DATA-DRIVEN DECISION MAKING IN NURSING HOME INVESTMENTS

BANA 620 Data Mining & Predictive Analytics for Business

CSUN Masters of Business Analytics

May 7, 2024

Submitted by: DJK Predictive Partners

Denise Marie Becerra

Josh Dennis

Kanak Sharma

TABLE OF CONTENT

Executive Summary

This executive summary presents DJK Predictive Partners' comprehensive findings, insights, and recommendations stemming from the predictive analysis proposal aiming to analyze the profitability of nursing homes across the United States of America during 2015 to 2021. The objective of this analysis accounts for the identification of potential factors affecting the financial performance of nursing homes to determine the most significant factors impacting *Net Income* through the use of machine learning algorithms. In addition, this analysis will describe trends observed in the data during pre and post COVID-19 pandemic and how the pandemic impacted the profitability of nursing homes across America.

Introduction

"According to Centers for Medicare & Medicaid Services (CMS) data, in July 2022, approximately 1.2 million people resided in more than 15,000 certified nursing homes" (HHS.org). As birth rates in the United States of America decline, as reported by the CDC in for National Center for Health Statistics, and a larger need for sustainable housing increases, via EPA Smart Growth and Affordable Housing, Americans are faced with the economic burden associated with the aging generation of Baby Boomers. Baby Boomers are individuals born between 1946 to 1964 and account for 76 million of the population. For many elderly Americans in need of financial stability and affordable housing, it is imperative to find solutions in the nursing home business sector. This report aims to tackle the proposed research question: What factors impact the *Net Income* of nursing homes pre & post COVID-19 pandemic?

Methodology

Data Collection

To begin this analysis, accessing data that is complete, accurate, and cost efficient can be difficult to find. However, various types of primary data were collected to conduct this analysis from the Centers for Medicare & Medicaid Services including:

Types of Reports	**Summary of Content**
Cost Reports	Detailed reports regarding financial aspects of the business including Facility Names, Addresses, Fiscal Year Dates, Control Types, Gross Revenue, Number of Beds, Overhead Salary Costs, Wages and Salaries, Liabilities, Total Assets, etc. from 2015-2021.
Health Deficiencies	Inspection Reports regarding facilities and reported incidents with patients, complaints, dates of inspection surveys, types of surveys conducted, etc. from 2015-2021
Provider Information	Reference reports regarding provider demographics and characteristics such as Provider Names, Addresses, Ownership Types, Weighted Health Survey Scores, Total Fine Amounts, Overall Facility Ratings, Facility Certification Types, Number of Federally Certified Beds, etc. from 2015-2021
Quality MSR	Reports on various metrics on quality of care by quarter, location of facility, file dates, provider names, etc. from 2015-2021.
Penalties	Reports related to fines, penalty types, location of facility, penalty file dates, etc. from 2015-2021.

From these reports, only relevant data was compiled into a singular final dataset to create a more comprehensive dataset that enhances accuracy and reliability as well as streamlining the data cleaning process as this is the most time consuming portion of predictive modeling. By filtering out data that has a high percentage of missing values, irrelevant data, or unique identifier features, the complexity of the data is minimized to be able to support a better understanding of robust insights to make the best informed business decisions.

Data Analysis Techniques

Descriptive Analytics

To understand the data being used, a statistical summary was produced using the method .describe(). From this, the team was able to calculate various statistics of the numerical columns in the dataset. This technique gave a quick overview of the data in addition to giving insights on the distribution of the data, identifying outliers, and detecting issues that would impact the integrity of the data such as missing or null values. The team was able to identify the mean, standard deviation, minimum and maximum values in the data set, as well as percentiles.

For the categorical data, visualizations were created to gauge an understanding of the data including graphs such as bar charts, pie charts, and histograms. Bar charts represent categorical data with rectangular bars showing the different length of the bars related to the frequency of data in that particular category. Pie charts are circular charts that are divided into triangular slices that show proportions of the categorical data compared to one another and also consider the frequency in each category. Lastly, histograms, which are usually used

for numerical data, can separate the categorical data into bars similar to bar charts, however, shows the distribution of categories relevant to the frequency of that category.

Feature Engineering

Feature engineering allows for the selection of relevant features for modeling. Features can be transformed or aggregated from its raw form to reduce dimensionality, address non-linear relationships among data, create more digestible insights, and improve model performance. For feature engineering, DJK began with manual feature selection, which consists of hand picking variables of interest that could have an impact on the target variable, *Net Income.* Additionally, many variables observed in the dataset were vastly composed of null values, therefore these variables were manually excluded. After selecting variables intuitively, the team processed a correlation matrix analysis to determine if the chosen variables were related to the target variable. It was determined that out of the 28 manually selected variables, 22 of those variables would impact the target variable during the modeling process. The remainder that would help describe trends in data that are of categorical nature.

Machine Learning Algorithms

When answering a predictive analytics question, differentiation between classification and regression problems is a necessity. In this analysis, the proposed problem statement is one of continuous nature. From this, DJK determined that regression models needed to be implemented as the target variable is *Net Income*. There are various regression models that could be implemented to predict a continuous variable. The ones selected for this analysis were Multi-Linear Regression, Lasso Regression, and K-Nearest Neighbor.

Multi-Linear Regression is a statistical modeling technique that identifies relationships between one dependent variable (usually denoted by *y*)and two or more independent variables (often denoted by x_1, x_2, x_3, etc). (Fuentes). The algorithmic equation is as follows:

$$\hat{y} = w_0 + w_1 X_1 + w_2 X_2 + \ldots + w_n X_n$$

where:

– w_0, w_1, . . . , w_n are the coefficients or weights of the model.

– X_1, X_2, . . . , X_n are the input variables.

$$RSS = \sum_{i=1}^{N} (y_i - \hat{y}_i)^2$$

where:

– N is the number of observations in the training set.

– y_i represents the actual values of the target.

– $\hat{y}_i$ are the predicted values.

In a Regularized Lasso Regression, a penalty is added to the absolute value of the magnitude of coefficients and can result in setting coefficients to zero creating a simpler model excluding certain factors entirely. By imposing this penalty, the machine learning algorithm prevents overfitting of the model, which is a technique known as Regularization (Fuentes). The algorithmic equation is a follows:

$$\text{Objective}_{L1} = \text{Loss Function} + \lambda \sum_{i=1}^{n} |w_i|$$

Where:

- Loss Function is the loss you are trying to minimize (e.g., Mean Squared Error for regression problems). For Mean Squared Error, the loss function is $\frac{1}{m}\sum_{j=1}^{m}(\tilde{y}^{(j)} - \hat{y}^{(j)})^2$, where *m* is the number of samples,

y^j is the actual value and $\hat{y}^j$ is the predicted value.

- w_i represents the coefficients or weights of the model.
- n is the number of features.

- λ is the regularization strength (hyperparameter).

The last model, K-Nearest Neighbor, is a non-parametric algorithm that can be used for either classification or regression problems. “The algorithm works on a simple principle: it classifies a data point based on how its neighbors are classified. In kNN, ’neighbors’ are the data points in the training set that are nearest to the point in question” (Fuentes).

$$\text{Euclidean distance} = \sqrt{\sum_{i=1}^{n} \left(x_i - y_i\right)^2}$$

Algorithm Steps

1. Choose the number of k and a distance metric (e.g., Euclidean distance).

2. Find the k nearest neighbors of the sample that we want to classify.

3. Assign the class label by majority vote. For regression tasks, assign the average of the k nearest neighbors.

Software & Programming Languages

Python, which is a high level, interpreted programming language utilized due to its simplicity and readability, accompanied with Visual Studio Code as an integrated development environment (IDE), were utilized to conduct the necessary data preprocessing and modeling required for a data analysis project due to its versatility, open source accessibility and user friendly interface.

Data Description

Data Characteristics

To achieve a general understanding of the dataset being used, DJK Predictive Partners (also referred to as DJK or Predictive Partners) meticulously cleaned the final dataset which included 28 features with 102,420 observations, or instances. The variables were primarily composed of numerical values, however there were some of categorical nature. The numerical variables included unique identifiers, which were later dropped, dates, which were transformed to only account for the year, amounts spent on salaries, costs, various sources of revenue, number of beds at each facility, overall ratings of facilities, and health survey scores. Categorical variables consisted of facility names, addresses and medicare vs medicaid institution status. Numerical variables were classified as integers or floats (decimal values), while categorical data were classified as objects.

Data Cleaning

The first significant hurdle in data preprocessing was combining multiple years of data from separate files into one master dataset. This proved difficult as formatting was not consistent across all periods. As the majority of the analysis would be based upon the Cost Reports, this data preparation took precedent. With the use of a data dictionary, columns with different names but representing the same data were combined by hand. Once combined, descriptive analysis was conducted to discover more about each variable. Ultimately, 39 variables in the Cost Reports were deemed to have bad data, as they had a significant proportion of null values, many of which were greater than 99 percent blank. These were then removed from the dataframe. Prior to the next step, an index was added to the data frame and utilized the fiscal year end data column to extract the reporting year. In an effort to reduce the dimensionality of the dataset, the data was then split across four categories; property data, income, rent roll, and balance sheet items. Once these new data

frames were cleaned and prepared independently, they were remerged. Additionally, it was deemed that the Provider Info reports had information that may be valuable to the analysis. These reports were merged with the Cost Report dataset, using the year and provider CCN as the key.

Outliers were identified through the use of boxplots and domain knowledge. Extreme outliers were handled by performing winsorization, which is a preprocessing technique in predictive analytics. It provides a max and min cap on extreme outliers based on a specific percentile threshold, which in turn reduces the influence of the outliers in the model and analysis. The dataset presented many negative values in *Net Income* due to the nature of the target variable. Profitability within any business can reflect positive or negative values in most business scenarios. However, from a data modeling perspective, a threshold for handling negative values must be identified and implemented to avoid damaging the integrity of the model accuracy. From this, DJK settled on a threshold of 30 percent inclusion rate when it came to negative values. Similarly, the exclusion or imputation of null or NA values in continuous variables posed a threat to the performance of the model. Therefore, the team decided to impute the null values with the mean of each particular variable to avoid deletion during instances where the dataset would lose a significant amount of data related to the target variable. It is important to note that some of the variable data types needed to be changed to enhance interpretability. These were changed from objects to integers so that the models would be able to capture the numerical data correctly.

