CALIFORNIA STATE UNIVERSITY, NORTHRIDGE

Project Report

By

Lekhana Chandra Palamuri

Sri Sai Manogna Gollapalle

Namratha Eti

Under the Guidance of

Professor. Akash Gupta

Spring 2024

BANA620-Data Mining

Submission Date- 5/7/2024

Comprehensive Report on Nursing Homes

Summary

An actual study of the data from the healthcare facility is presented in this project report over the years. Some of the key findings point to the fact that there is significant variation in the gross revenue, net income, and patient revenue among various facility types in various regions. Recommendations target improved operational effectiveness more specifically with respect to facilities that contain lower revenues and higher costs. The review brings out information from a number of the healthcare facilities with respect to key financial performance measures such as trends of total income, operating expenses, and net income over some time.

Introduction

The healthcare sector within the United States functions in a complex financial environment characterized by variable funding sources, regulatory pressures, and high operational costs. The issue of financial sustainability, therefore, is a paramount concern to the health care administrators. Subsequently, will have to measure the ability to be able to offer quality care that maintains its current level, takes care of all staff concerns, and invests in the technology that might be needed. This report will dig into the financial performance of over 106,000 health institutions based on data drawn between the years 2015 and 2021. The report will aim to establish the rural-urban financial trends and differences within the health facilities. This is an effort to identify the fundamental causes of the nation's financial health and, by properly analyzing the financial data, offer some insights that can help strategists and policymakers make decisions that will help the country move forward financially. These analyses will be needed with ongoing reforms in healthcare and other financial "unknowns" that come along with public health threats like the COVID-19 pandemic.

Project Objectives

- How does the financial efficiency of rural hospitals compare to urban hospitals in terms of net income, gross revenue, over the years?

- How does the location (rural vs. urban) influence the financial outcomes and patient service revenues of hospitals?

- Can we predict the financial stability of hospitals based on historical revenue, expenses, and other operational metrics?

Methodology

The methodology is based on a multistep approach and details of the data collection, analysis, and interpretation. The project interpretation employs statistical and machine learning tools in specifying a few.

Data Collection:

Data were drawn from the cost reports, which derive from health care facilities and have metrics of details on finances and operation. In this stage, several datasets were summarized in order to achieve one large dataset that had a consistent format across all of them. The imputation and removal of records with excessive missing values were performed at the end of this stage as a missing data handling technique.

Data Analysis Techniques:

Descriptive statistics: Some simple summary statistics, including the mean, median, and standard deviation, have been computed in order to get a bearing on the distribution of some key metrics like gross revenue, net income, and number of beds, among the facilities.

Visual Analysis: The histograms and boxplots did succeed in showing the distribution and trend over the years.

Correlation Analysis: This is used in order to understand the relationship between financial metrics with each other and the characteristics of the facility.

Machine Learning Algorithms:

Random Forest and Decision Trees: Use both the decision trees and random forests for predicting the categorical dependent variable against input features, respectively. It was, among other metrics, evaluated based on accuracy and the Area Under Curve (AUC) of the Receiver Operating Characteristic (ROC) curve.

Logistic Regression: It is a tool used in predicting binary outcomes. The accuracy of the logistic regression model is realized at a level of determination using the confusion matrix.

Software and Programming Languages:

Python: Principal language of work in this project, used for data manipulation, analysis, and modelling.

- Pandas & NumPy: For data manipulation and numerical computations.
- Matplotlib & Seaborn: Used to create plots and other visualizations in advance to explore data.
- Scikit-learn: Scikit-learn was used for implementing and evaluating machine learning models such as Decision Trees, Random Forests, and Logistic Regression.

Data Description

The dataset used in the project is the comprehensive operational and financial data of the health facility in the United States over the year 2015 to 2021, having 106,269 records where each record has 193 attributes with detailed information on different dimensions of the health facility with respect to geography, finance performance, patient populations, and types of service offers.

Characteristics of Data:

Geographic Information: City, state, and county can also allow regional analysis.

Financial Metrics: It covers issues of gross revenue, net patient revenue, and other indicators showing the economic status of the facilities.

