Hospital Cost Reports
FY 2014-2018

Jon Kim | Springboard

Context

Background

- Each year, Medicare-certified hospitals and providers are required to submit an annual cost report, called CMS-2552-10 form to Medicare Administrative Contractor (MAC)
- MAC will then report the data to Healthcare Cost Report Information System (HCRIS).
- CMS compiles these data and provides a dataset that aggregates at the hospital-level, and published them to data.cms.gov

Goal

- Target Variable: Net Income
- Choose a final regression model to predict the target variable at the hospital-level

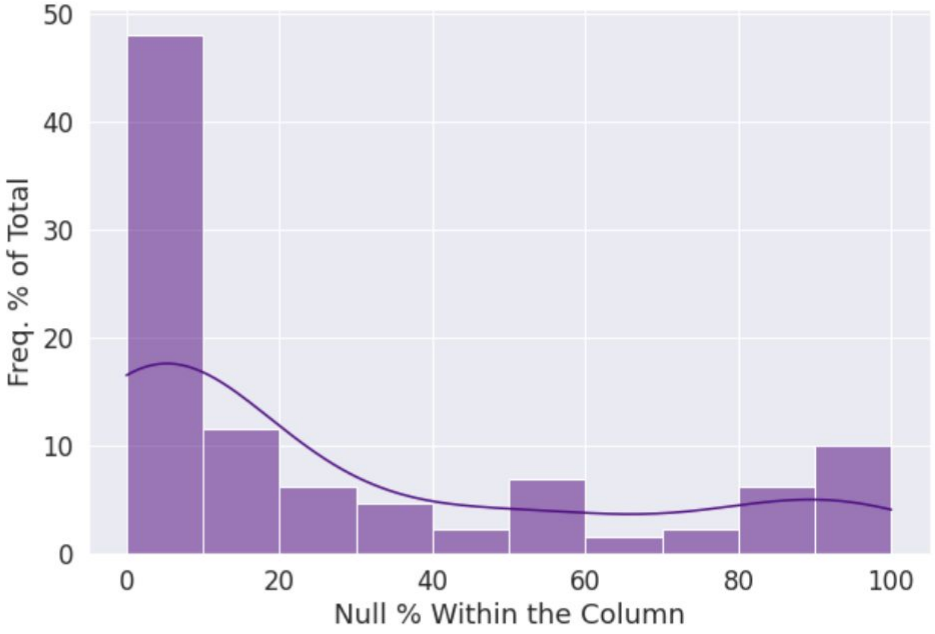
- I chose this as the target variable since Net Income (as opposed to purely revenue) better indicates profitability, and the ultimate goal of this project is to build a predictive model for profitability.

Data Wrangling

Null Cleaning

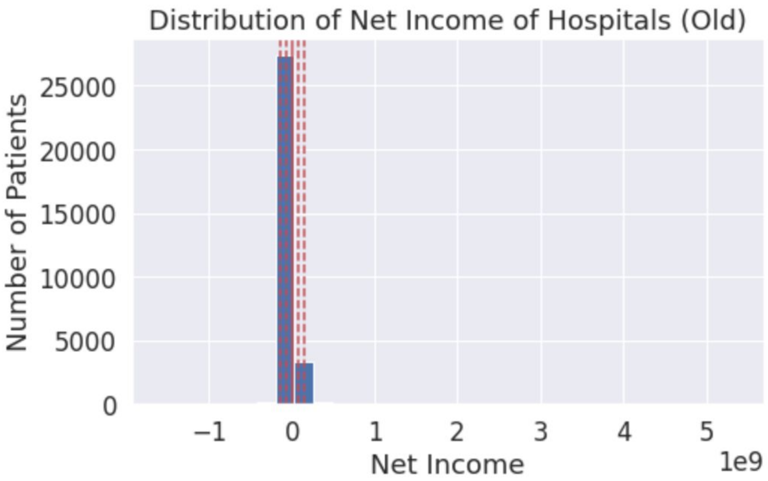
- High number of nulls, though thankfully, the majority of the columns had less than 20% in null values
- Set of four arbitrary bins of null percentages: >= 80%, 40-80%, 10-40%, and <10%.
- The number of columns reduced from 129 to 115

