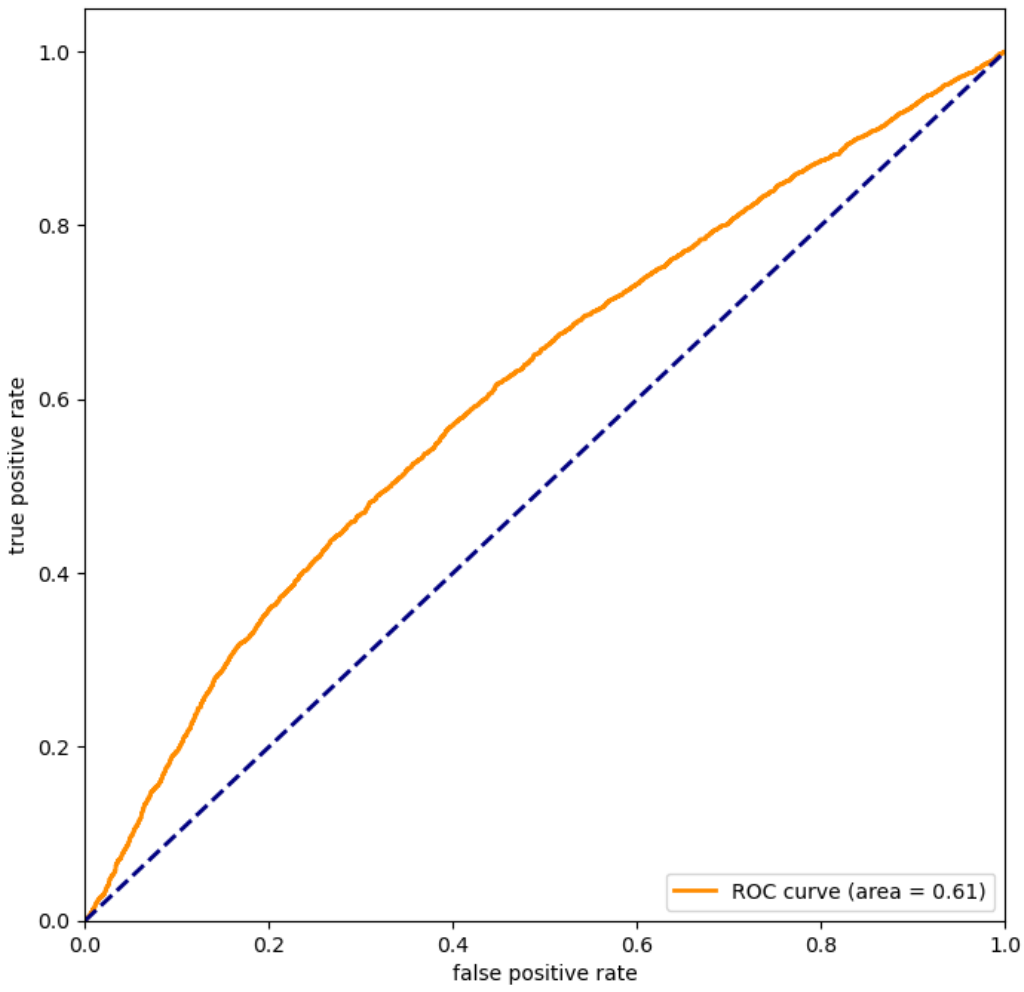
Modeling

The data was prepared for modeling by one hot encoding the categorical features, scaling the numerical features and creating a train test split of 80/20 stratified on the readmission variable. The training set had 52,979 rows and the test had 13.245.

AUC was selected as the success metric for models. Because of report fatigue in clinical settings, false positives are important to avoid, but false negatives are costly in both patient safety and quality of life as well as CMS reimbursement. AUC provides information on the balance between these.

Four model types were tried: naive bayes, logistic regression, random forests and ada boosted random forests. Each model had several common hyperparameters tuned with grid search and common values for datasets of this size. Naive bayes provided a baseline AUC of 0.61 that other models were judged against.

