Predicting Readmission of Diabetes Patients

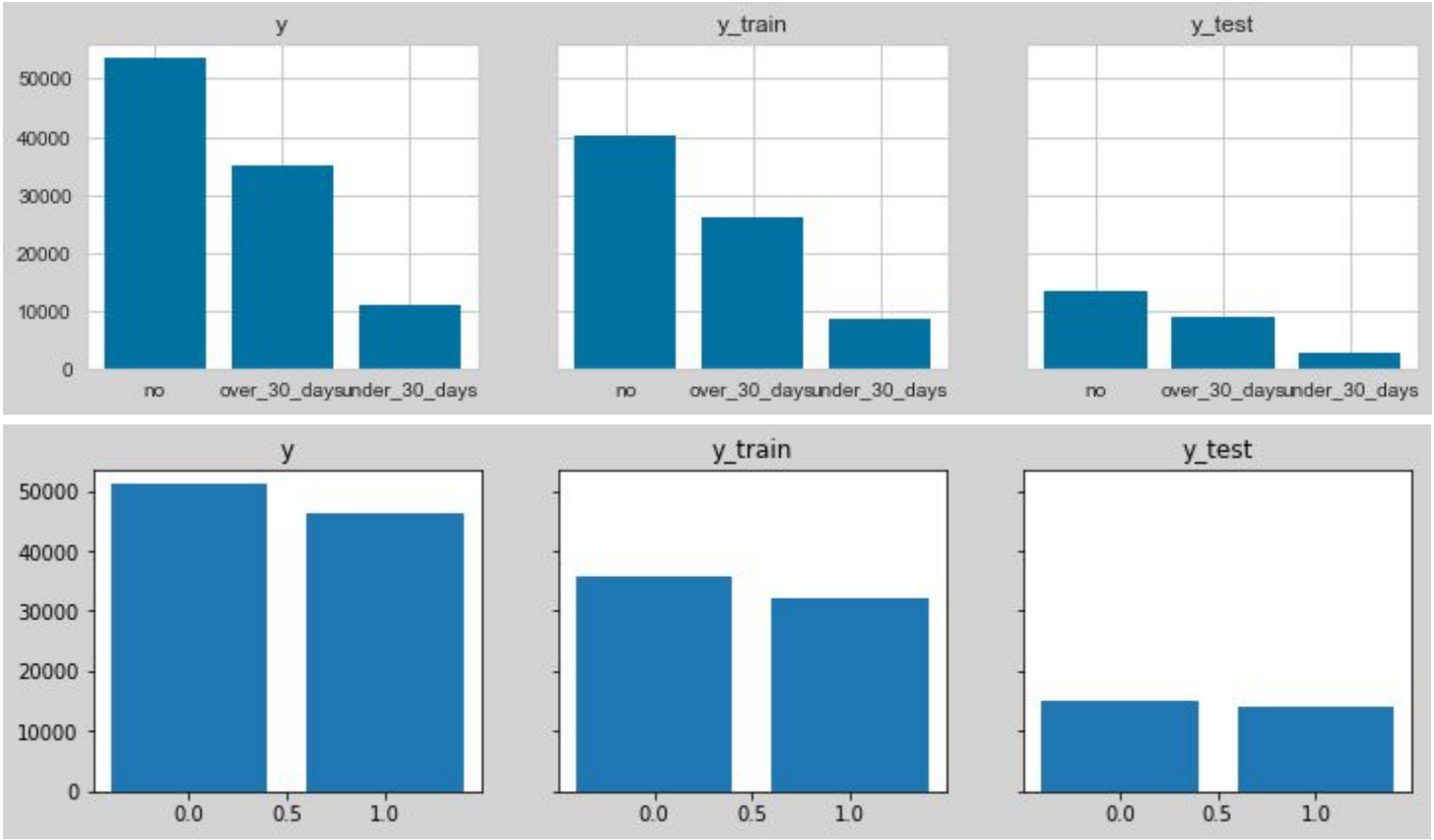
Kyle Hayes, Kaleb Nyquist

Question & Dataset

In this project, we see if were are able to predict (at a rate better than purely random) the readmittance of diabetes patients given data both about the patient and the treatment the patient received.