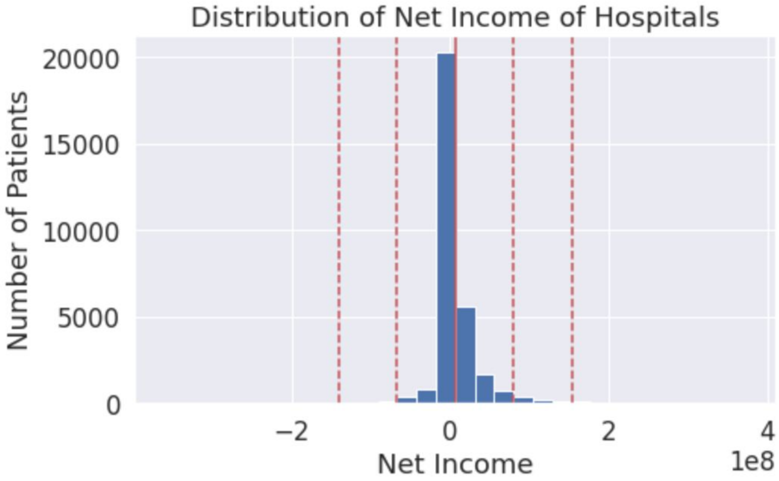
(Fig. 2) Null % Histogram:
What does the percentage distribution look like for the "Null %" column?



Target Distribution

- The distribution of the target feature itself - Net Income - originally had a strong skew to the right due to strong outliers
- Filtered out any hospitals with a Net Income z-score of less than 5
- Resulted in a much more visibly "bell-shaped" curve





Feature Correlation

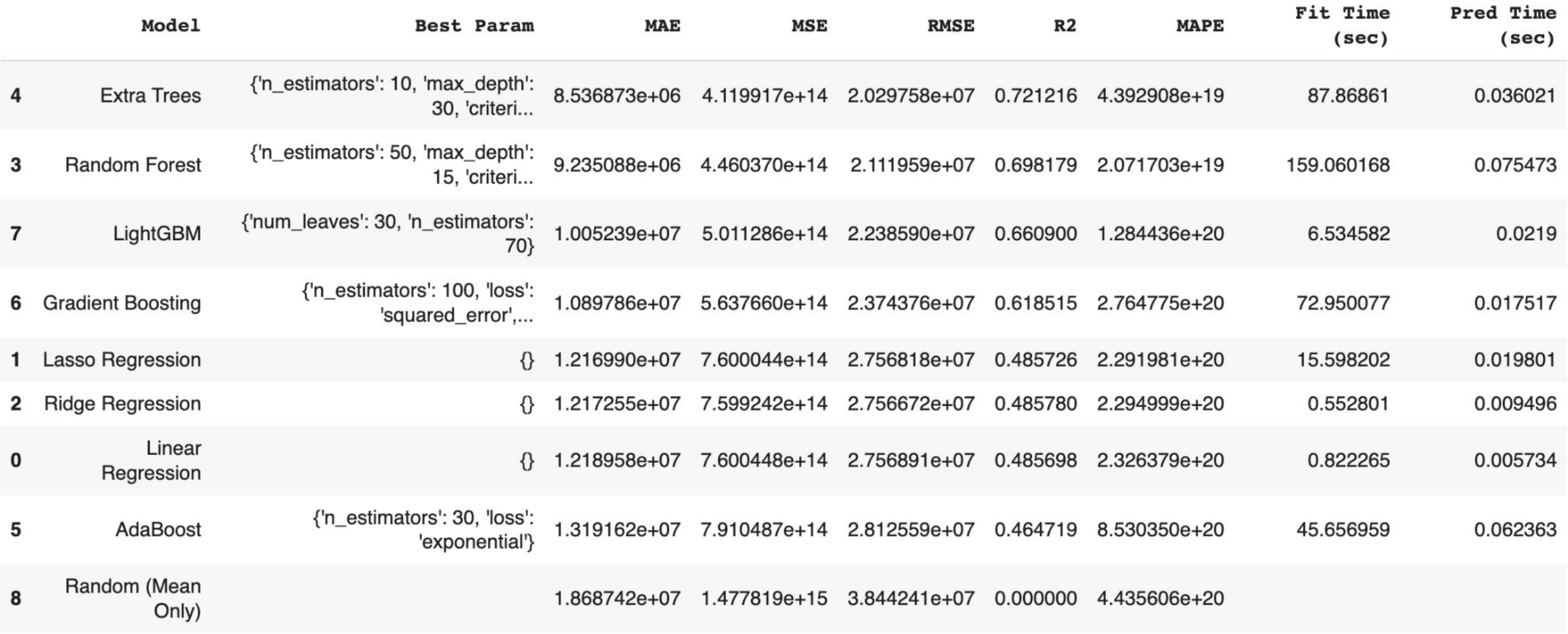
- Several features were perfectly correlated with at least one other feature. Their removal reduced the total feature set by only five.
- Other miscellaneous adjustments

Scaling

- The categorical features of the final dataset were dummified and scaled for both the training and testing sets.