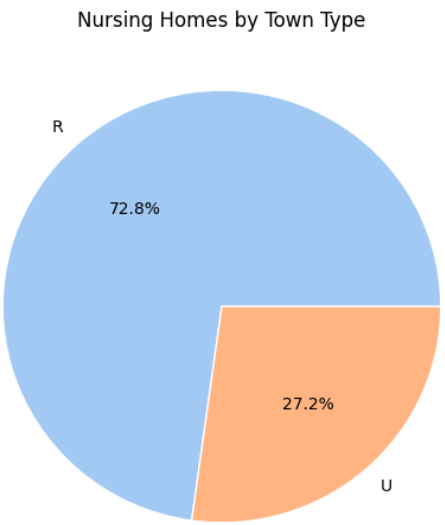
Exploratory Data Analysis

For the numerical values, a summary statistics was performed to provide insights on the range and dispersion of values. Provided below (Figure 1) is an example of a summary statistics using the .describe() method in python on 5 numerical features of interest:

	count	mean	std	min
Gross_Revenue	102214.0	1.181310e+07	1.333882e+07	-16696967.0
Overhead_Non_Salary_Costs	102420.0	5.785627e+06	5.315069e+06	1.0
Total_Costs	102356.0	1.308935e+06	1.126040e+06	48.0
Wage_related_Costs_core	101812.0	8.287623e+05	1.113373e+06	-134421.0
Net_Income	102140.0	6.935344e+04	2.454103e+06	-167469610.0

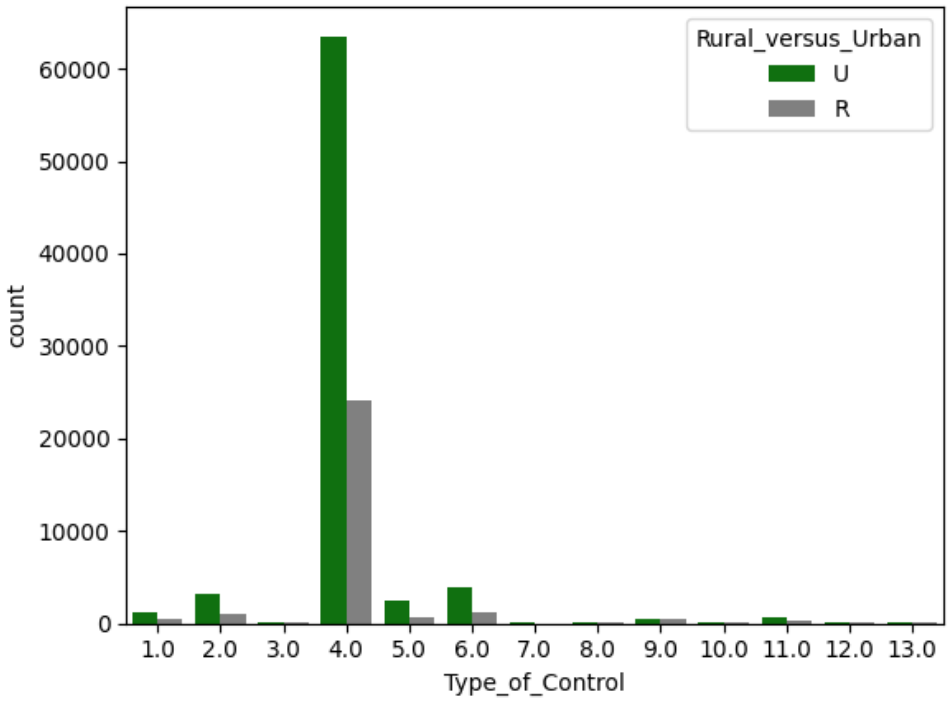
	25%	50%	75%	max
Gross_Revenue	5629118.75	9356288.5	14797444.50	1.300407e+09
Overhead_Non_Salary_Costs	2838789.25	4622817.5	7074050.25	2.177784e+08
Total_Costs	589853.25	1052547.0	1702053.25	3.551056e+07
Wage_related_Costs_core	332752.75	570726.0	957057.25	4.054870e+07
Net_Income	-388385.00	64840.0	555882.00	2.096005e+08

To further explore the trends in the data, visualizations were created. This pie chart shows the frequency of nursing facilities in the United States on the basis of rural locations versus urban locations. From the graph below, Predictive Partners concluded that about 73 percent of the nursing home facilities in the United States were located in rural areas, while only 27 percent were located in urban areas (Figure 2).



Rural areas are defined as areas of a country with low population density. Examples of these areas in the United State of America include the Appalachian Region (Kentucky, West Virginia, Tennessee, and North Carolina), the Great Plains (Kansas, Nebraska, parts of Texas and Oklahoma), and the Mississippi Delta (Mississippi, Arkansas, and Louisiana) (U.S. Census Bureau). Urban areas have higher population density rates when compared to rural areas, according to the Census Bureau, and the most recognizable urban areas are Los Angeles, California, San Francisco, California, New York City, New York, Washington D.C, Chicago, Illinois, and Houston, Texas.

Additionally, the bar chart below shows the distribution of facility control types across rural and urban areas on the basis of frequency where each value, 1 - 13, represents

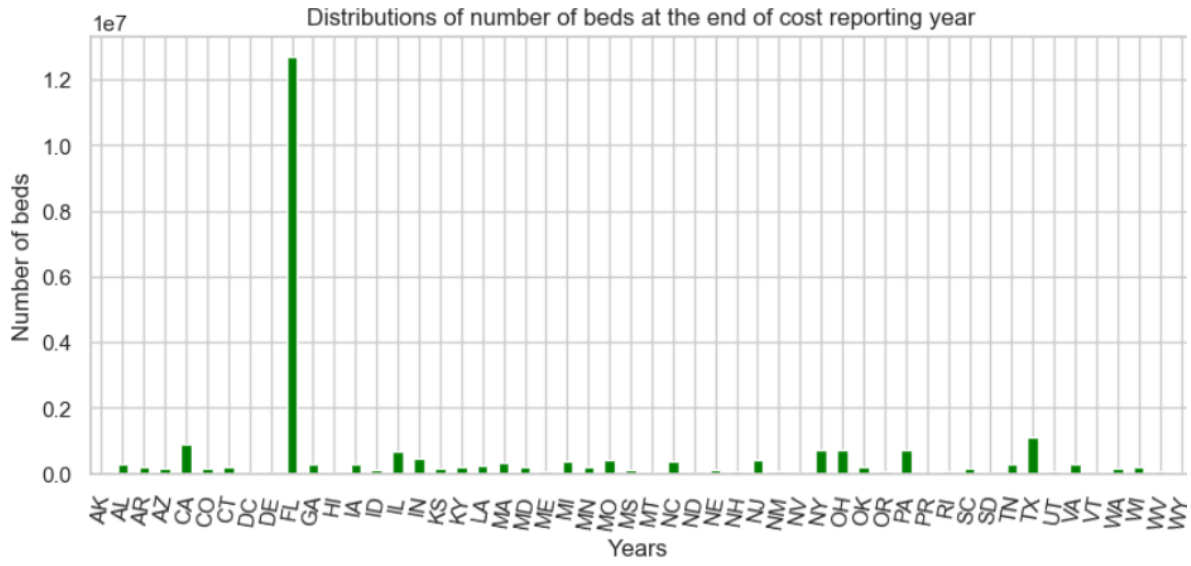


(Figure 3)

different types of ownership. The most significant were types 4, 6, 2, 5, and 1 which correspond to Proprietary Corporations, Proprietary-Other, Voluntary Nonprofit-Other,

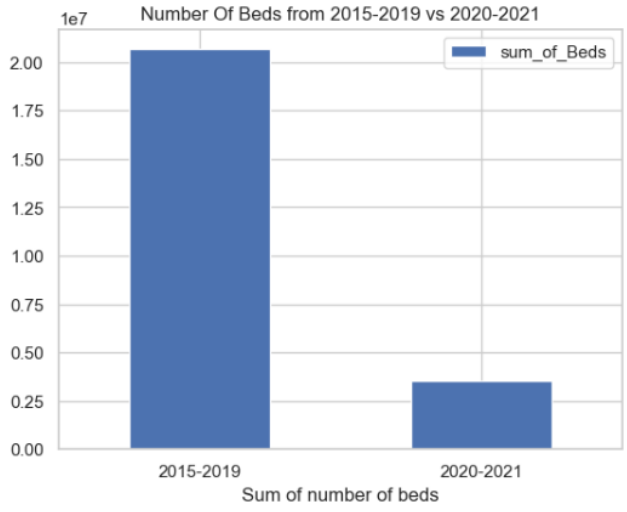
Proprietary-Partnership, and Voluntary Nonprofit-Church. From this graph it was determined that both urban and rural areas have the most significant frequency of corporation owned nursing facilities in the United States.

(Figure 4)

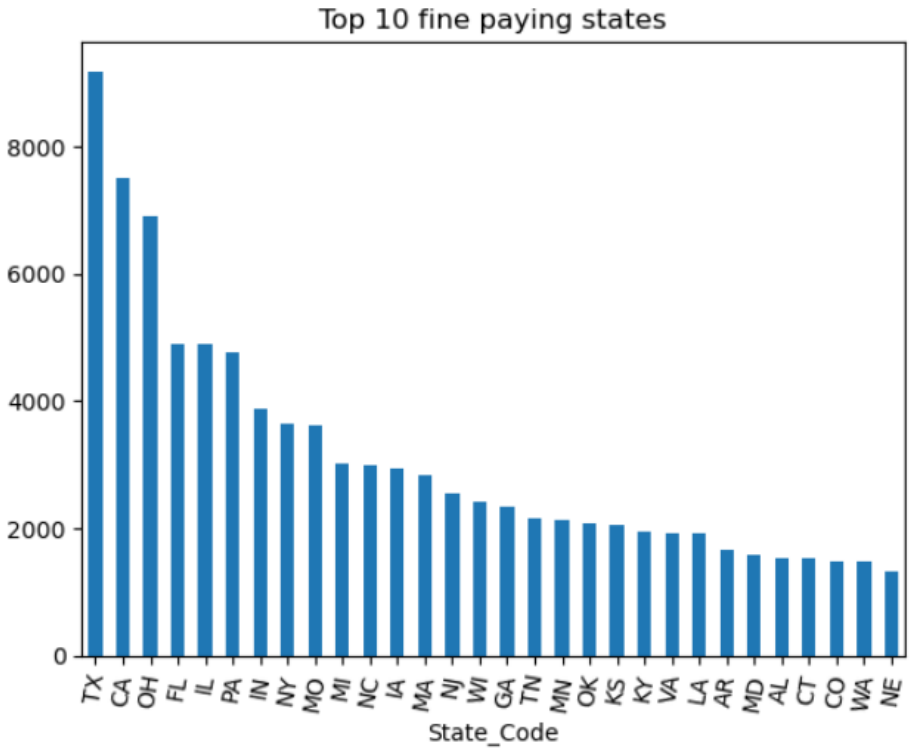


In regards to distribution among the number of beds within nursing home facilities by states, Florida outperforms all other states. This could be due to the assumption that Florida is the most desired location for retirements in the United States. Texas, California, New York, and Ohio are the next significant states, however, their bed count does not compare to that of Florida's.

The following graph represents the summation of beds pre and post covid 19 pandemic. As the COVID-19 pandemic significantly affected the elderly community, it is easily interpreted that the number of beds available in nursing care facilities decreased immensely as recovery from the infection took much longer in those over the age of 65 (CDC).

