Hospital Provider CMS Cost Reports for FY 2014-2018: A Summary Report

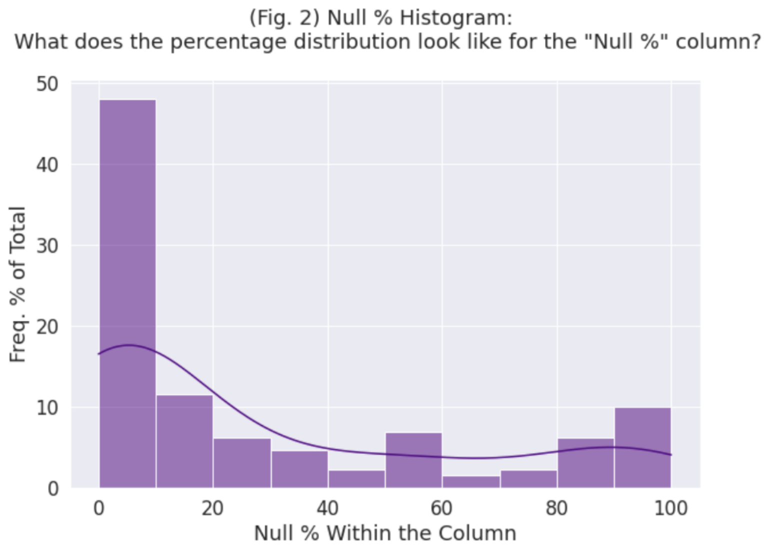
Jon Kim | July 2022

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

Goal. This project will analyze the available data for five of the most recent years - FY 2014 to 2018 - and use a final regression model to predict the Net Income variable at the hospital-level.

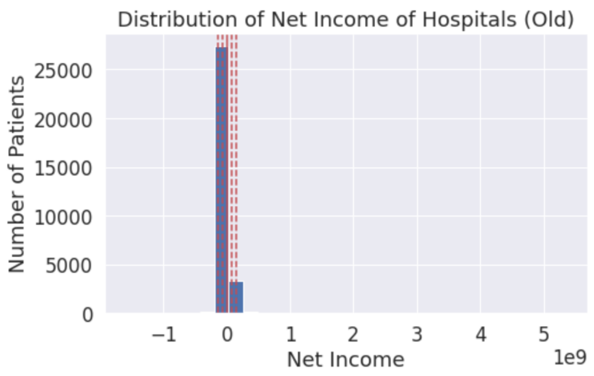
Data Wrangling.

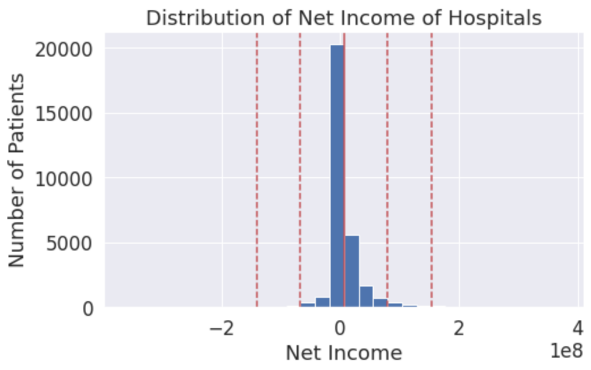
Null Cleaning. At the beginning of the EDA stage, the high number of nulls within the dataset became immediately obvious, though thankfully, the majority of the columns had less than 20% in null values (See below).



My null strategy simply involved a careful consideration of columns within a set of four arbitrary bins of null percentages: >= 80%, 40-80%, 10-40%, and <10%. Each of these bins was subject to a tailored approach of null cleaning, though I naturally omitted more of the columns outright for those with the highest percentage of nulls. After the null cleaning and imputations, the number of columns reduced from 129 to 115 – a smaller than expected change, but a necessary transformation for the next step.

Target Distribution. The distribution of the target feature itself - Net Income - originally had a strong skew to the right due to strong outliers (See left). To alleviate this skew and better force a normal distribution, I filtered out any hospitals with a Net Income z-score of less than 5, which means that only those within 5 standard deviations from the mean were included. The new distribution of the Net Income (See right) became much more visibly “bell-shaped,” with a normal distribution verified using the statsmodel normaltest() function.





