

**PREDICTIVE ANALYTICS FOR NURSING HOME INVESTMENT**

BANA 620 SPRING 2024







INDEX

1. EXECUTIVE SUMMARY
2. INTRODUCTION
3. METHODOLOGY
4. DATA DESCRIPTION
5. ANALYSIS AND FINDINGS
6. DISCUSSION
7. RECOMMENDATION
8. CONCLUSION
9. APPENDIX

10.REFERENCES

1.EXECUTIVE SUMMARY

**1.1 Overview**

Our team used predictive algorithms to assess the viability of investments and offer financial advice to clients who were interested in investing in nursing homes in the United States. We employed feature engineering to generate significant financial ratios and we used a mix of these ratios to generate the target variable called the "Investment Choice" , and we processed the data to resolve missing values, outliers, and inconsistencies using cost report data from 2015–2021, collected from CMS. We used PCA and scaling techniques to control the dimensionality of the data and normalize the inputs. Three machine learning models (KNN, Random Forest, and Logistic Regression) were chosen, trained on 75% of the data, and assessed with 75% of the data going toward training, 10% going toward validation, and 15% going toward testing. Accuracy, precision, recall, F1 score, confusion matrix, and ROC AUC were used to evaluate the performance of the model, and cross-validation was used to improve the model and reduce overfitting. The Random Forest model was refined and implemented in a production setting because it showed the best accuracy and AUC ROC. Using predictive analysis we can help our clients decide if they want to invest in nursing homes.

**1.2 Key findings:**

* The results show that there are good investment prospects in Florida, Ohio, California, where prospective returns are highest, hence we advise investing in these states' nursing facilities.
* We found that there was a 33% loss in Net income from 2020 to 2021 due to Covid-19.
* Urban areas have better occupancy compared to rural areas.
* Bed occupancy rate reduced by 15% from 2015 to 2021.
* Covid 19 increased the operating cost per resident.
* Due to covid 19 , debt increased by 100% in 2020, making financial loss to almost all the nursing homes across the US due to increased costs.
* We also saw a decrease in staff especially nurses during covid 19 but that began to improve after 2020.

**1.3 Recommendations:**

* Strategic investments in high-performing facilities
* Enhancing Operational Efficiency by maintainingbed occupancy rates to be higher than 80%.
* Crisis Preparedness and Response should be done by lowering the influence on operations from unforeseeable pandemics.
* Monitoring and Evaluationwill helporganize yearly strategy sessions to fine-tune tactics.

# 2.INTRODUCTION

## Nursing Home

Due to the significant number of baby boomers exploring nursing home options, our company has been tasked with determining whether it is advisable to suggest our client invest in nursing homes. As a member of the data analysis team, we are required to perform an analysis to address this question.

Our team has acquired a comprehensive dataset of U.S. nursing homes for the years 2015-2021. These datasets allow us to compare the average performance of nursing homes nationwide using different metrics.

**Objective**:

We calculated financial ratios from the variables in cost reports to create a target variable called Investment choice which is a categorical variable. Using this we aim to answer a few questions that will aid the client to make investment decisions in nursing facilities.

1. Based on the analysis, is it advisable for clients to invest in nursing homes, and if so, what criteria should guide these investment decisions?
2. What is the overall financial performance of nursing homes in the United States from 2015 to 2021?
3. Which factors most significantly affect the financial performance of nursing homes?
4. How did the COVID-19 pandemic impact the operational and financial metrics of nursing homes?

# 3. METHODOLOGY

**3.1 Data Collection Process:**

Our analysis made use of a broad dataset spanning the financial performance of nursing homes in the United States from 2015 to 2021. The annual expense report files, which were generated and supplied by healthcare databases, were the source of the data. Individual CSV files including a variety of financial and operational variables relevant to nursing homes were created from the dataset for each year.

**3.2 Data Preprocessing:**

The data preprocessing steps were crucial in ensuring the quality and consistency of the analysis. These steps included:

* Normalization of Column Names
* Merging Data
* Handling Missing Values
* Feature Engineering

**3.3 Data Analysis Techniques:**

To gain insights, the analytical approach we used descriptive statistics and machine learning algorithms.

* Descriptive Statistics: To comprehend the distribution and core tendencies of important financial indicators, basic statistical measures including mean, median, and standard deviation were produced.
* Correlation Analysis: To determine and show the connections between various operational and financial measures, a correlation matrix was created.
* Algorithms for Machine Learning: Classification Model To ascertain investment eligibility based on predetermined financial health standards, a binomial classification model was created. Using this concept, nursing facilities can be categorized as either suitable or not for investment.
* Validation of the Model: Cross-validation methods were used to validate the model in order to guarantee its reliability and correctness. To improve the performance of the model, the K-Nearest Neighbors (KNN) algorithm was specifically tested for a range of values of 'n\_neighbors'.
* Principal Component Analysis (PCA): By concentrating on the most important features, PCA was used to decrease the dataset's dimensionality, increasing the computing efficiency of machine learning methods and possibly improving model performance.

**3.4 Software and Programming Languages:**

* Python's powerful tools and frameworks for data analysis and machine learning were widely employed throughout the project.
* Pandas with NumPy: For working with data and performing mathematical operations.
* Matplotlib and Seaborn: These tools are used for data visualization, which includes making plots and heatmaps to show conclusions.
* Scikit-learn: Used for PCA, cross-validation, and the implementation of machine learning algorithms.
* Jupyter Notebook: Jupyter Notebooks provide an interactive coding environment that is ideal for exploratory data research. All code was generated and performed within these notebooks.

# 4.DATA DESCRIPTION

**4.1.Data Exploration:**

Analyze the cost reports data for the past 7 years. We had 192 columns of data which was reduced to 63 columns as we selected the common columns that were significant in cost reports data.

**4. 2. Data Preprocessing:**

* **A) Missing values**: This involves handling missing values using Mean imputation for Numerical columns and Mode imputation for Categorical columns.
* **B) Outliers:** Handling outliers by removing extreme values and replacing other outliers with closest values.
* **C) Inconsistent data:** We removed inconsistent data, two columns with same variable had inconsistency in naming:
* **D) Normalization:** We used Normalization to remove data inconsistency by replacing space between variable names with underscore.