Operational Metrics: Details about how many beds and the types of service that are offered in the facility will give one a clear picture of size and scope of operations.

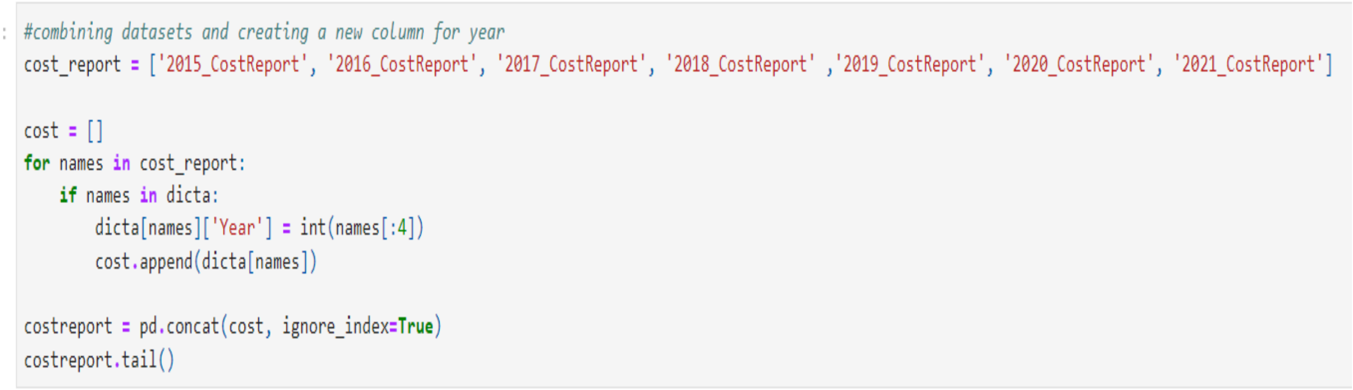
Data Cleaning and Preprocessing:

● **Datasets**: Using the panda module, the study simple merging cost report datasets from several years into a single data frame with a new column called "Year".

● **Managing Missing Values:** Any columns with more than 50% missing values will be deleted from analysis, as will the occurrence in which those values are absent.

● **Duplicate Columns**: After the dataset was cleaned, any two columns containing the

Identical information was found and eliminated.



```
#combining datasets and creating a new column for year
cost_report = ['2015_CostReport', '2016_CostReport', '2017_CostReport', '2018_CostReport' ,'2019_CostReport', '2020_CostReport', '2021_CostReport']

cost = []
for names in cost_report:
    if names in dicta:
        dicta[names]['Year'] = int(names[:4])
        cost.append(dicta[names])

costreport = pd.concat(cost, ignore_index=True)
costreport.tail()
```

[ ]:

	rpt_rec_num	Provider_CCN	Facility_Name	Street_Address	City	State_Code	Zip_Code	County	Medicare_CBSA_Number	Rural_versus_Urban	...	Less Contractual Allowance and discounts on patients' accounts	Net Patient Revenue	Less Total Operating Expense	Net Income from service to patients	Total Other Income	I
106264	1365869	NaN	NaN	NaN	SEATTLE	NaN	NaN	KING	NaN	NaN	...	690392.0	8661500.0	9750271.0	-1088771.0	129009.0	-95
106265	1365888	NaN	NaN	NaN	DALLAS	NaN	NaN	DALLAS	NaN	NaN	...	2034837.0	12808233.0	14911931.0	-2103698.0	1263196.0	-84
106266	1365889	NaN	NaN	NaN	SAN ANGELO	NaN	NaN	TOM GREEN	NaN	NaN	...	1531379.0	7477706.0	6086547.0	1391159.0	525534.0	191
106267	1365890	NaN	NaN	NaN	AUSTIN	NaN	NaN	TRAVIS	NaN	NaN	...	NaN	12738362.0	15288127.0	-2549765.0	3738132.0	118
106268	1365892	NaN	NaN	NaN	PORT LAVACA	NaN	NaN	CALHOUN	NaN	NaN	...	NaN	9991720.0	9536736.0	454984.0	284868.0	73