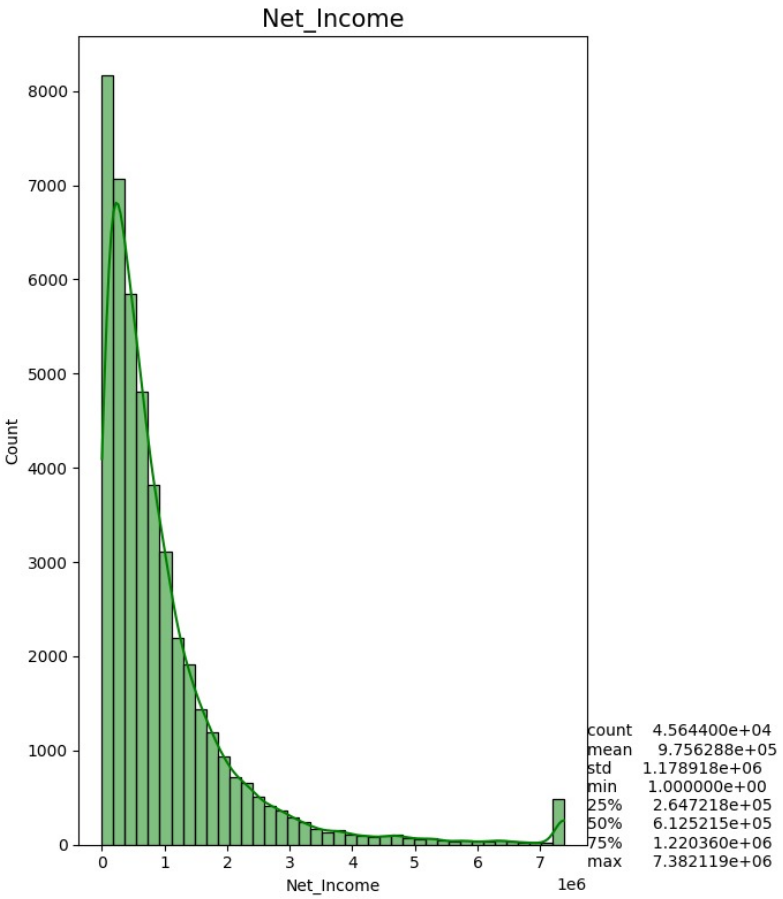


(Figure 5)



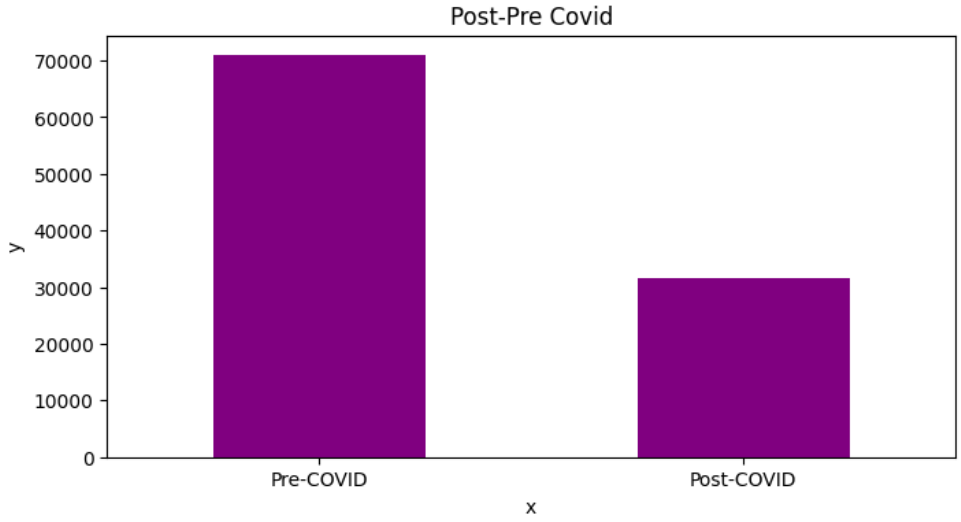
(Figure 6)

Despite Florida being the state with the highest number of beds in the country, it is notable that Texas is the highest paying state when it comes to paying out penalty fines in nursing home facilities in the visualization above.



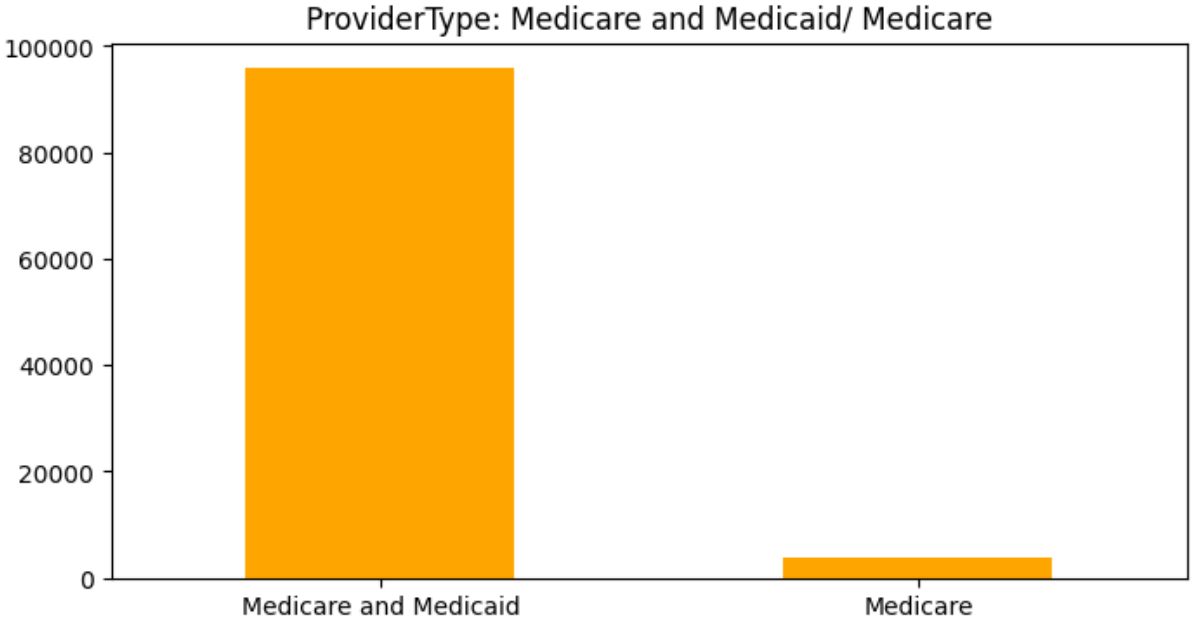
(Figure 7)

Although the skewness of the target variable is extremely high towards right thus it can be established that the target variable is right- skewed. This means the data is not normally distributed which leads to interpretability issues, so transformation is needed for this data to increase model performance, validate the model assumptions, and facilitate statistical inferences. In Figure 8 below, it is evident that the COVID-19 pandemic had a substantial impact on the nursing home facilities.



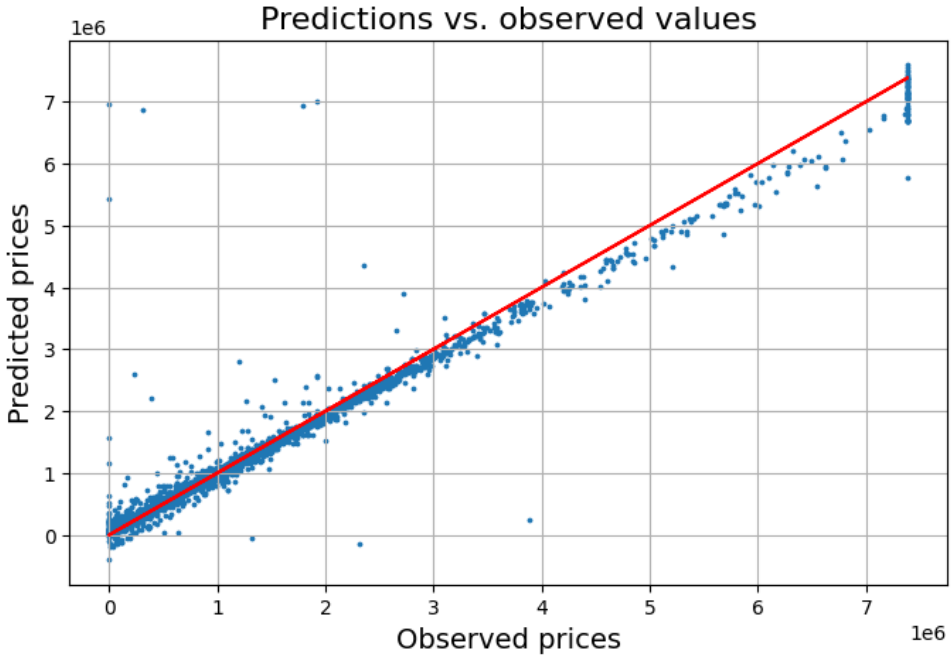
(Figure 8)

When examining facilities that accept certain providers, it is clear that the target variable, *Net Income,* is related to the acceptance of medicare and medicaid at nursing home facilities. Acceptance of these types of providers could increase the profitability of nursing home facilities. (Figure 9)



Analysis & Findings

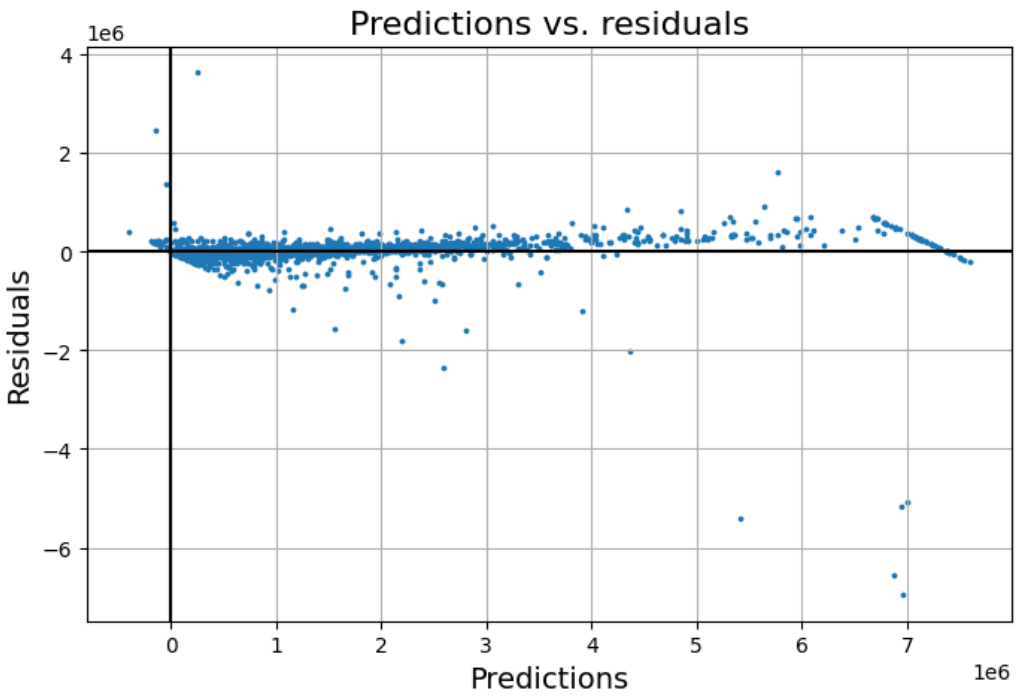
Training Results



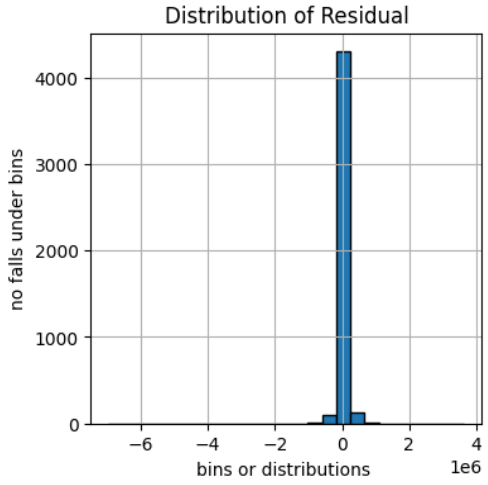
(Figure 10)

In the training set, the model performs extremely well. As can be seen in the prediction versus observed values table above, it is clear that most predicted values are closely plotted around the diagonal. This indicates that the model is predicting values closely to the actual value. However, there are multiple outliers that must be recognized. These deviations from the trend line are likely a result of outliers in the dataset itself, as model complexity and fit are not a concern.