The dataset we use is *Diabetes 130-US hospitals for years 1999-2008*, available in the Public Domain through UCI's Machine Learning Repository and Kaggle. This dataset was first created for the open access journal article by Strack et al. "Impact of HbA1c Measurement on Hospital Readmission Rates", itself a subset of the Health Facts Database curated by the Cerner Corporation. The creation of this 100,000+ record dataset was a query of diabetes diagnoses among 74,000,000 visitation records meeting certain criteria (such as length of stay being 24 hours or greater, or laboratory examinations being performed on the patients).

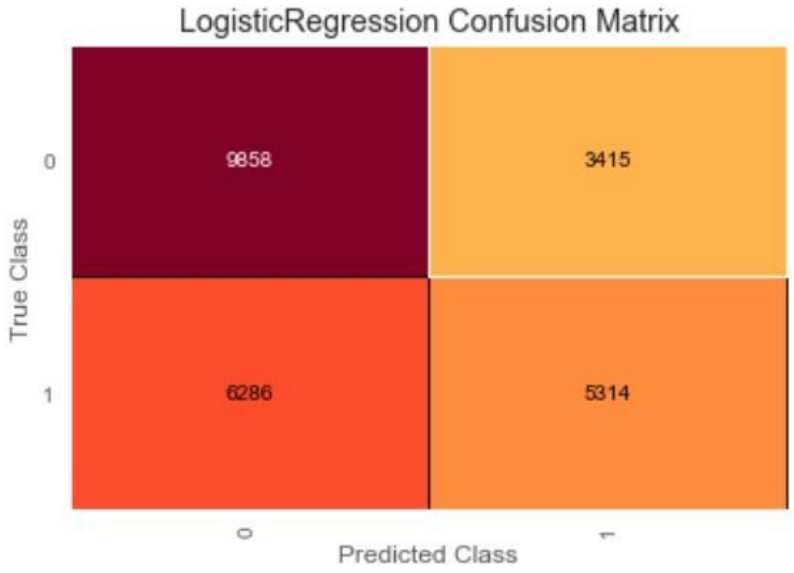
Data Cleaning, Note #1: “Natural Balancing”



Features

Demographics	Visitation Data	Diagnoses	Medication Data	Diabetes Specific Information	Testing
'female', 'Age', 'african_american', 'asian', 'hispanic', 'other',	'time_in_hospital', 'num_lab_procedures', 'num_procedures', 'num_medications', 'number_outpatient', 'number_emergency', 'number_inpatient', 'number_diagnoses', 'change',	**ICD #1-19**	'Num_meds', 'num_down', 'Num_up',	'Diabetesmed', 'glu_200_300', 'glu_over_300', 'Glu_norm', 'insulin_down', 'insulin_steady', 'insulin_up'	'a1_7_8', 'a1_norm', 'al_over_8',
			Up, Down or Steady: Metformin, Repaglinide, Glimepiride Glipizide Glyburide Pioglitazone Rosiglitazone		

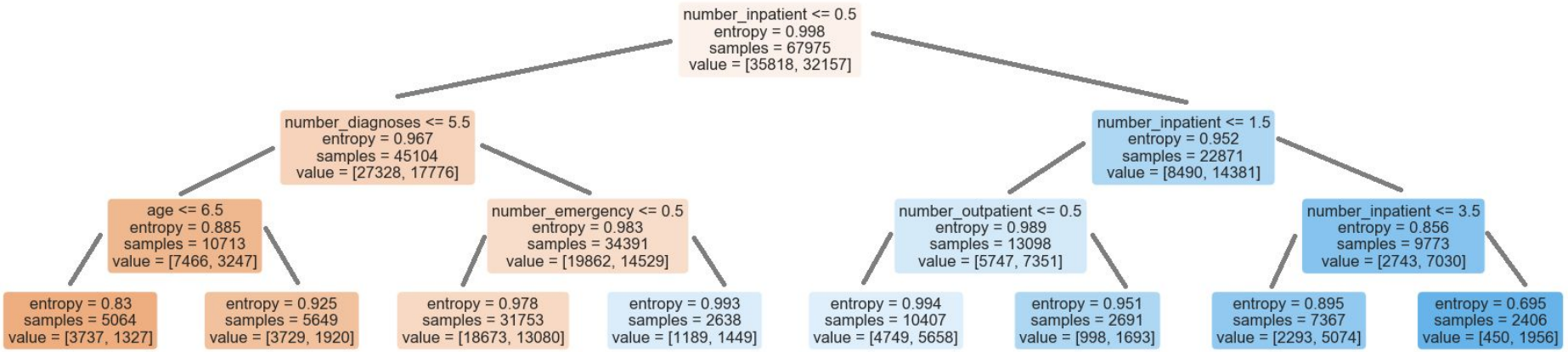
Data Cleaning, Note #2: “I Don’t See Dead People”