Modeling

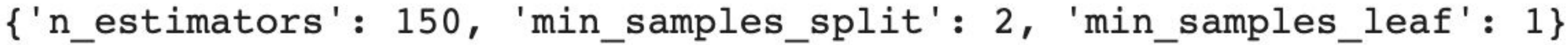
Model Comparison



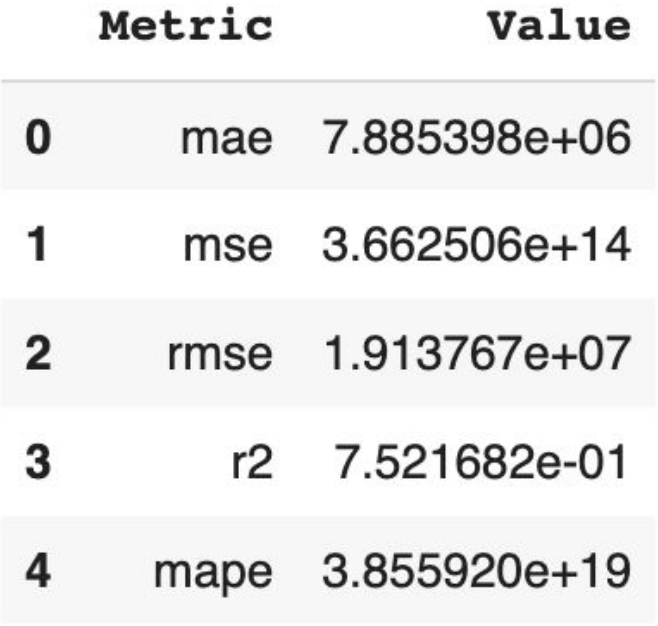
	Model	Best Param	MAE	MSE	RMSE	R2	MAPE	Fit Time (sec)	Pred Time (sec)
4	Extra Trees	{'n_estimators': 10, 'max_depth': 30, 'criteri...	8.536873e+06	4.119917e+14	2.029758e+07	0.721216	4.392908e+19	87.86861	0.036021
3	Random Forest	{'n_estimators': 50, 'max_depth': 15, 'criteri...	9.235088e+06	4.460370e+14	2.111959e+07	0.698179	2.071703e+19	159.060168	0.075473
7	LightGBM	{'num_leaves': 30, 'n_estimators': 70}	1.005239e+07	5.011286e+14	2.238590e+07	0.660900	1.284436e+20	6.534582	0.0219
6	Gradient Boosting	{'n_estimators': 100, 'loss': 'squared_error',...	1.089786e+07	5.637660e+14	2.374376e+07	0.618515	2.764775e+20	72.950077	0.017517
1	Lasso Regression	{}	1.216990e+07	7.600044e+14	2.756818e+07	0.485726	2.291981e+20	15.598202	0.019801
2	Ridge Regression	{}	1.217255e+07	7.599242e+14	2.756672e+07	0.485780	2.294999e+20	0.552801	0.009496
0	Linear Regression	{}	1.218958e+07	7.600448e+14	2.756891e+07	0.485698	2.326379e+20	0.822265	0.005734
5	AdaBoost	{'n_estimators': 30, 'loss': 'exponential'}	1.319162e+07	7.910487e+14	2.812559e+07	0.464719	8.530350e+20	45.656959	0.062363
8	Random (Mean Only)		1.868742e+07	1.477819e+15	3.844241e+07	0.000000	4.435606e+20		

Note: The empty braces under "Best Param" for the Linear Regression and variants mean that the default parameters were used.