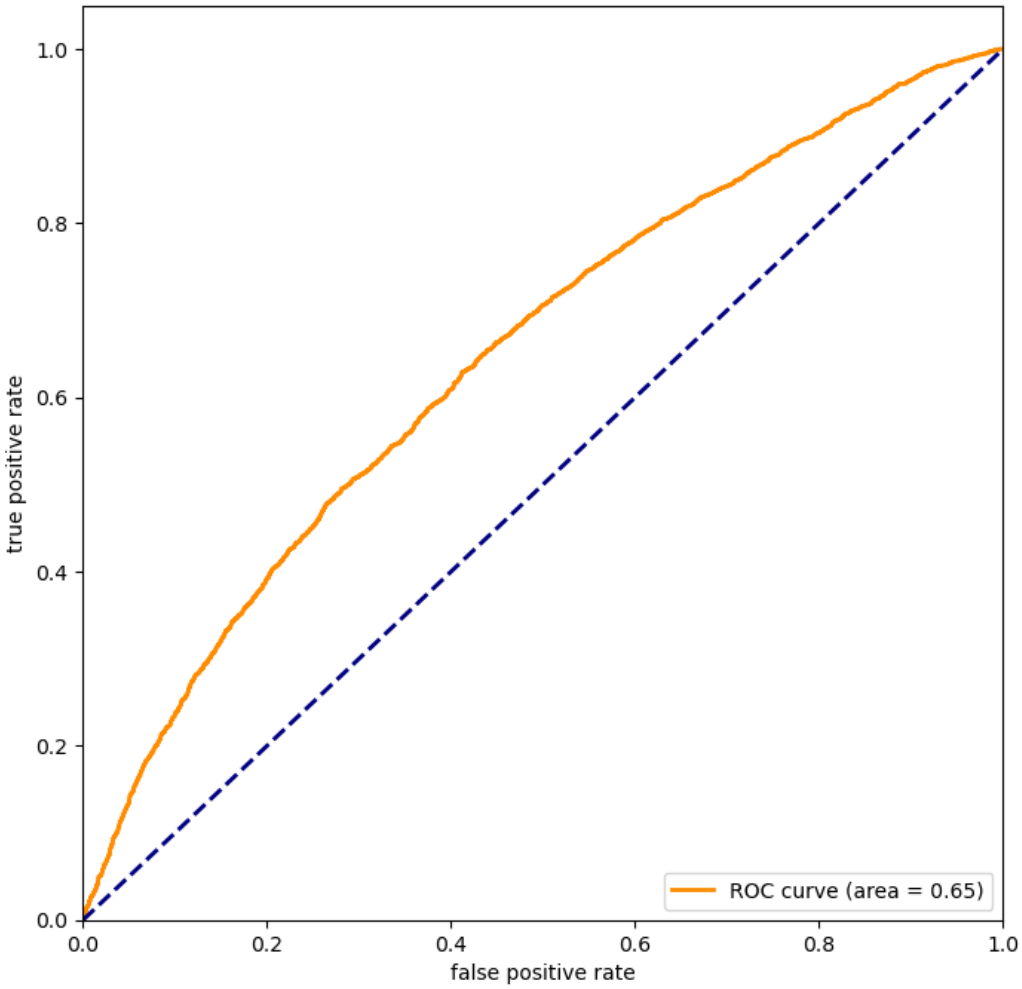
```
Best parameters: {}
```




```
GaussianNB()
```

The random forest model with a depth of 20 and 10 samples per split was the final model selected with an AUC of 0.65.

```
Best parameters: {'max_depth': 20, 'min_samples_leaf': 1, 'min_samples_split': 10, 'n_estimators': 200}
```



	feature	importance
1	num_lab_procedures	0.124377
3	num_medications	0.105879
0	time_in_hospital	0.073771
8	inpatientTF	0.061215
4	number_diagnoses	0.061063
2	num_procedures	0.051217
7	emergencyTF	0.024267
6	outpatientTF	0.022997
11	gender_Male	0.020099
10	race_Caucasian	0.019506