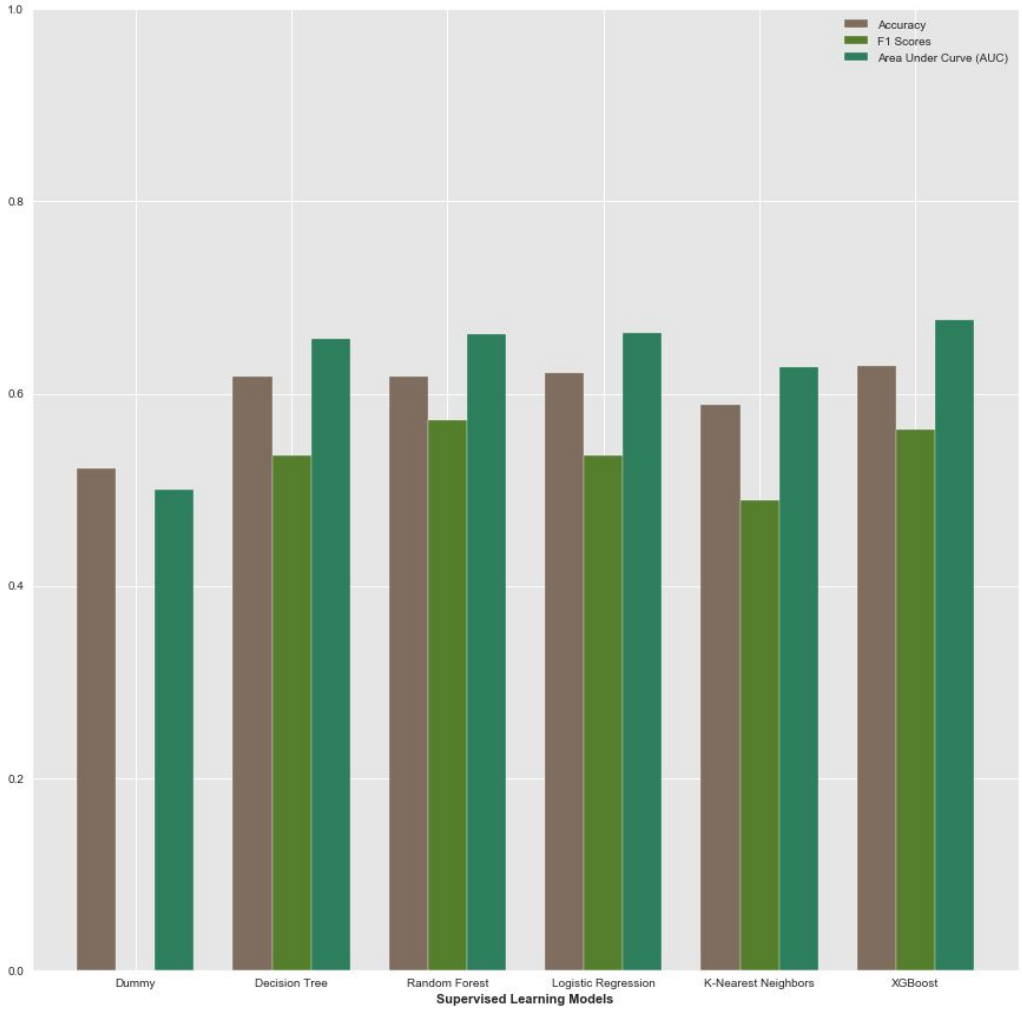


Supervised Learning Techniques Used

- Dummy
- Decision Tree
- Random Forest
- Logistic Regression
- K-Nearest Neighbors
- XGBoost

...a peek inside the “black box”




Accuracy
F1 Scores
Area Under Curve (AUC)
1.0
0.8
0.6
0.4
0.2
0.0
Dummy
Decision Tree
Random Forest
Logistic Regression
K-Nearest Neighbors
XGBoost
Supervised Learning Models