**4.3. Feature Engineering:**

We performed feature engineering to get our target variable **Investment choice**, which is a mix of six financial ratios that we calculated as shown below.In the analysis of nursing homes' financial performance and operational efficiency, several key features were engineered to provide insights into their profitability, financial health, and operational effectiveness. Stakeholders might find possible areas for investment and operational improvement by looking at these metrics.

A) Return on Investment (RoI) = Net income/ total Assets

B) Operating Margin = Operating Income/ Net Patient Revenue

C) Debt to Equity Ratio = Total Liabilities / Total fund

D) Current Assets to Current liabilities = Total Current Assets / Total Current liabilities

E) Bed Occupancy Rate = Total Days Total / (number of beds x 365)

F) Days in accounts receivable = Accounts Receivable /(Net Patient Revenue x 365)

**4.3.1. Profitability:**

Return on Investment (ROI):It is a measure of profitability compared to total investments. It is calculated as the ratio of net income to total investment funds.

Significance: ROI is a universal measure of profitability and capital efficiency. A greater ROI suggests that nursing homes generate more income per dollar of investment, which is good news for potential investors.

Net Operating Margin: The percentage of income remaining after all expenses are paid. It is a measure of operational efficiency and pricing plan efficacy.

Significance: Operating margin indicates management's efficiency in regulating costs and expenses in relation to the facility's revenue. Higher margins indicate improved management performance and financial stability.

**4.3.2. Financial Performance:**

Debt to Equity Ratio (D/E):This ratio compares total liabilities to shareholders' equity, demonstrating the extent to which a nursing home finances its operations through debt rather than fully owned money.

Significance: The D/E ratio is an important indicator for determining the balance of debt and equity in a nursing home's finances. A smaller ratio is often regarded as good since it reflects less reliance on debt, which reduces financial risk.

Current Ratio:This ratio assesses a nursing home's ability to repay short-term liabilities with short-term assets.

Significance: The current ratio measures liquidity and short-term financial health. A ratio greater than one indicates that the nursing home has sufficient assets to meet its short-term commitments, which is critical for maintaining financial stability and operational continuity.

**4.3.4. Operational Efficiency**

Bed Occupancy Rate:Over a certain time period, this rate indicates the proportion of beds that are occupied by patients relative to all available beds.

Significance: The bed occupancy rate is an essential operational measure that shows how well the facility's resources are being used. Elevated occupant rates typically signify efficient administration and a market need for the assisted living facility's offerings.

Days in Accounts Receivable (DAR): Interpretation: This measure shows the typical time it takes for a nursing facility to get past due payments.

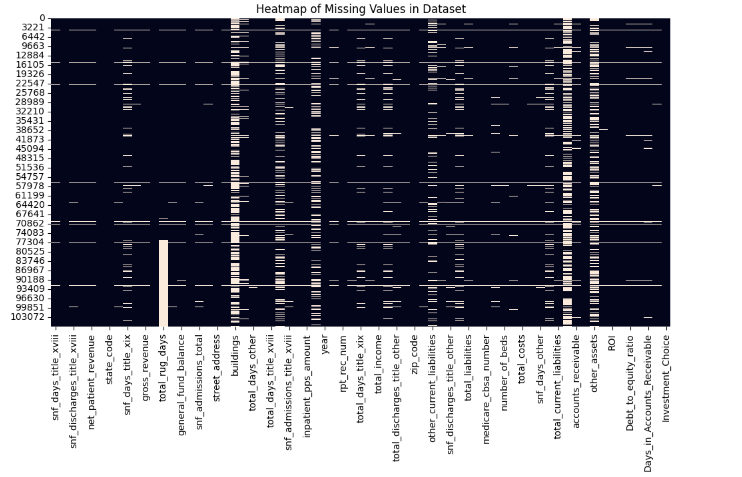
Significance: DAR is a key metric for assessing how well cash flow management and billing activities are working. Reduced values signify an enhanced revenue collection efficiency of the nursing home, so bolstering operational liquidity and mitigating the likelihood of cash flow disruptions.

5.ANALYSIS AND FINDINGS

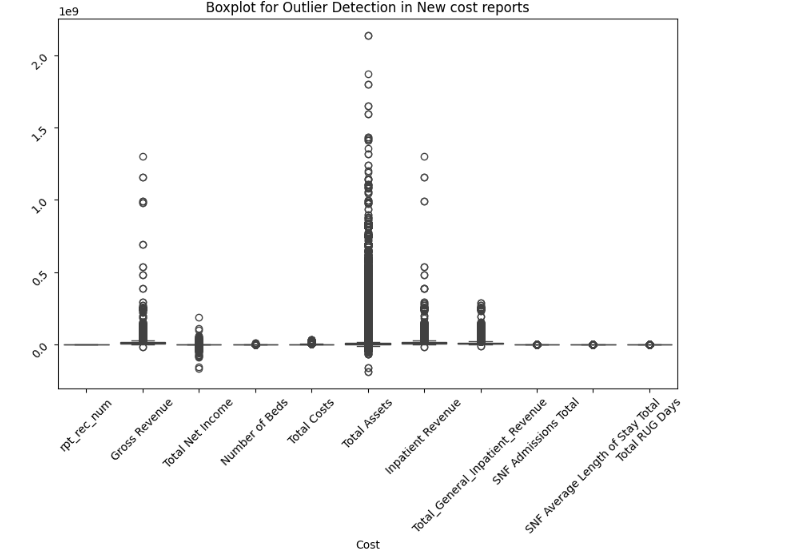
Our analysis consisted of the following:

**5.1. Data Preprocessing Analysis:**

1. **Handling missing values by mean imputation and mode imputation.**



1. **Outliers were handled by removing extreme values and replacing other outliers by closest values.**



1. **Inconsistent Data was found where the same variables had similar names causing inconsistency.**