It should be noted that the predicted values slowly trend lower as the observed values increase. While this is not ideal, the linear relationships between the observed and predicted values lean lowerer but have linearity.



(Figure 11)




```
(eval_df["residuals"] <=0).mean()
```

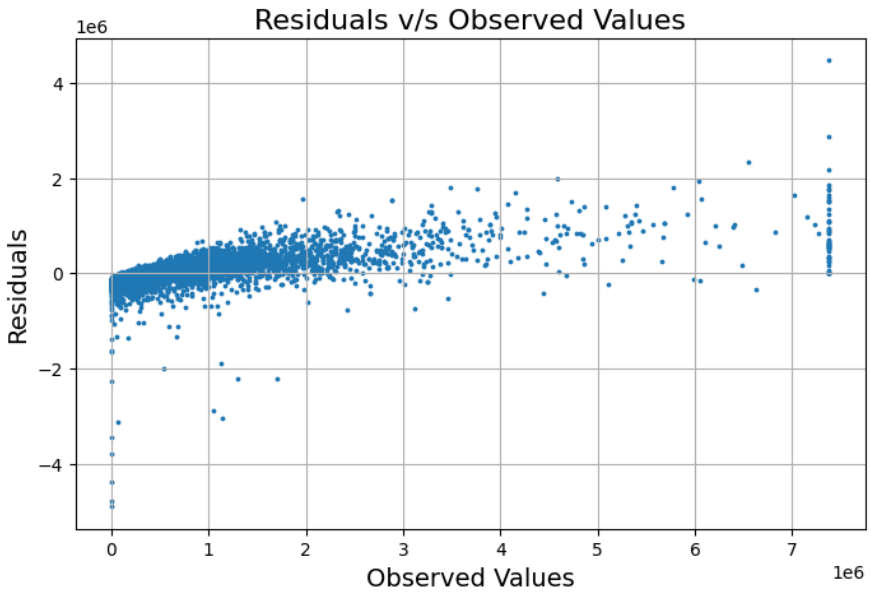
```
0.5001095290251917
```

(Figure 12 & 13)

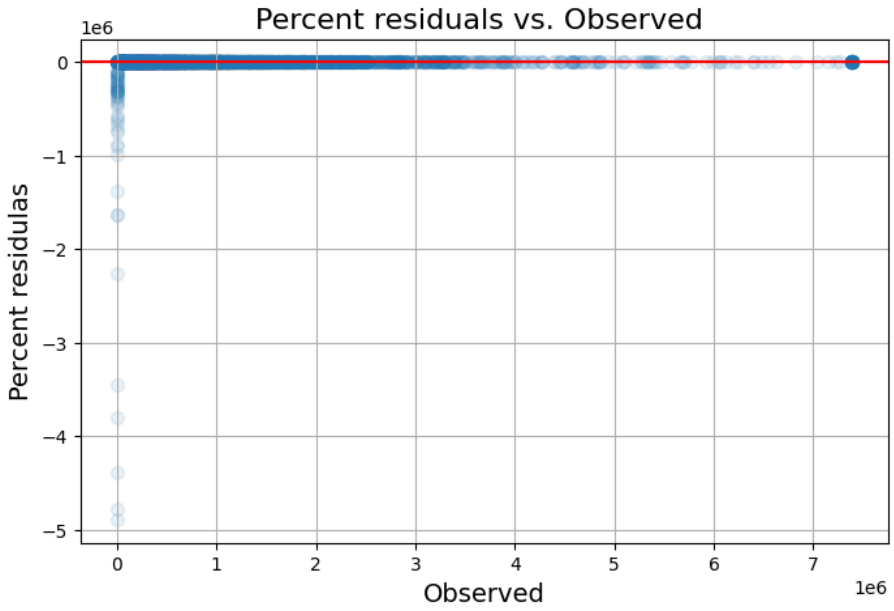
When analyzing the residuals of the predicted values, it is positive to see that there is not a significant pattern or trend of the residuals. This indicates our model is performing nominally well at predicting our target value. There are no immediately noticeable patterns or funnel-shape, which would indicate the model is experiencing heteroscedasticity. As this is not the case, we can confidently say that the size of residual is not dependent upon the

predictor variable. It is worth noting the residuals are split almost perfectly around 0, as can be seen above. This is yet another indicator that our model is accurately predicting our target.

Testing Results

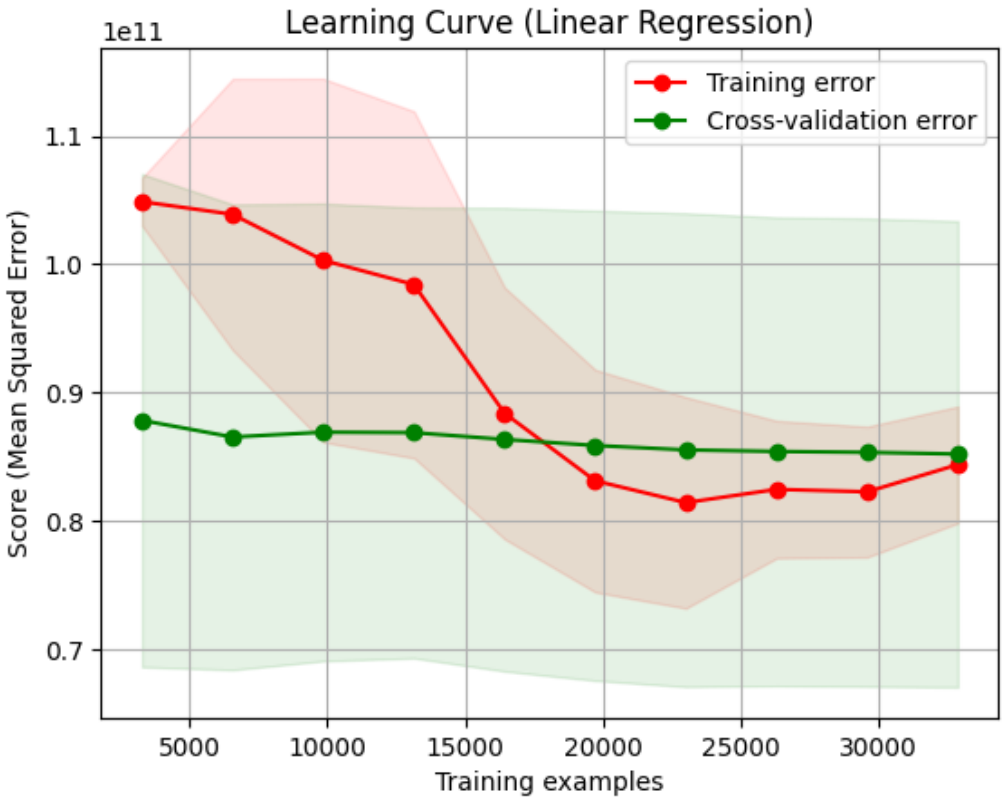


(Figure 14)



(Figure 15)

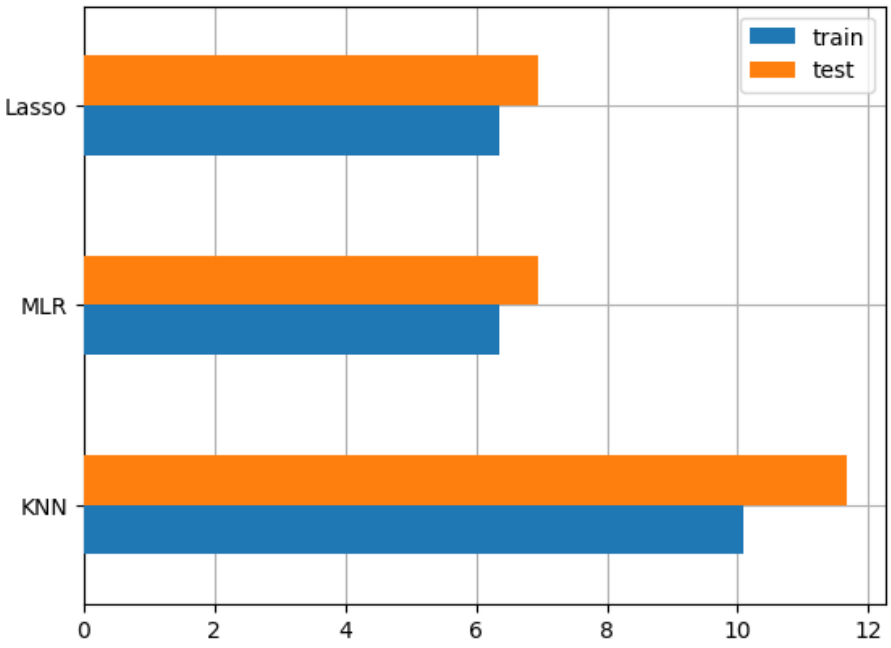
The plot above is showing the residuals of the predictions made from the test data against the actual observed values. The distribution of residuals is encouraging, as it is evenly distributed around the 0 line, and there is no notable pattern or trend. However, as with the previous charts, there are a sizable number of outliers which should be addressed.



(Figure 16)

We also used a learning curve plot to analyze the performance of our model. As can be seen above, the mean squared error using training data begins high, above 1.0 prior to 10,000 iterations. As the iterations increase, the error decreases, showing that the model is refining its predictive accuracy. However, it is concerning that the error begins to elevate as the iterations surpass 30,000, indicating the model is moving toward overfitting.

Regarding the cross-validation error, it is trending lower as iterations increase, however not at the rate that would be preferable. But it is positive to the accuracy of the model when facing unseen data that the error decreases as the examples increase.



(Figure 17 & 18)

	y_true	pred_MLR	pred_Lasso	pred_KNN
43579	1480144.0	1281703.6	1283126.7	941254.4
19651	1273797.0	1228844.7	1228431.7	950020.2
40975	458415.0	474009.1	473878.8	479128.9
4141	12500.0	50014.0	50689.8	300823.2
29308	476744.0	510244.4	510358.2	451794.1
7948	238863.0	259539.4	259853.0	371102.9
20496	47688.0	98695.4	98801.5	463490.5
44537	813093.0	809637.1	808831.0	584577.3
22747	418403.0	441085.8	441369.9	368981.5
13299	424730.0	408209.0	408226.2	375320.1

Ultimately, DKJ utilized 3 predictive model types to fit the data (Multiple Linear Regression, Lasso Regression, K-Nearest Neighbors). It was determined that KNN had the highest degree of performance, as noted in the graph above. KNN will prove extremely useful

to the investor, but will prove most valuable once a potential property is selected. That property can then be compared against similar properties as determined by the KNN analysis and insight can be extracted with a closer look at those similar businesses.