Feature Correlation. Several features were perfectly correlated with at least one other feature. Due to the issues that arise from perfect correlation, I removed these variables which reduced the total feature set by only five. After gathering a more complete list of features, I removed a couple whose meaning did not seem relevant to the project, such as CCN Facility Type, as well as some adjustments to some other features with categorical values that were unclear or ambiguous when later dummified.

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

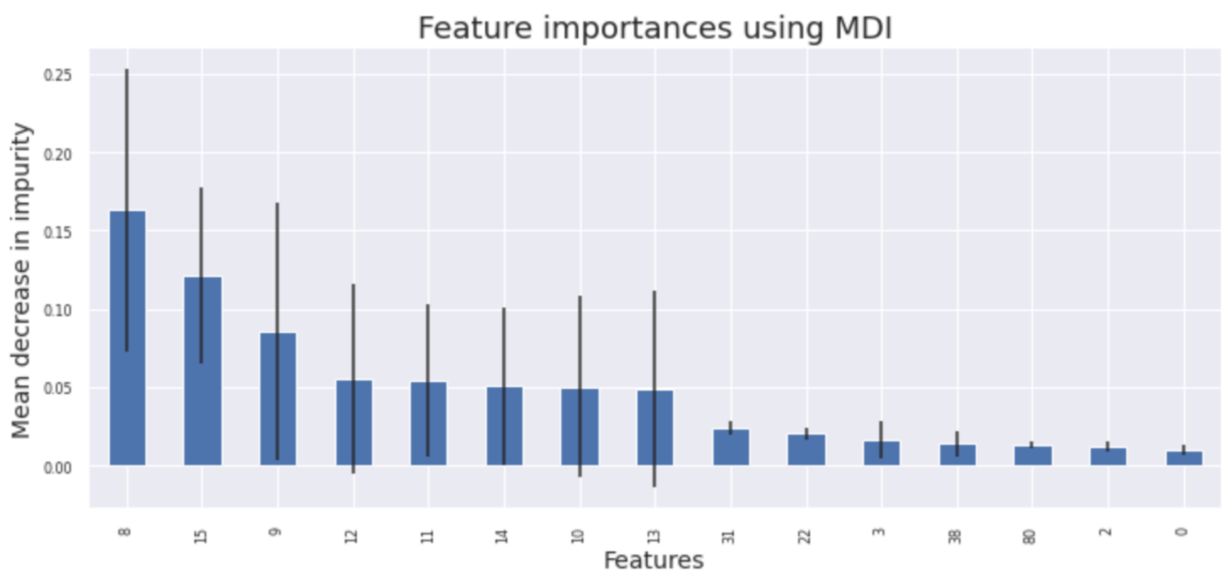
Modeling. I compared several different regression models (See below) using several different metrics, with an emphasis on root mean squared error (RMSE). I created a function that iterates and performs a randomized search of arbitrarily set parameters through each model, with a cross-validation of five folds.

	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.

For most of the metrics, including RMSE, the Extra Trees Regressor performs the best, in that it outputs the lowest value of these loss functions. However, one major concern was the still high value of the RMSE, which is at times almost tenfold the actual values in unit size. To help determine whether the high values for the loss functions are due to modeling errors or some inherent flaw in the dataset, I added a “random” model that only predicted the mean of the target value, and hypothesized that, if the Random model outperformed any of the other models, then the issue likely points to a modeling error. I found that the Random model performed the worst in almost all metrics, with the obvious exception of R-squared. Thus, I decided to move forward with the Extra Trees Regressor, with further recommendations that I include for next steps.

In terms of the feature importances, the top few have clear connections to the target variable of Net Income, as almost all of them relate to either revenue or cost (net income usually reflects revenue less cost in some way). I had originally kept these features since they did not have too high of a correlation with the target variable, unlike others such as Total Net Income which was nearly perfectly correlated



```
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. Some recommendations for further improvement:

- 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.