A white background with many text

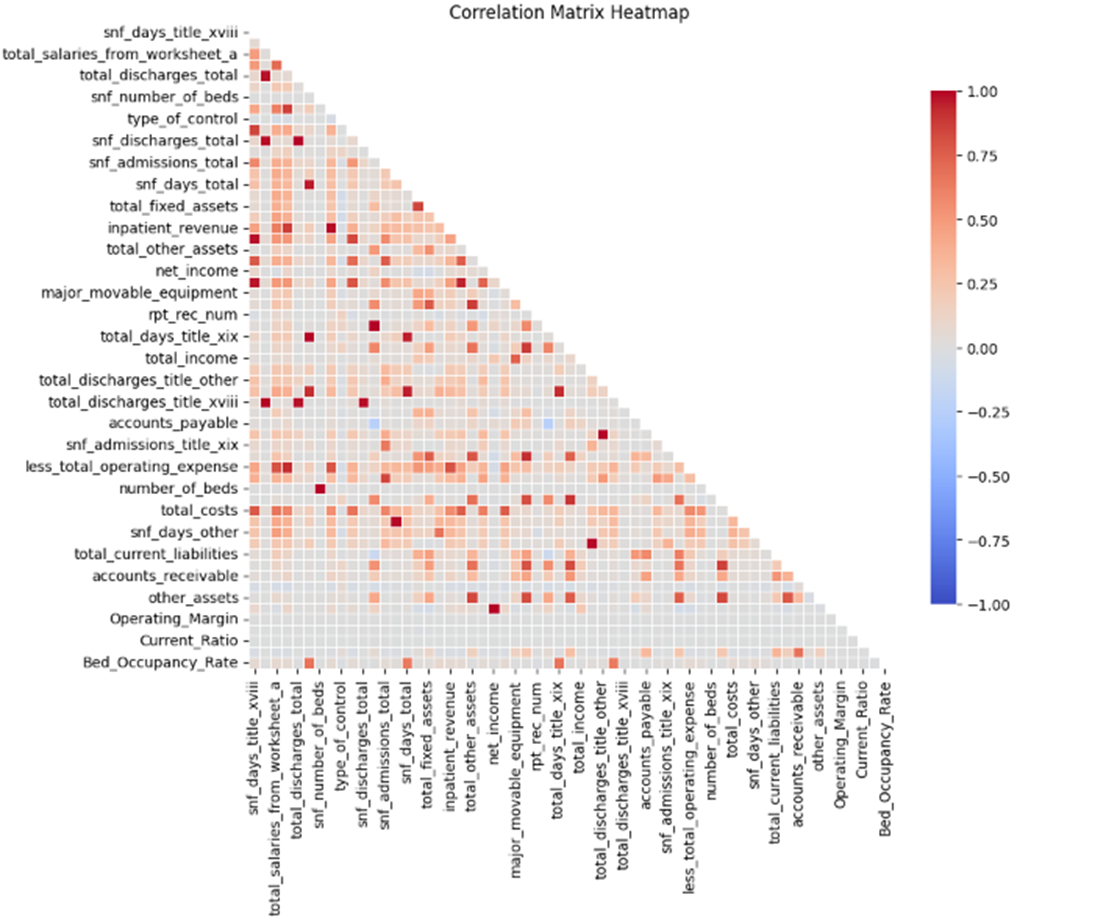
Description automatically generated with medium confidence

* We replaced the spaces with underscore so we can avoid data inconsistency :



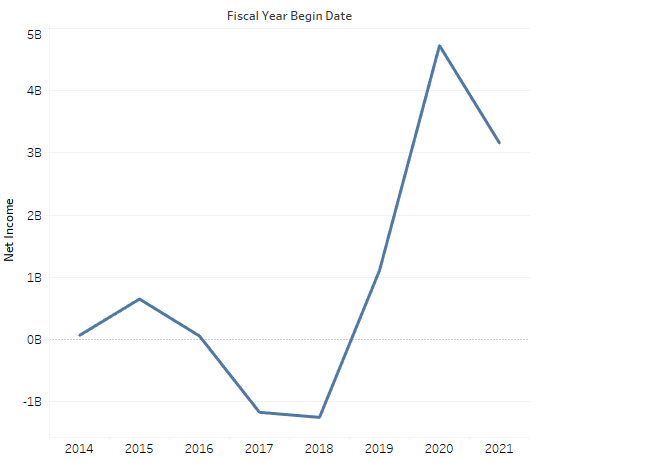
1. **Correlation Matrix:**

* We created a correlation matrix of all the variables and analyzed that to find the following.
* Red: Indicates a positive correlation, where values closer to +1.00 suggest a strong positive relationship.
* Blue: Indicates a negative correlation, where values closer to -1.00 suggest a strong negative relationship.
* White or Pale Colors: Indicate little to no correlation between the variables.
* Net Income: The heatmap indicates that there are positive connections of varied degrees between the variable "net\_income" and a number of other financial measures. This implies that higher values in relevant financial measurements, such as total assets and perhaps total revenue, may be linked to higher net income.



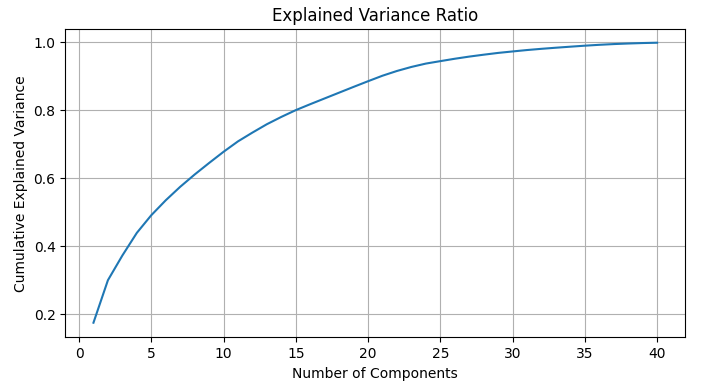
**5.2. Exploratory Data Analysis**

1. **Net income over years :** We find that there was a significant growth from 2018 to 2019 in net income but after covid 19 we can see a decline in net income.