Discussion

Key Findings

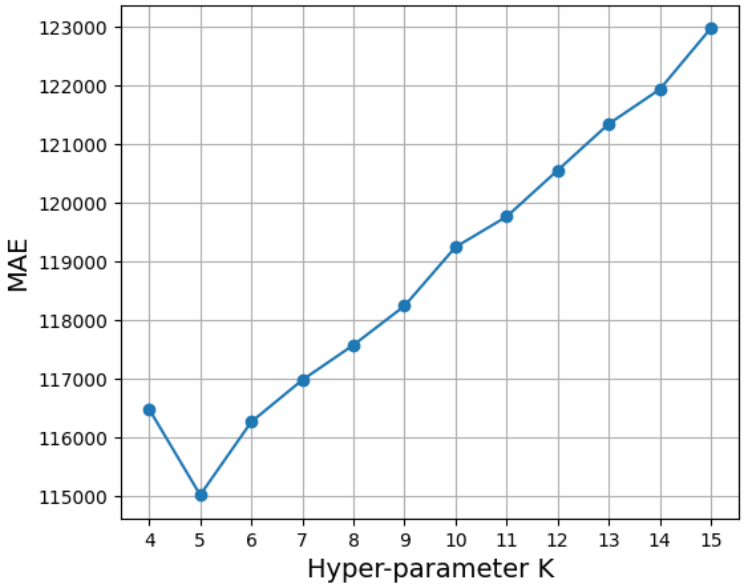
DKJ feels that there are potential limitations of this model which should be addressed. As this model is intended to predict a continuous variable, the team feels that there is an acceptable amount of error in its predictive ability. However, as the source data, specifically the dependent variable, has significant variance, it is expected that the model will also require acceptance of significant error. In future research, DKJ believes it would be beneficial to narrow the scope of the sample, including only data which more closely relates to the property type of interest for the prospective buyer. While this will also decrease the size of the training data, it may prove a worthwhile trade off to reduce the variability of the predicted variable.

MODEL TUNING AND IMPROVING PERFORMANCE

As it has been determined, K-Nearest Neighbors (KNN) is an example of a non-parametric model, meaning it has no parameters to be learned from data. Predictive Partners decided on a k = 15 value in the model earlier. The hyperparameter is optimization —finding the set of best hyperparameters, or, more often, a set of good hyperparameters so our model performs better.

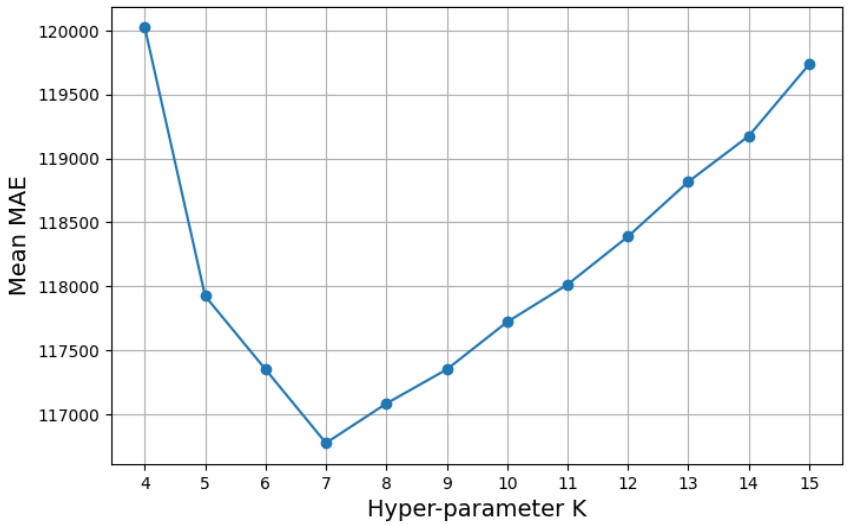
So, the use of MAE (Mean Absolute Error) to measure the performance of the model was utilized after performing a single hyperparameter. The below graph shows where the

MAE is really less at k =5. But as it can be seen that the fluctuation in the graph may be due to randomness associated with creating the training and validation set.



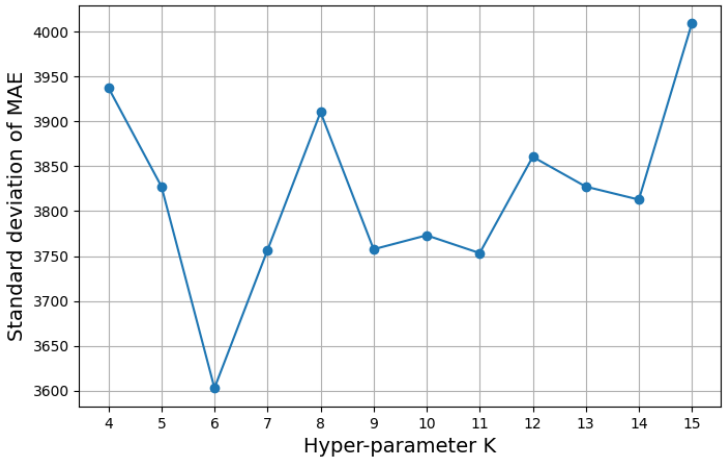
(Figure 19)

Additionally, the team has a need to address the cross validation values. After doing so in the below graph , it is depicted that k=7 is the best value for the model.



(Figure 20)

Another important statistic that was collected from the cross-validation procedure is the standard deviation of each of the 10 MAE estimations (10 estimations for each candidate value). The graph is provided below:



(Figure 21)

The values that can be seen in this plot are estimations of what is known as the variance of the model – the sensitivity of the model to small fluctuations in the dataset. Intuitively, it measures how big the change is in the results of the model when the training samples change; the high variance is generally not desirable, as it would mean the model is not stable. The model shows a relatively low variance and that the model with the lowest variance is the one with K=6. From this, it can be concluded that the optimal parameter for KNN models is K=6.

Recommendations

In analyzing the coefficients of the MLR and Lasso regressions, there were a few specific factors that seemed to have the most significant impact on the predicted value. As is expected, various subsets on income showed to have a significant impact, as well as the overall expenses. However, it is worth noting that payroll related expenses seemed to have

the most significant influence on the *Net Income*, compared to the other expense categories. This information could prove useful, as investors should be looking for investment opportunities in areas where average salary for healthcare professionals remains low.

Conclusion

In general, senior housing facilities prove to be a fruitful investment, but not without its dangers. While some properties observed in this analysis are largely successful, the impact of external forces, such as a pandemic, can have disastrous impacts on the financial standing of a property. Any potential investor should be aware of the impact on payroll and wages on the economic performance of a property. Additionally, once a potential property is selected, information gleaned from the KNN analysis can prove very useful in the decision of whether or not to invest.

Appendices

Figure 1

	count	mean	std	min
Gross_Revenue	102214.0	1.181310e+07	1.333882e+07	-16696967.0
Overhead_Non_Salary_Costs	102420.0	5.785627e+06	5.315069e+06	1.0
Total_Costs	102356.0	1.308935e+06	1.126040e+06	48.0
Wage_related_Costs_core	101812.0	8.287623e+05	1.113373e+06	-134421.0
Net_Income	102140.0	6.935344e+04	2.454103e+06	-167469610.0

	25%	50%	75%	max
Gross_Revenue	5629118.75	9356288.5	14797444.50	1.300407e+09
Overhead_Non_Salary_Costs	2838789.25	4622817.5	7074050.25	2.177784e+08
Total_Costs	589853.25	1052547.0	1702053.25	3.551056e+07
Wage_related_Costs_core	332752.75	570726.0	957057.25	4.054870e+07
Net_Income	-388385.00	64840.0	555882.00	2.096005e+08