Final Model: Extra Trees Regressor

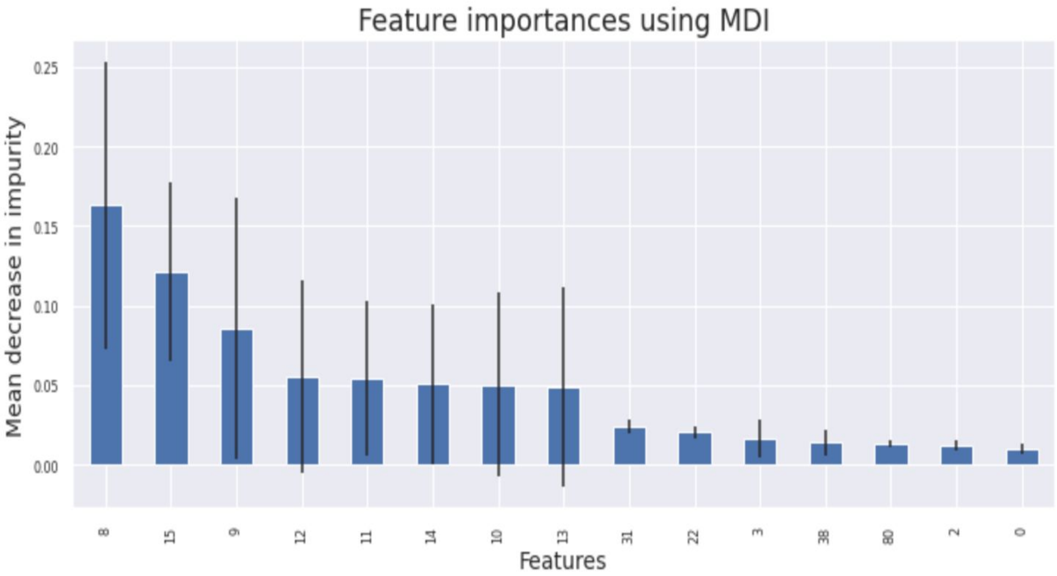


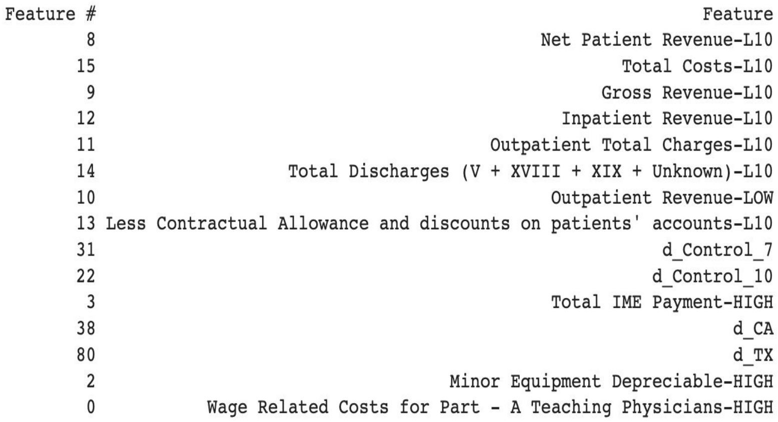
```
{'n_estimators': 150, 'min_samples_split': 2, 'min_samples_leaf': 1}
```



	Metric	Value
0	mae	7.885398e+06
1	mse	3.662506e+14
2	rmse	1.913767e+07
3	r2	7.521682e-01
4	mape	3.855920e+19

Feature Importances





```
Feature #                                                         Feature
        8                                         Net Patient Revenue-L10
       15                                                 Total Costs-L10
        9                                               Gross Revenue-L10
       12                                           Inpatient Revenue-L10
       11                                     Outpatient Total Charges-L10
       14                  Total Discharges (V + XVIII + XIX + Unknown)-L10
       10                                          Outpatient Revenue-LOW
       13 Less Contractual Allowance and discounts on patients' accounts-L10
       31                                                     d_Control_7
       22                                                    d_Control_10
        3                                          Total IME Payment-HIGH
       38                                                            d_CA
       80                                                            d_TX
        2                                   Minor Equipment Depreciable-HIGH
        0           Wage Related Costs for Part - A Teaching Physicians-HIGH
```

Next Steps

Further Development

Some recommendations for further improvement:

- Include additional estimators for the final model, with a broader range for the hyperparameter set.
- Converting the target variable to a classification problem, by perhaps binning the continuous values or even signifying a “positive” versus “negative” Net Income.
- Time series analysis will help predict the Net Income specifically for the next year, by accounting for possible trends year-over-year.
-

Questions?