****

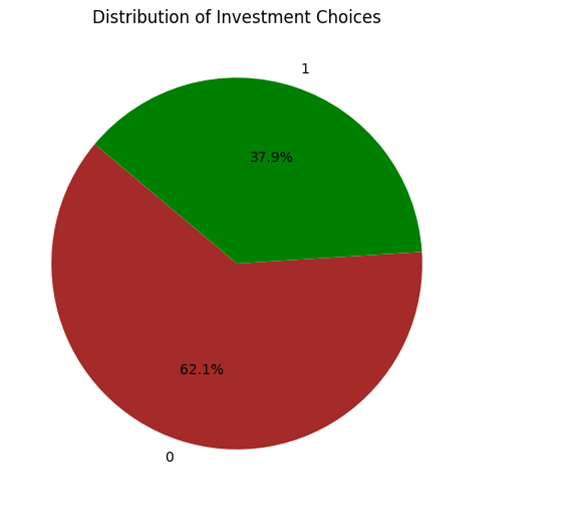
1. **Choosing number of components**

* X-Axis (Number of Components): Represents the number of principal components considered in the PCA.
* Y-Axis (Cumulative Explained Variance): Represents the cumulative proportion of the dataset's total variance that is explained by the first n components.
* Approximately 99% Variance: The curve gets closer to capturing roughly 99% of the total variance by about 40 components. This may be the best cutoff, depending on the degree of variance retention that is thought to be required for your analysis.



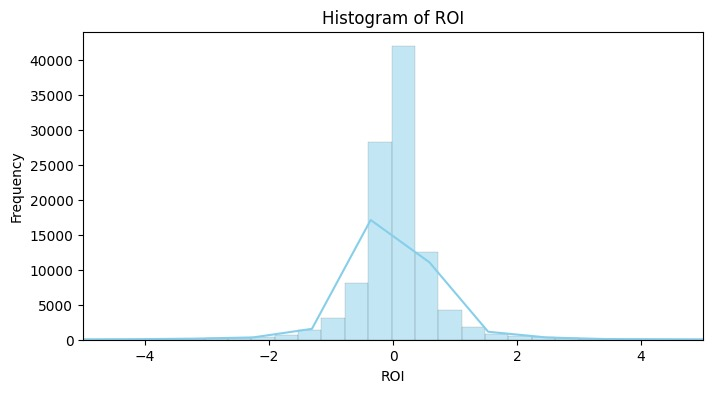
1. **Target variable- Investment Choice:**

* The distribution of investment options divided into two categories—0 and 1—is displayed in the pie chart you provide. Based on the data in the chart, the following analysis has been conducted:
* Description of the Chart: Category 0 (Red): Indicates the group of investment options deemed "Not Suitable" for investing. With 62.1% of the total, this category represents a bigger chunk of the pie.
* Classification 1 (Green): Indicates the category that has been deemed "Suitable" for investment. 37.9% of the total is accounted for by this category.

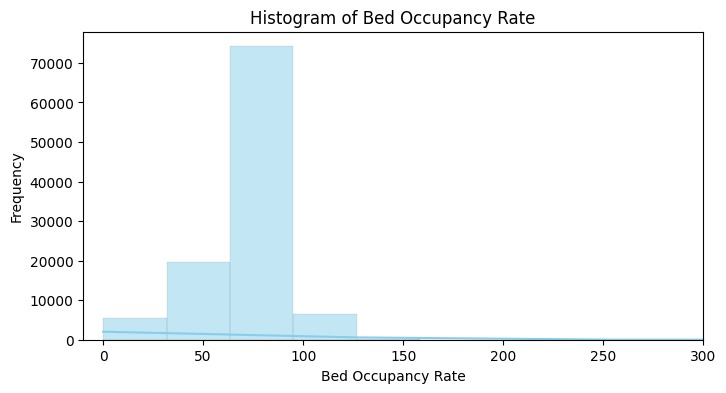


1. **Ratios created for Target Variable:**

* The most frequent ROI value falls between -2 and 0. This means that most investments resulted in a loss.
* The ROI is right skewed. This means that there were more investments with a negative ROI than a positive ROI, but there were also a few outliers with a very positive ROI.

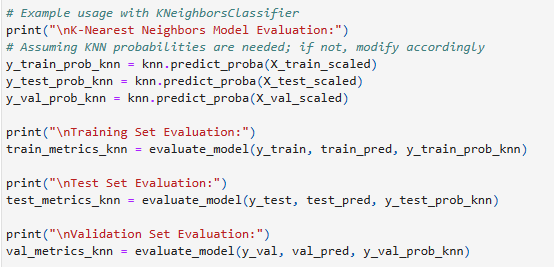
****

* According to the distribution, most facilities appear to be running at full capacity, which is an important sign for healthcare facility design and management. Healthcare stakeholders may find this data useful in evaluating patient demand, operational efficiency, and the need for growth or extra resources.

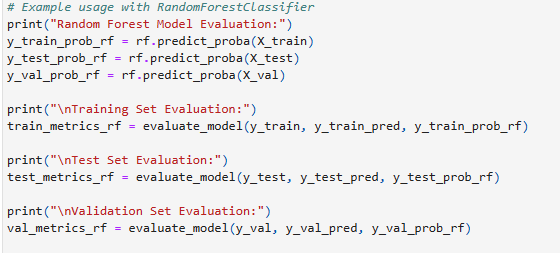
****

**5.3. Model employment**

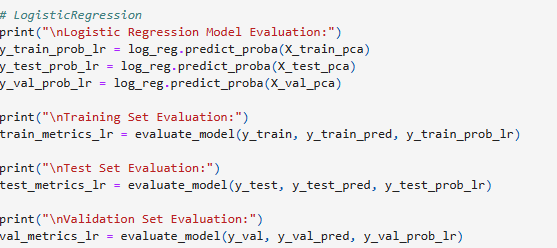
1. **KNN :** KNN is a flexible classifier that, in spite of its ease of use, may be very successful in solving some kinds of classification issues, such as binomial classification assignments, which involve selecting nursing home investments based on comparable past data.