Figure 2

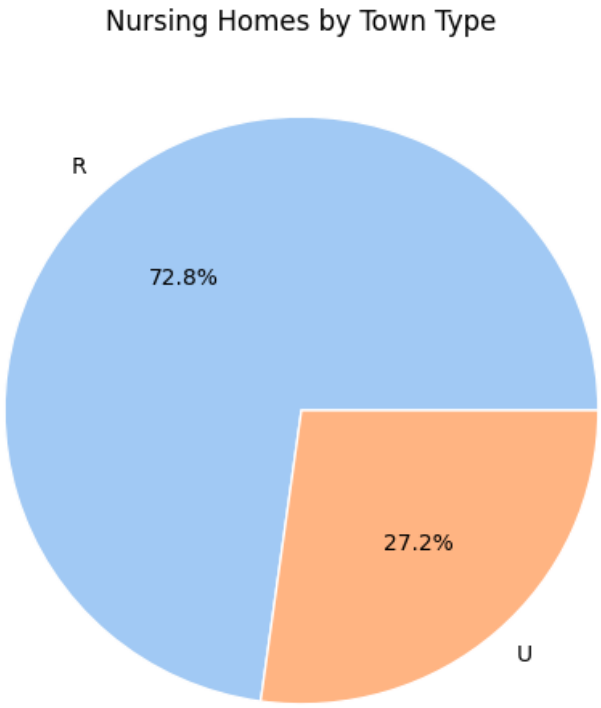


Figure 3

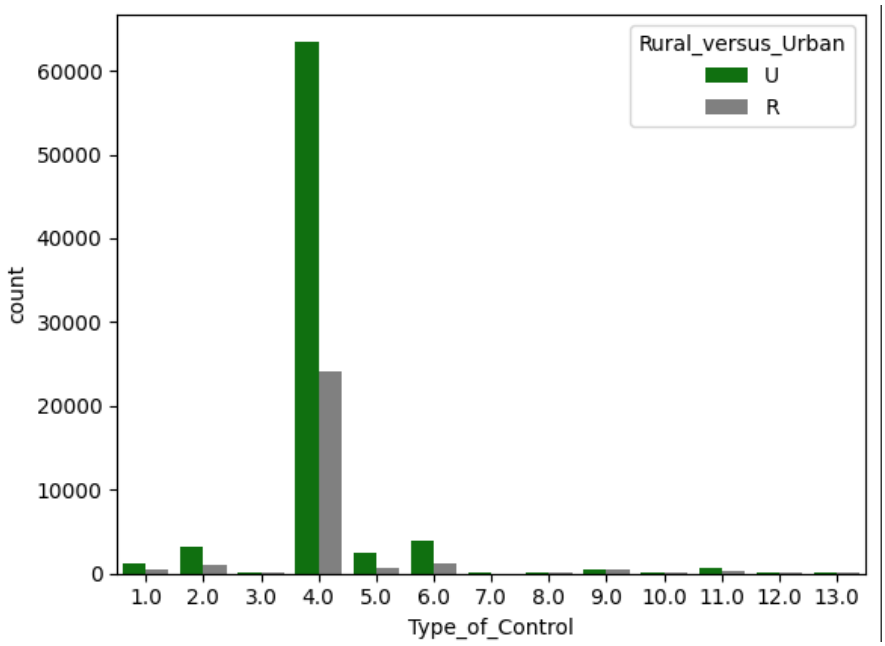


Figure 4

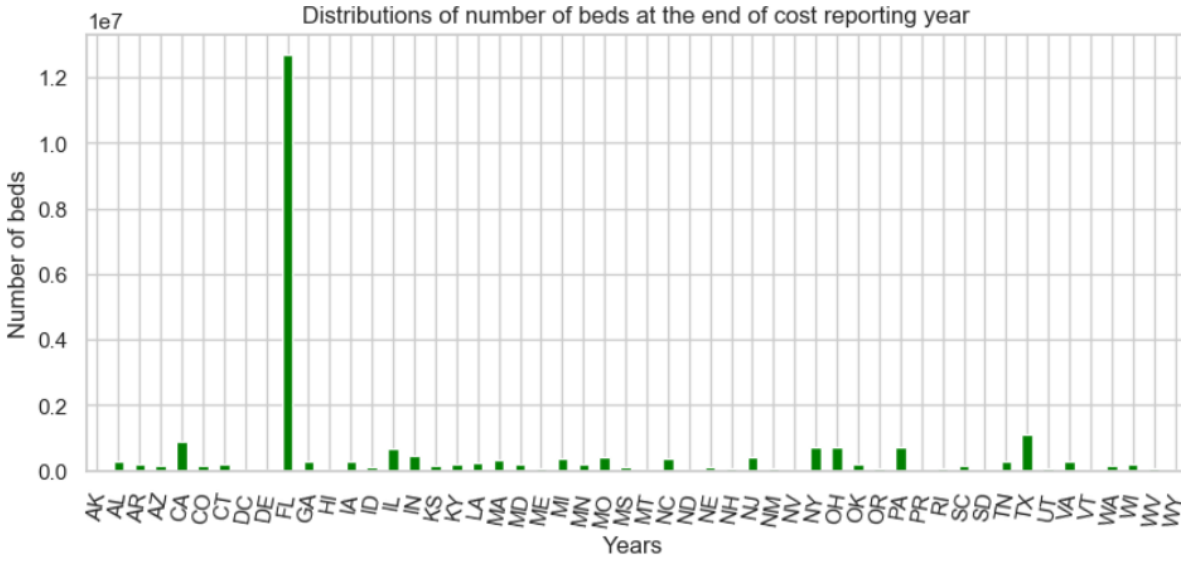


Figure 5

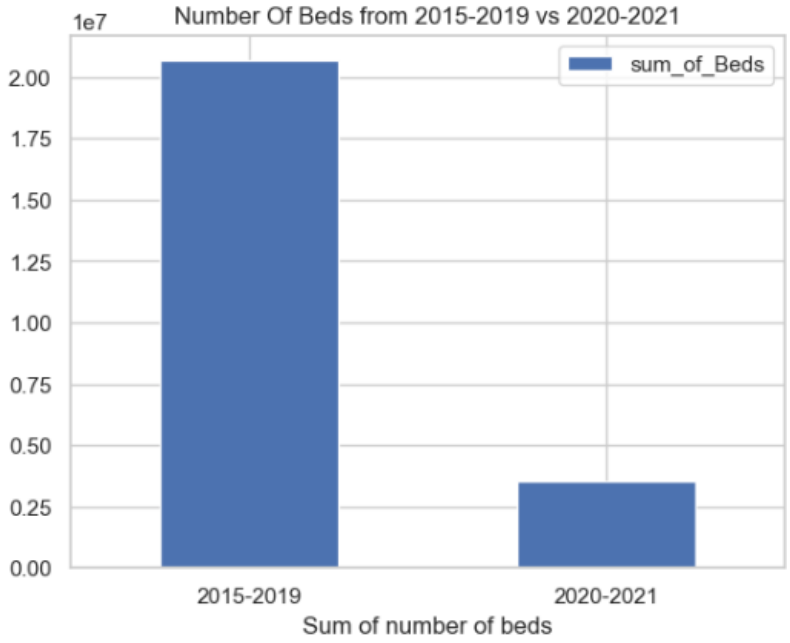


Figure 6

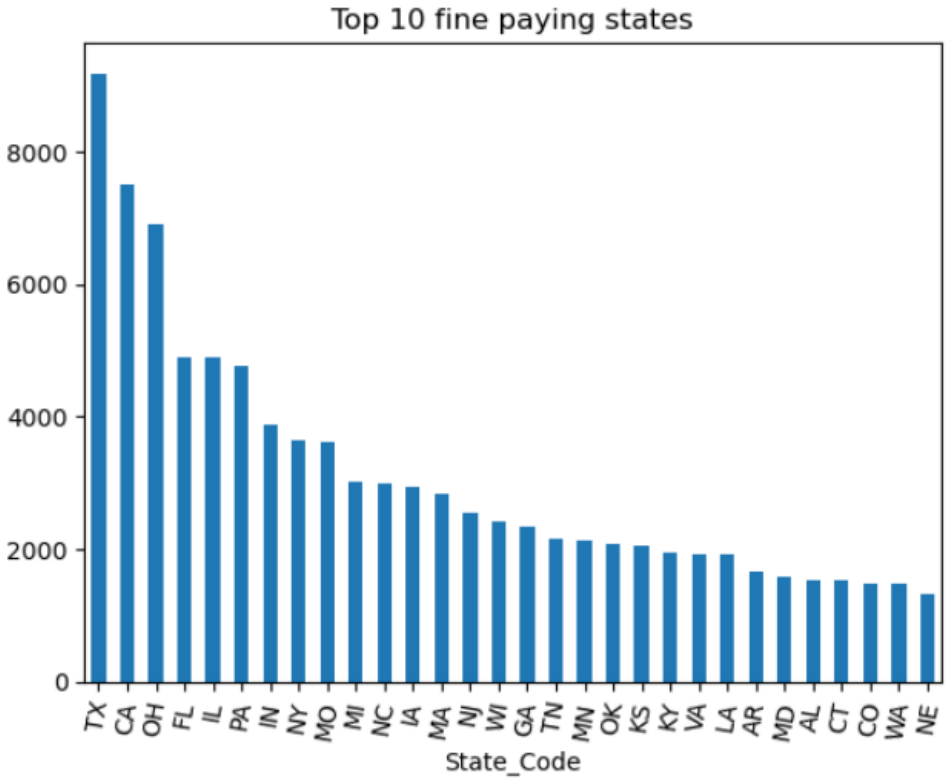


Figure 7

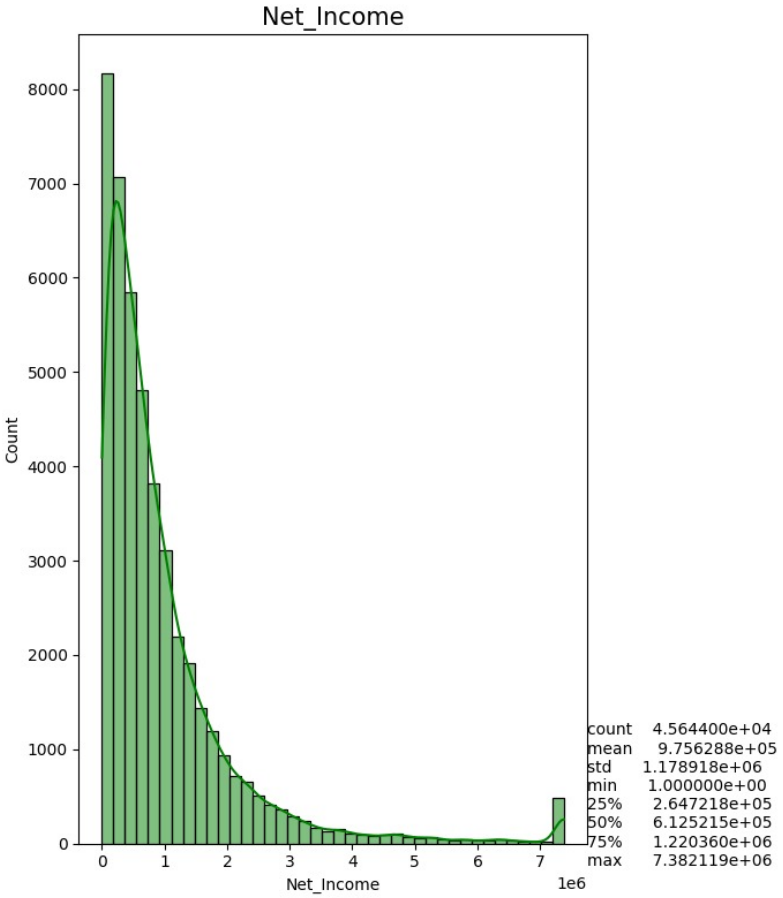


Figure 8