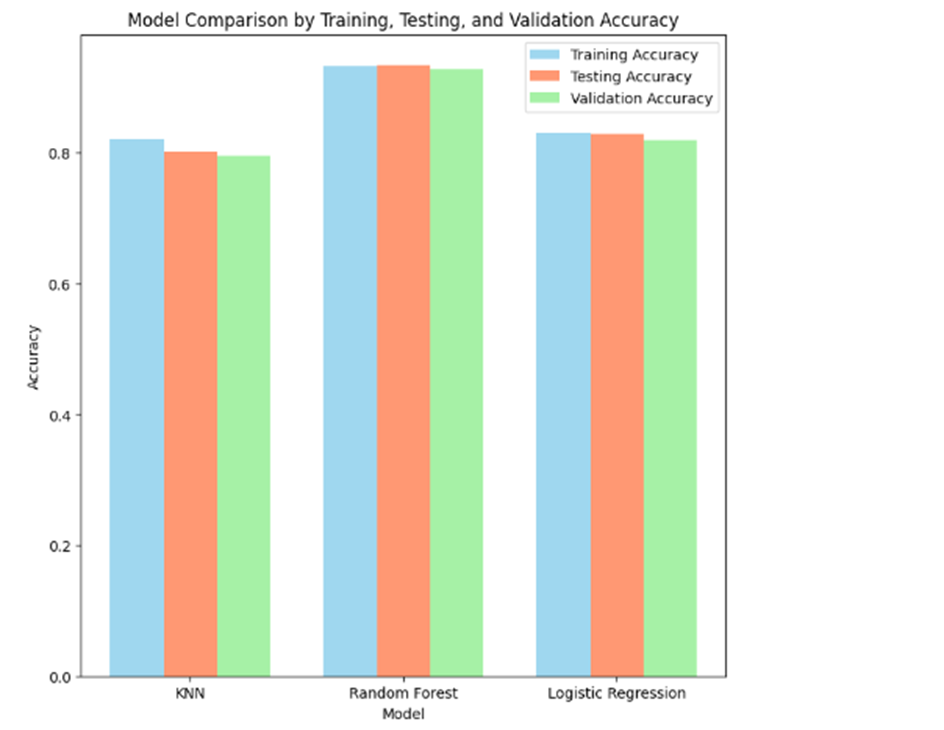
1. **Random Forest :** Because of its high accuracy, resilience to overfitting, and capacity to manage big datasets including a combination of numerical and categorical information, Random Forest is a potent and effective machine learning approach for binomial classification issues.



1. **Logistic regression:** For binomial classification tasks, logistic regression is a reliable and simple approach that provides unambiguous probabilistic interpretations and effective computing. Due to its broad application in a variety of fields, such as social sciences, healthcare, and finance, it offers valuable insights into the variables determining binary outcomes.



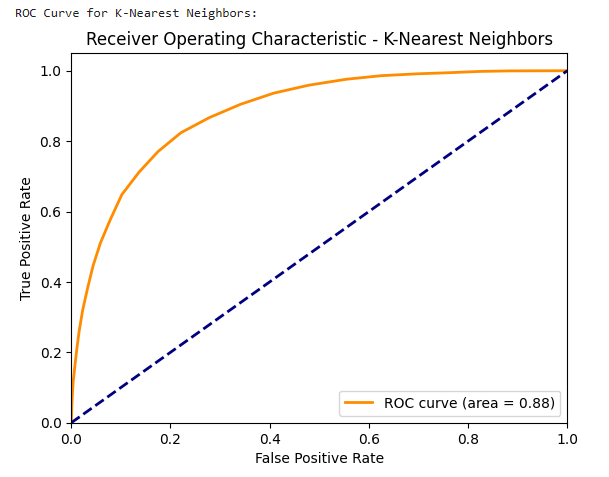
**Model Comparison:** Random forest has the highest accuracy in training, testing and validation set.