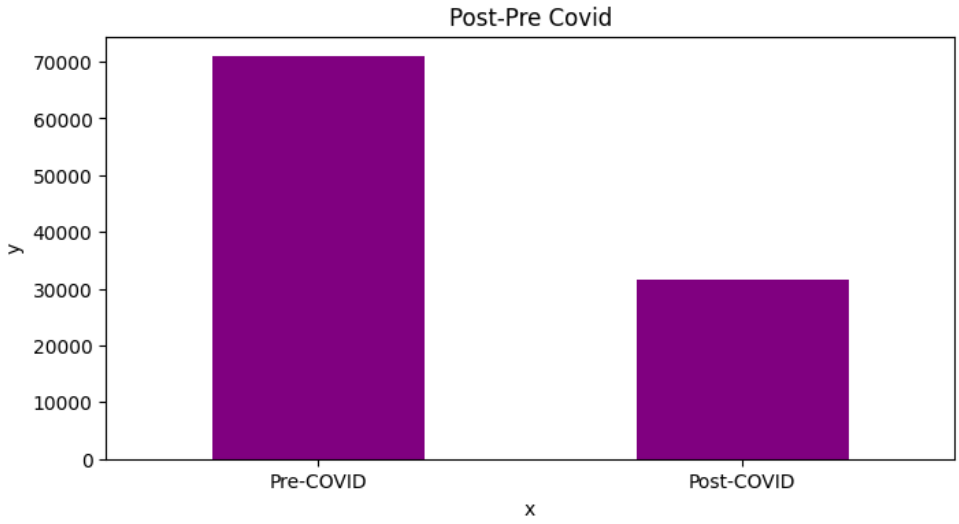


Figure 9

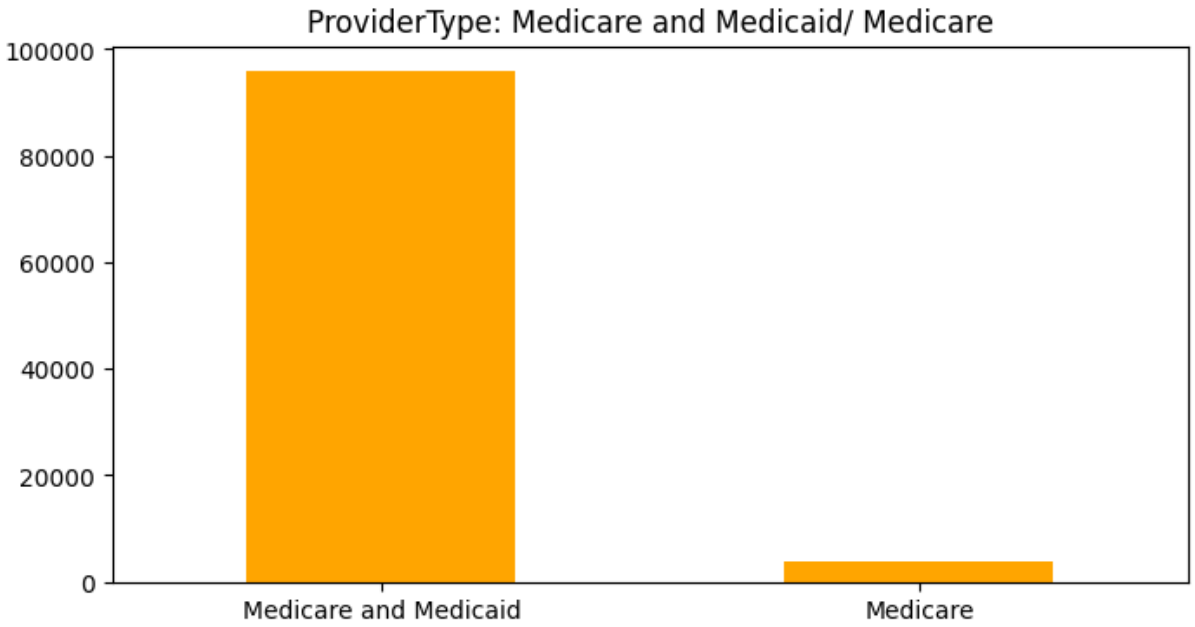


Figure 10

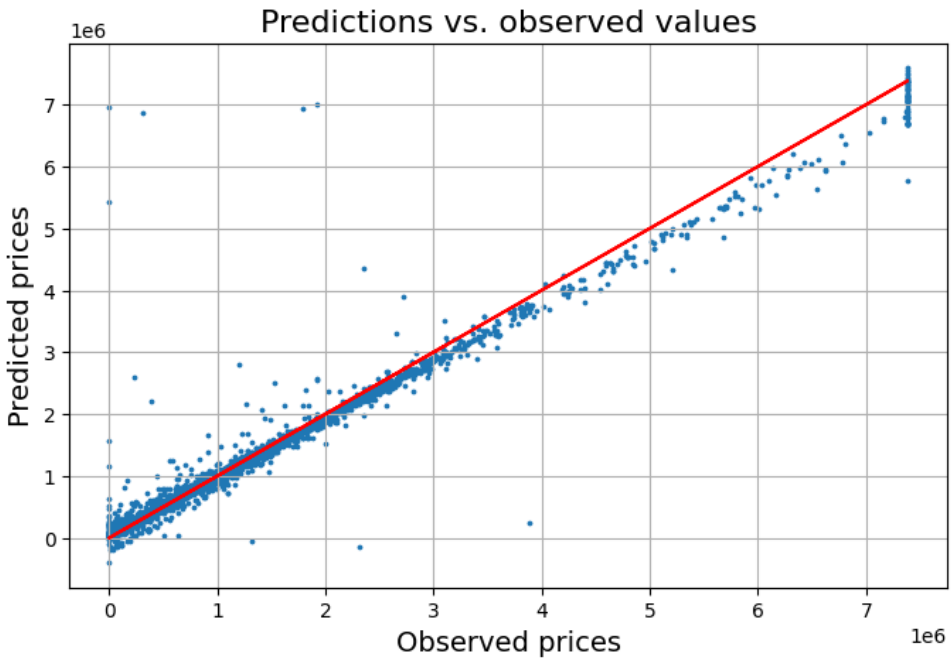


Figure 11

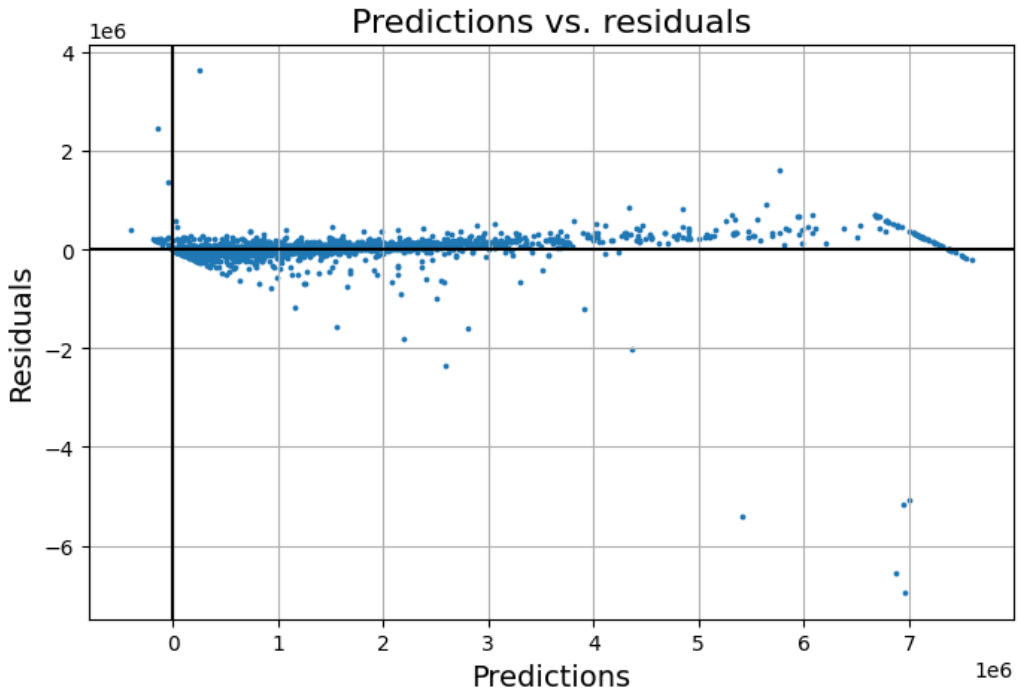
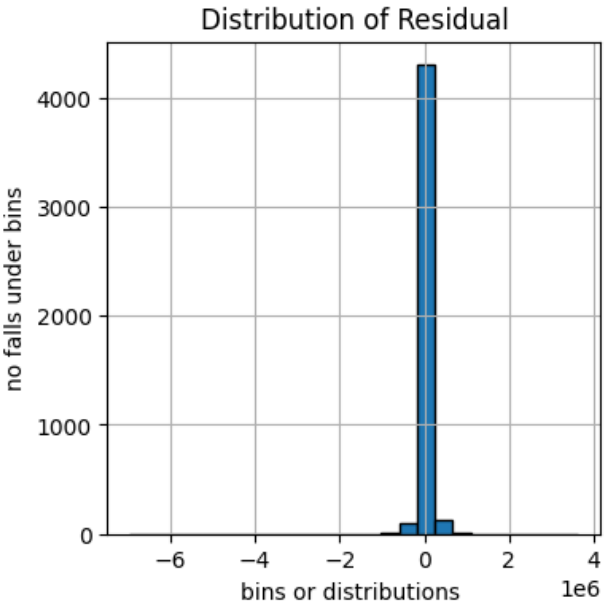


Figure 12

Figure 13




```
(eval_df["residuals"] <=0).mean()
```

```
0.5001095290251917
```