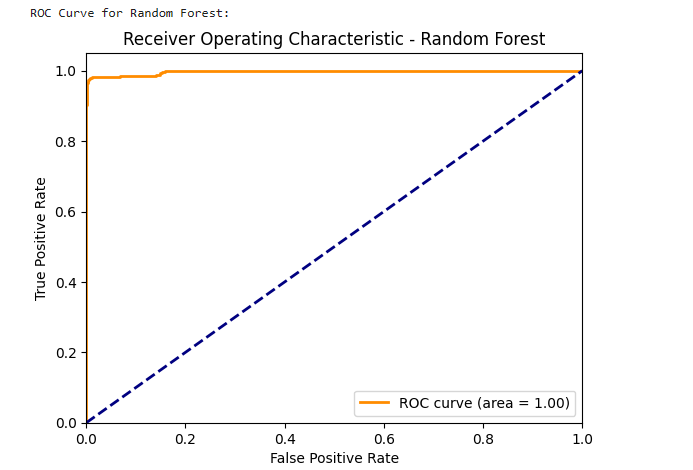
**5.4. Model Evaluation**

| **Logistic Regression Model Evaluation:** | **K-Nearest Neighbors**  **Model Evaluation** | **Random Forest Model**  **Evaluation:** |
| --- | --- | --- |
| **Training Set Evaluation:**  Accuracy: 0.9194  Precision: 0.8345  Recall: 0.9824  F1 Score: 0.9024  **Confusion Matrix:**  [[43565 5894]  [ 533 29709]]  ROC AUC: 0.9258 | **Training Set Evaluation:**  Accuracy: 0.9194  Precision: 0.8345  Recall: 0.9824  F1 Score: 0.9024  **Confusion Matrix:**  [[43565 5894]  [ 533 29709]]  ROC AUC: 0.9053 | **Training Set Evaluation:**  Accuracy: 0.9194  Precision: 0.8345  Recall: 0.9824  F1 Score: 0.9024  **Confusion Matrix:**  [[43565 5894]  [ 533 29709]]  ROC AUC: 0.9922 |
| **Test Set Evaluation:**  Accuracy: 0.9211  Precision: 0.8351  Recall: 0.9836  F1 Score: 0.9033  **Confusion Matrix:**  [[8806 1160]  [ 98 5876]]  ROC AUC: 0.9267 | **Test Set Evaluation:**  Accuracy: 0.9211  Precision: 0.8351  Recall: 0.9836  F1 Score: 0.9033  **Confusion Matrix:**  [[8806 1160]  [ 98 5876]]  ROC AUC: 0.8820 | **Test Set Evaluation:**  Accuracy: 0.9211  Precision: 0.8351  Recall: 0.9836  F1 Score: 0.9033  **Confusion Matrix:**  [[8806 1160]  [ 98 5876]]  ROC AUC: 0.9928 |
| **Validation Set Evaluation:**  Accuracy: 0.9132  Precision: 0.8240  Recall: 0.9815  F1 Score: 0.8959  **Confusion Matrix:**  [[5735 848]  [ 75 3970]]  ROC AUC: 0.9217 | **Validation Set Evaluation:**  Accuracy: 0.9132  Precision: 0.8240  Recall: 0.9815  F1 Score: 0.8959  **Confusion Matrix:**  [[5735 848]  [ 75 3970]]  ROC AUC: 0.8757 | **Validation Set Evaluation:**  Accuracy: 0.9132  Precision: 0.8240  Recall: 0.9815  F1 Score: 0.8959  **Confusion Matrix:**  [[5735 848]  [ 75 3970]]  ROC AUC: 0.9913 |

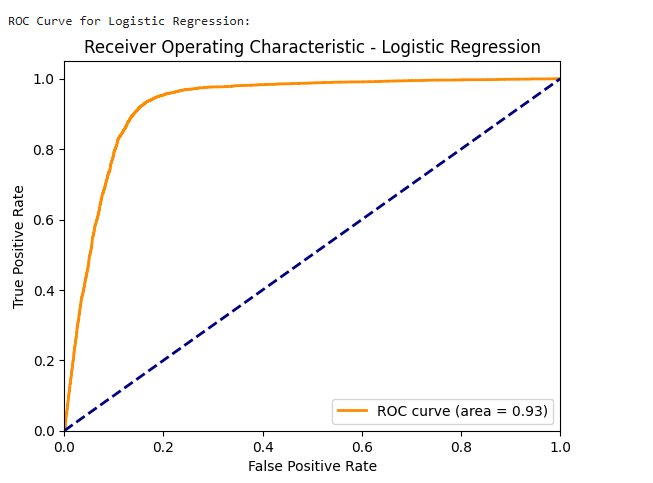
1. **Accuracy:** It measures the proportion of true results so we can see that the Random forest model has the highest accuracy.
2. **Precision**:It is the ratio of correctly predicted positive observations , showing how relevant the model’s classifications are.All 3 models had a similar value.
3. **Recall:** It indicates how well the model can find the positive samples. This value was the highest for a random forest model.
4. **Confusion matrix:** Describes the number of correct and incorrect predictions with count and we can see that Random forest has the most accurate matrix.
5. **F1 score:** Conveys the balance between the precision and the recall.
6. **AUC ROC Curve:** The AUC gives a single measure of how well the prediction model is capable of distinguishing between classes. As seen in the plots below we can see the random forest model is near perfect score. Then we have the logistic model with AUC ROC score of 0.92. The KNN model has the least score of 0.87. But all three models have a good score and help us in predicting analytics.



* The area under the ROC curve for the K-Nearest Neighbors (KNN) model is 0.88, which indicates a good performance for the model. The model is good at distinguishing between true positives and false positives.

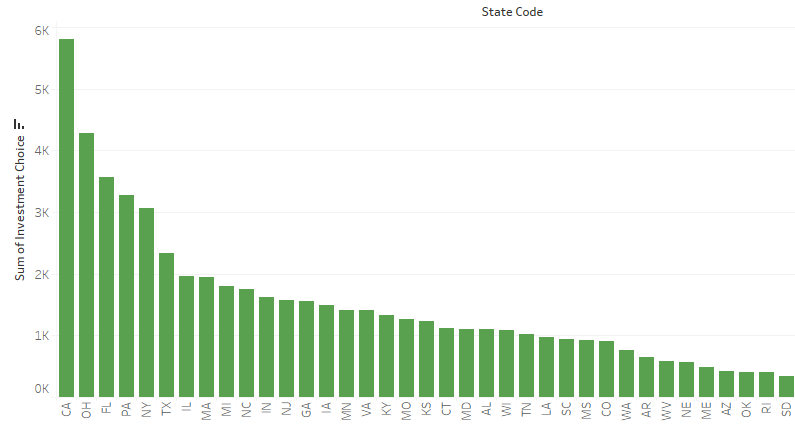


* The area under ROC curve is 1.0, indicating perfect performance for the random forest model in classifying between positive and negative classes.



* The area under ROC curve (AUC) 0.93, which indicates a strong performance for the logistic regression model in classifying between positive and negative classes.

**5.5 Best states to invest in :**



* We found that the best states to invest in based on Investment Choice are

1. California
2. Ohio
3. Florida

6. DISCUSSION

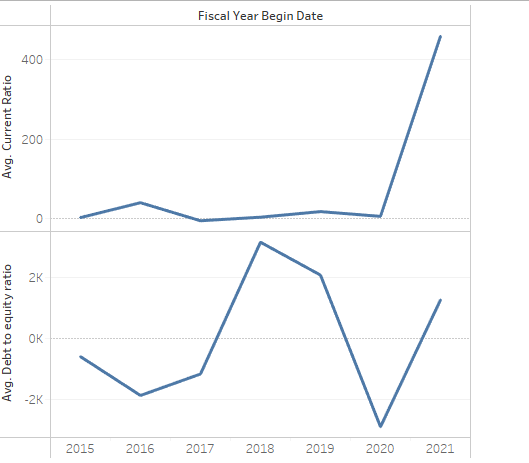
Let us discuss the questions we aimed to answer by performing predictive analytics:

1. **Based on the analysis, is it advisable for clients to invest in nursing homes, and if so, what criteria should guide these investment decisions?**

We can see that out of all the facilities in our dataset 37.9% are a good investment because they bring in a profit. A majority of these facilities are in Florida, Ohio and California. So clients should focus on nursing homes in these three states the most.

1. **What is the overall financial performance of nursing homes in the United States from 2015 to 2021?**

Nursing homes have a good financial performance after 2018 till 2020. Current Ratio and Debt to Equity Ratio will be helpful to measure the financial performance. We see a sharp decline after 2020.

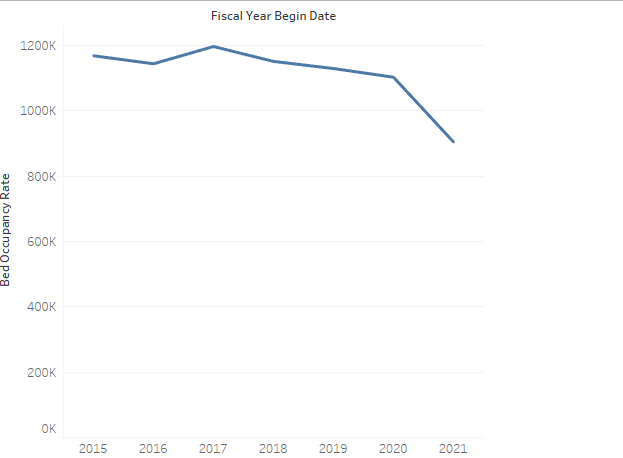
****

1. **Which factors most significantly affect the financial performance of nursing homes?**

We found that net income impacted the financial performance of nursing homes significantly. We can see that bed occupancy rates dropped by 15%.

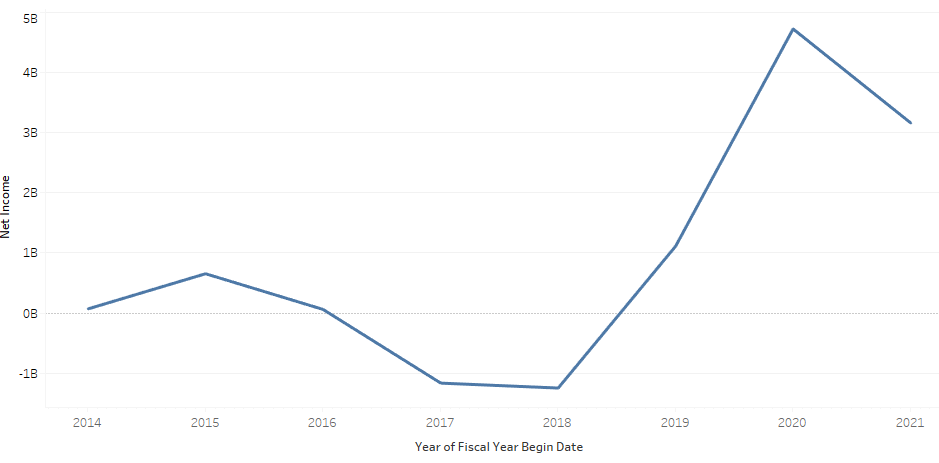
The chart that shows a declining trend in nursing homes' bed occupancy rates between 2015 and 2021 provides important information about the variables influencing their financial success. A confluence of demographic, economic, and regulatory variables is responsible for the reported fall, which became more noticeable after 2019.

Significant financial difficulties have been plaguing nursing homes, which have been made worse by changes such as growing operational expenses, and heightened competition from alternative care models. Furthermore, the COVID-19 pandemic has had a significant impact, resulting in increased operational difficulties and modifications to patient admission patterns. Nursing homes need to concentrate on improving operational effectiveness, upholding high standards of care, and strategically responding to shifting market and demographic situations in order to overcome these obstacles. Stabilizing occupancy rates and enhancing financial results may be achieved by focusing on the quality of care and investigating novel business models, such as integrated care services. These calculated actions are necessary to continue providing the aging population's vital care needs as well as to ensure financial sustainability.



1. **How did the COVID-19 pandemic impact the financial metrics of nursing homes?**

The significant variations in net income between 2014 and 2021 show how the COVID-19 epidemic has affected nursing facilities' financial measures. Prior to the pandemic, the industry experienced difficulties, as seen by a notable decline in net profits in 2018. Although the rebound in 2019 was encouraging, it brought to light persistent worries about the industry's financial soundness. When the pandemic struck in 2020, things changed in ways never seen before. These considerations highlighted the need for outside financial assistance to handle financial crises, even while they momentarily improved the financial situation of nursing facilities. Even if the 2021 decrease was not as bad as the 2018 one, it nevertheless highlights the brittleness of this financial upswing and highlights how unsustainable the financial gains from the epidemic were.

****

7.RECOMMENDATION

**1. Strategic investments in high-performing facilities:**

* Give top priority to funding nursing facilities with operating margins (>2%) and return on investment (>5%), as well as those with excellent financial health metrics such debt to equity ratios (below 1.5) and current ratios (above 1.5).
* Reduce investment risks by emphasizing facilities with sound financial standing.
* Boost revenue generation and operational efficiency to maximize returns on investments.

Plan of Implementation:

* Short-term: Within the following six months, thoroughly evaluate the finances of prospective investment targets in order to pinpoint high-performing facilities.
* Mid-term: Within a year, negotiate the purchase of a major interest in the facilities that have been identified.
* Long-term: Track and modify the investment portfolio each year in accordance with performance evaluations that are conducted.

**2. Enhancing Operational Efficiency:**

* Enhance bed occupancy rates to be higher than 80%.
* Cut the number of days in accounts receivable to under 80 days by improving patient care and billing procedures.
* Increased revenue is anticipated as a result of better use of resources and assets.
* increased cash flow and decreased credit risk due to unpaid receivables.