Figure 14

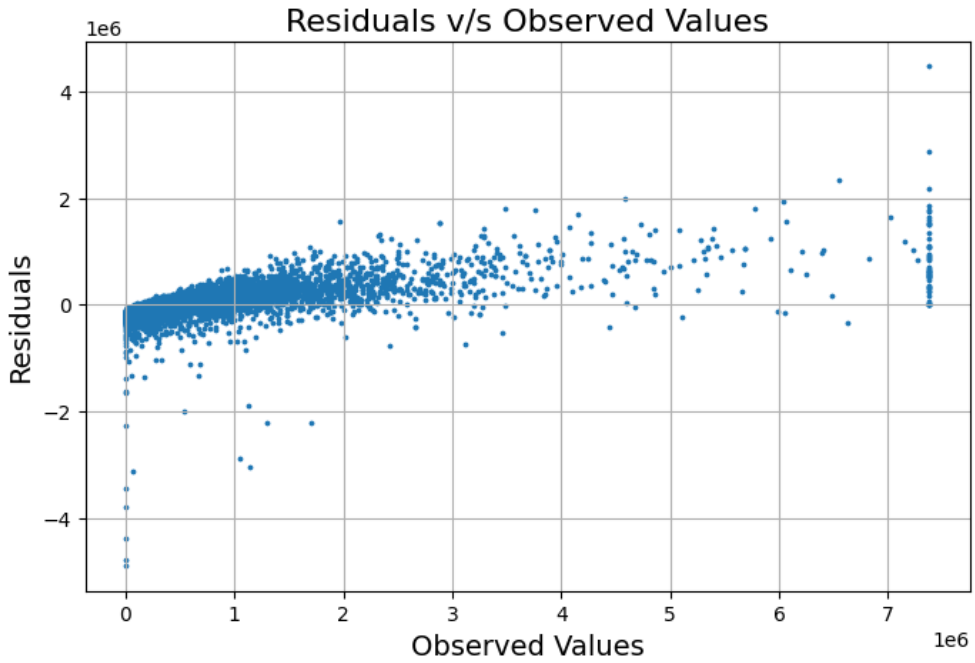

Residuals v/s Observed Values
1e6
Residuals
Observed Values
1e6

Figure 15

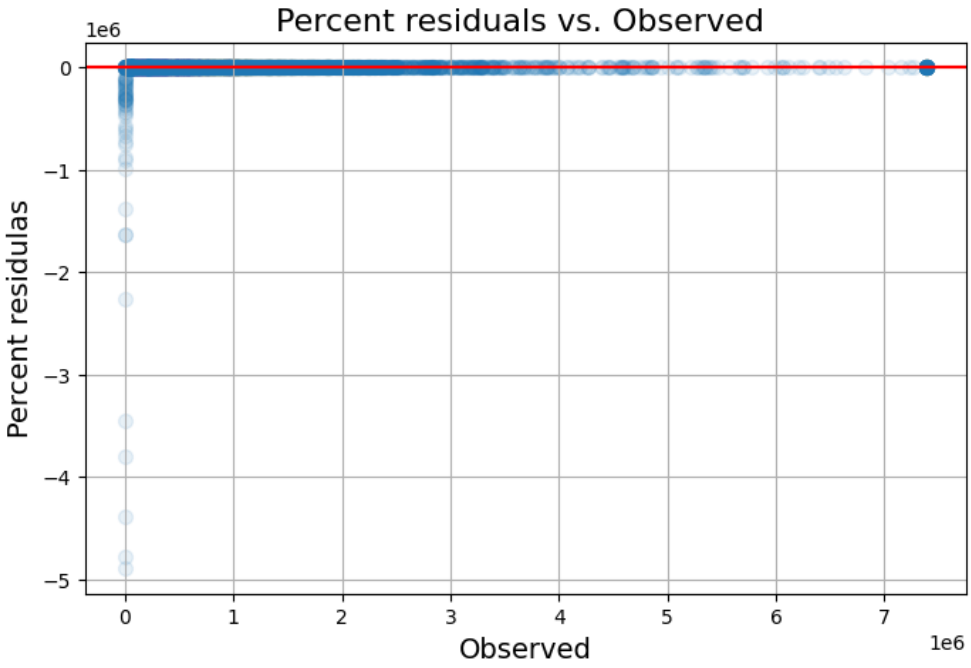

Percent residuals vs. Observed
1e6
Percent residulas
Observed
1e6

Figure 16

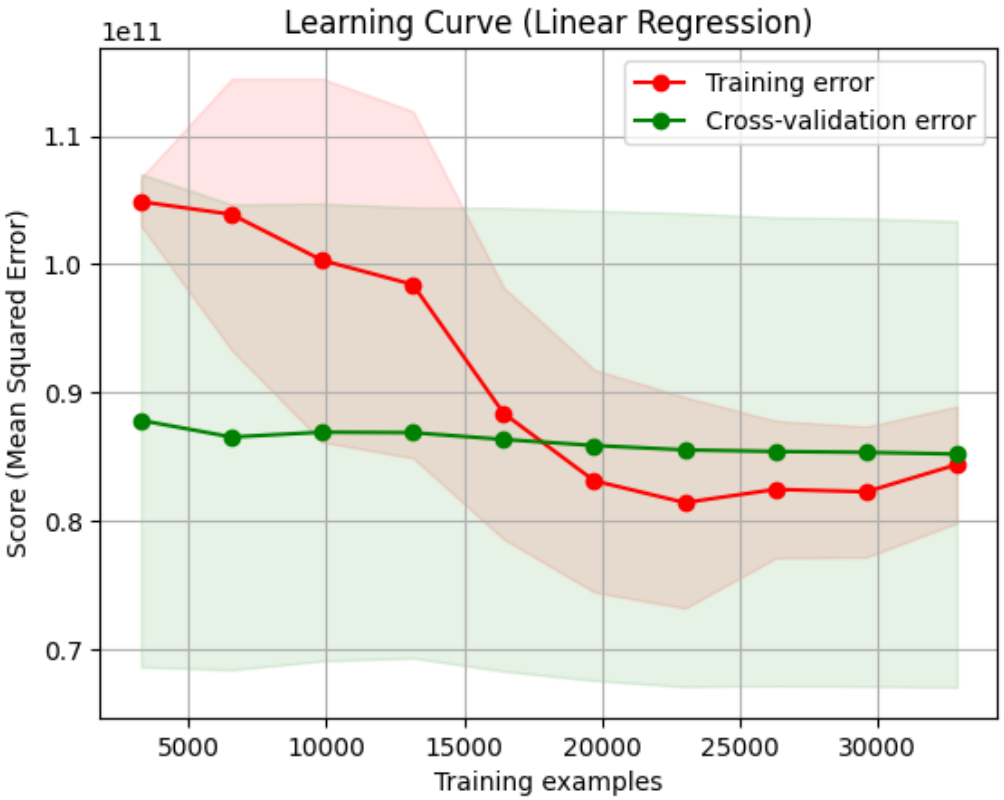


Figure 17

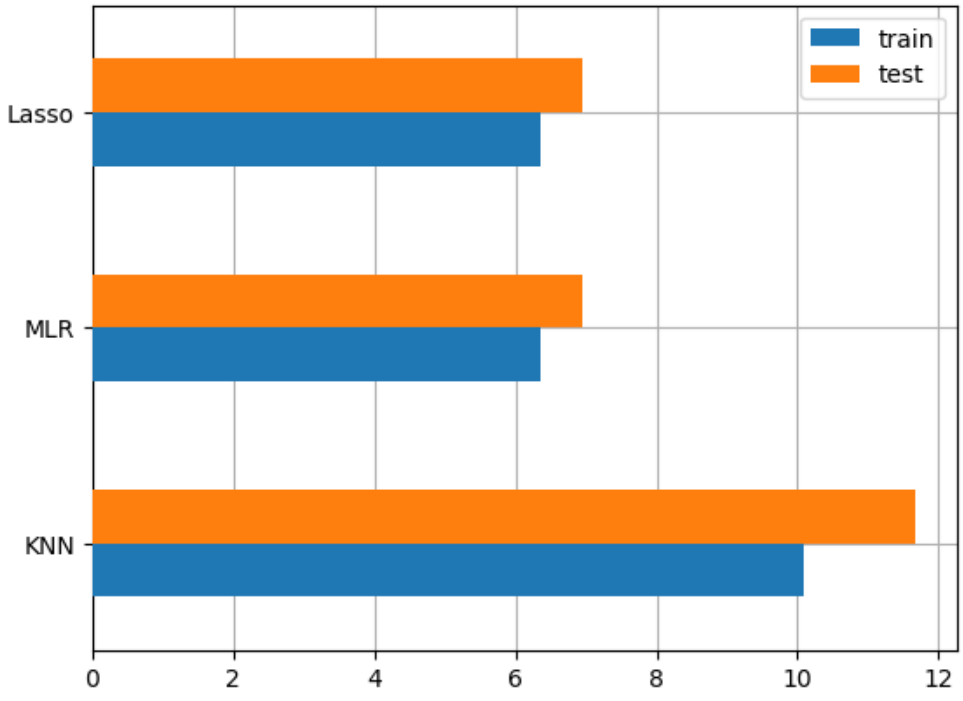


Figure 18

	y_true	pred_MLR	pred_Lasso	pred_KNN
43579	1480144.0	1281703.6	1283126.7	941254.4
19651	1273797.0	1228844.7	1228431.7	950020.2
40975	458415.0	474009.1	473878.8	479128.9
4141	12500.0	50014.0	50689.8	300823.2
29308	476744.0	510244.4	510358.2	451794.1
7948	238863.0	259539.4	259853.0	371102.9
20496	47688.0	98695.4	98801.5	463490.5
44537	813093.0	809637.1	808831.0	584577.3
22747	418403.0	441085.8	441369.9	368981.5
13299	424730.0	408209.0	408226.2	375320.1

Figure 19

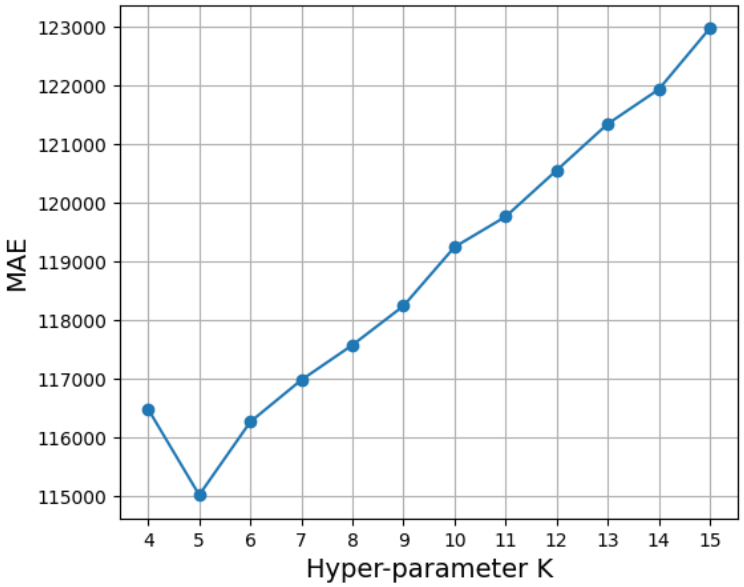


Figure 20

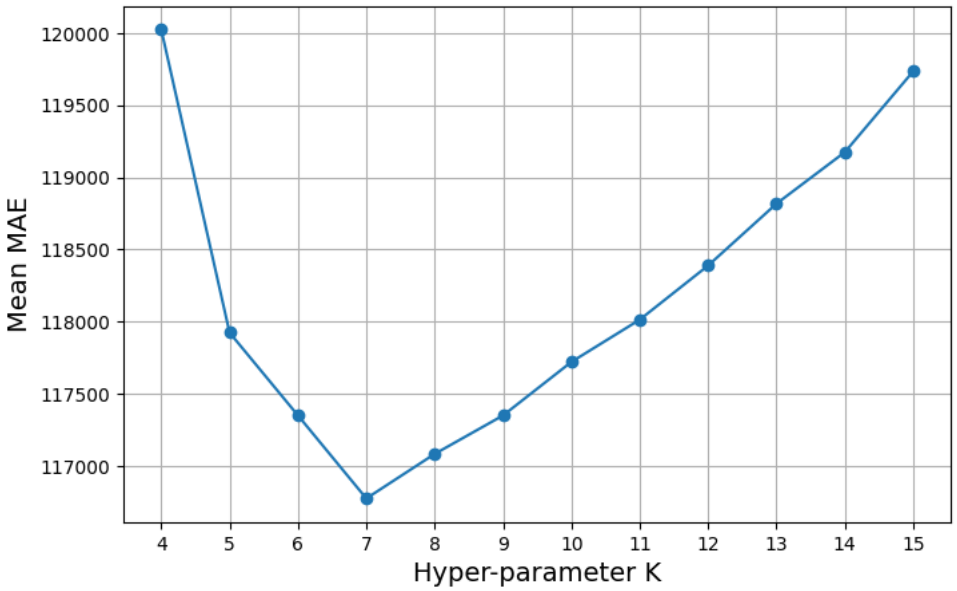
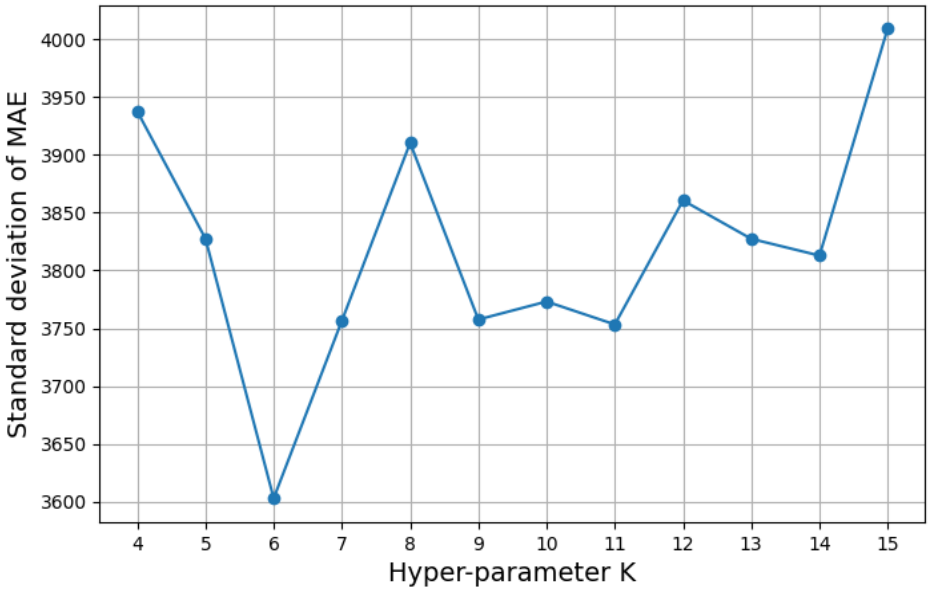


Figure 21



REFERENCES

- https://www.cdc.gov/nchs/fastats/nursing-home-care.htm

- https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1464018/
- https://www.britannica.com/topic/baby-boomers
- https://www.cdc.gov/nchs/pressroom/nchs_press_releases/2024/20240525.htm#:~:text=The%20general%20fertility%20rate%20in,consistently%20decreased%20by%202%25%20annually.
- https://www.epa.gov/smartgrowth/smart-growth-and-affordable-housing
- https://www.census.gov/data/datasets/time-series/demo/popest/2020s-national-detail.html
- https://saturncloud.io/blog/how-to-import-functions-from-another-jupyter-notebook/
- https://github.com/PacktPublishing/Hands-On-Predictive-Analytics-with-Python/blob/master/
- https://stackoverflow.com/questions/62669075/are-time-complexity-and-space-complexity-inversely-proportional
- https://www.geeksforgeeks.org/python-get-elements-till-particular-element-in-list/
- https://www.geeksforgeeks.org/python-get-elements-till-particular-element-in-list/
- https://www.geeksforgeeks.org/regularization-in-machine-learning/
- https://datascience.stackexchange.com/questions/40089/what-is-the-reason-behind-taking-log-transformation-of-few-continuous-variables
- https://www.kaggle.com/datasets/sadiquekhann/functions-and-loops-in-python
- https://stats.stackexchange.com/questions/545148/why-is-multicollinearity-different-than-correlation
- https://chat.openai.com/
- https://www.youtube.com/watch?v=rsyrZnZ8J2o