Plan of Implementation:

* Short-term: Within the next 12 months, put electronic health record systems and sophisticated scheduling into place to maximize occupancy rates and simplify billing procedures.
* Mid-term: To sustain efficiency improvements, train employees on new systems and principles of continuous improvement.

**3. Crisis Preparedness and Response:**

* Create strong systems and procedures to handle medical emergencies, with an emphasis on staff education, infection control, and emergency response.
* Lower the influence on operations from upcoming pandemics or health emergencies.

Plan of Implementation:

* Short-term: Within the next three months, evaluate the infrastructure deficiencies and current health and safety procedures.
* Mid-term: Over the following 18 months, improve the emergency response and health monitoring systems.
* Long-term: To guarantee preparedness and adherence to health laws, establish continuing training programs and health safety checks.

**4. Technological Advancements and Innovation**

* Make investments in cutting-edge technologies including data analytics systems, automated patient monitoring, and telemedicine.
* improved patient outcomes and treatment quality.
* automation leads to lower labor expenses and more operational efficiency.

Plan of Implementation:

* Short-term: Within six months, identify technology providers and launch test programs in a few chosen facilities.
* Mid-term: Within two years, implement all facilities with successful technology.
* Long-term: To remain at the forefront of healthcare innovation, constantly assess and incorporate new technology.

1. **Monitoring and Evaluation**

* Quarterly Reviews: Evaluate how implemented changes are affecting financial and operational indicators by conducting quarterly performance reviews.
* Organize yearly strategy sessions to fine-tune tactics in light of current market conditions and internal performance information.
* Investors and operators of nursing homes could anticipate significant gains in operational effectiveness, financial stability, and crisis resistance by putting these suggestions into practice.
* In addition to increasing financial returns, this strategic approach raises resident quality of care and boosts the industry's general viability and reputation in the nursing home sector.

8.CONCLUSION

We have learned a lot from our extensive research on using predictive analytics to finance nursing home investments. Notably, we have identified opportunities for targeted investments in Florida, Ohio, and California, among other states. According to our study's predicted models and historical data, these states provide the highest possible returns. The analysis of financial performance from 2015 to 2021 shows a significant variation that was largely caused by outside variables such as the COVID-19 pandemic, which presented previously unheard-of operational and financial difficulties for the industry.

The Random Forest model, which beat other predictive models like KNN and Logistic Regression in terms of accuracy and ROC AUC scores, allowed for a more in-depth knowledge of the elements that influence investment viability. The model's accuracy in classifying about 37.9% of nursing homes as desirable investments highlights how well cutting-edge machine learning approaches work to inform financial choices in the healthcare real estate market.

The pandemic's effects also highlighted how urgent it is for nursing homes to adjust to the quickly shifting economic and health landscapes. The investigation showed that the pandemic significantly reduced operational efficiencies, especially bed occupancy rates. This finding implies that future investments should take into account the facilities' ability to endure such crises in addition to their existing financial health.

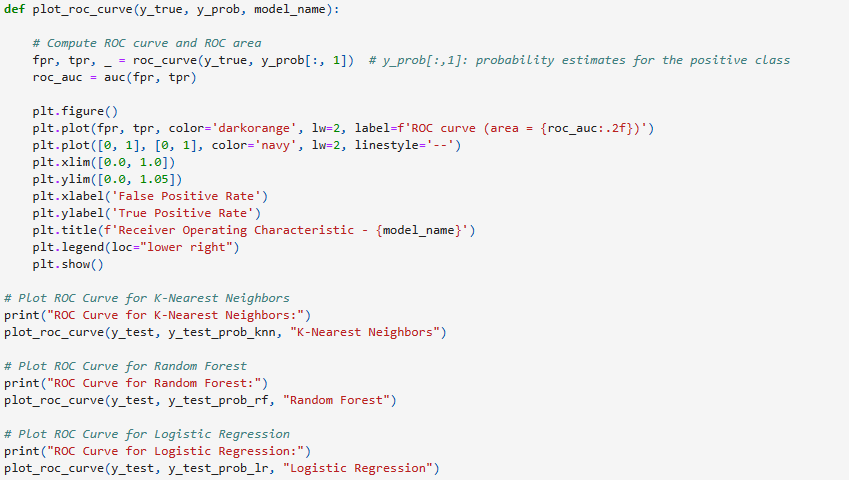
It is recommended that investors approach nursing home investments using the following approach:

* Investing selectively means concentrating on areas and establishments that, according to our prediction models, exhibit strong profitability and resilience.
* Operational Efficiency: Putting money into systems and technologies that improve efficiency means that facilities can either immediately or gradually return to their ideal occupancy rates.
* Crisis Preparedness: Assessing and funding establishments that have strong infection control, flexible emergency plans, and the ability to continue operations in the event of a health emergency.

In summary, the nursing home industry offers promising investment prospects; nevertheless, these are mainly limited to particular areas and rely on the facilities' capacity to efficiently manage operational efficiency and crisis response. Through the utilization of sophisticated analytics and a steadfast emphasis on sustainability and quality, investors can maximize their portfolios to augment returns and make a constructive impact on the elderly population's care. This strategic approach will support the sector's overall sustainability and resilience by optimizing financial returns while simultaneously improving the quality of care delivered.

9.APPENDIX

Code snippets for our KNN, Logistic Regression and Random forest Auc Roc curve. Rest of our code can be found in Investment.html file attached.



REFERENCES

* More information about the dataset can be found at :

[**https://data.cms.gov/provider-data/topics/nursing-homes**](https://data.cms.gov/provider-data/topics/nursing-homes)

* [**https://corporatefinanceinstitute.com/resources/accounting/financial-ratios/**](https://corporatefinanceinstitute.com/resources/accounting/financial-ratios/)
* **https://www.investopedia.com/articles/basics/10/guide-to-calculating-roi.asp**
* **https://gocardless.com/en-us/guides/posts/top-7-financial-ratios/**