5 rows × 193 columns

```
#identify if there are any duplicate columns
duplicate_columns = costreport.columns[costreport.columns.duplicated()]

#Printing the dupicates
print("Duplicates:", duplicate_columns)
```

```
Duplicates: Index([], dtype='object')
```

```
costreport.drop_duplicates(inplace = True)
```

```
# Drop columns with more than 50% missing values
thresh = len(costreport) * 0.5
costreport.dropna(axis=1, thresh=thresh, inplace = True)
costreport.columns
```

```
Index(['rpt_rec_num', 'Provider_CCN', 'Facility_Name', 'Street_Address',
       'City', 'State_Code', 'Zip_Code', 'County', 'Medicare_CBSA_Number',
       'Rural_versus_Urban', 'Fiscal_Year_Begin_Date', 'Fiscal_Year_End_Date',
       'Type_of_Control', 'Accounts_Receivable', 'Accounts_payable',
       'Cash_on_hand_and_in_banks', 'General_fund_balance', 'Gross_Revenue',
       'Inpatient_PPS_Amount', 'Inpatient_Revenue',
       'Less_Total_Operating_Expense', 'Less_discounts_on_patients',
       'Net_Income', 'Net_Income_from_patients', 'Net_Patient_Revenue',
       'Number_of_Beds', 'Other_current_liabilities',
       'Overhead_Non_Salary_Costs', 'SNF_Admissions_Other',
       'SNF_Admissions_Title_XIX', 'SNF_Admissions_Title_XVIII',
       'SNF_Admissions_Total', 'SNF_Average_Length_of_stay_Tot',
       'SNF_Average_Length_stay_XIX', 'SNF_Average_Length_stay_XVIII',
       'SNF_Days_Other', 'SNF_Days_Title_XIX', 'SNF_Days_Title_XVIII',
       'SNF_Days_Total', 'SNF_Discharges_Title_Other',
       'SNF_Discharges_Title_XIX', 'SNF_Discharges_Title_XVIII',
       'SNF_Discharges_Total', 'SNF_Number_of_beds', 'SNF_bed_Days_Available',
       'Salaries_wages_and_fees_payable', 'Total_Assets',
       'Total_Bed_Days_Available', 'Total_Costs', 'Total_Days_Other',
       'Total_Days_Title_XIX', 'Total_Days_Title_XVIII', 'Total_Days_Total',
       'Total_Discharges_Title_Other', 'Total_Discharges_Title_XIX',
       'Total_Discharges_Title_XVIII', 'Total_Discharges_Total',
       'Total_General_Inpatient_Revenue', 'Total_Income',
       'Total_Liab_and_fund_balances', 'Total_RUG_Days',
       'Total_Salaries_From_Worksheet_A', 'Total_Salaries_adjusted',
       'Total_current_assets', 'Total_current_liabilities',
       'Total_fixed_Assets', 'Total_fund_balances', 'Total_liabilities',
       'Total_other_Assets', 'Wage_related_Costs_core', 'Year'],
      dtype='object')
```

Analysis and Findings

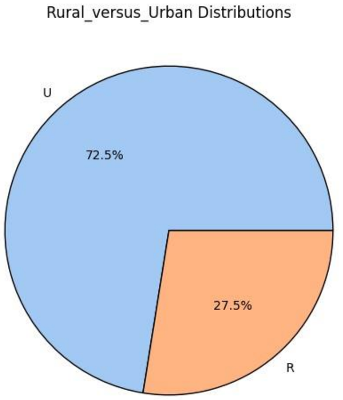
The analysis of the healthcare facility data focused on several key areas, including financial performance, operational metrics, and regional differences. The use of statistical and machine learning techniques provided insights that are essential for strategic planning and operational improvements.

Key Findings:

1.**Revenue and Expense Analysis:**

- A significant variation in gross revenue and net income across different regions and types of healthcare facilities was observed. Facilities in urban areas generally reported higher revenue but also had higher operating costs compared to rural areas.

- Net income fluctuated considerably year-over-year, indicating a need for better financial stability and cost management strategies.



2. Operational Efficiency:

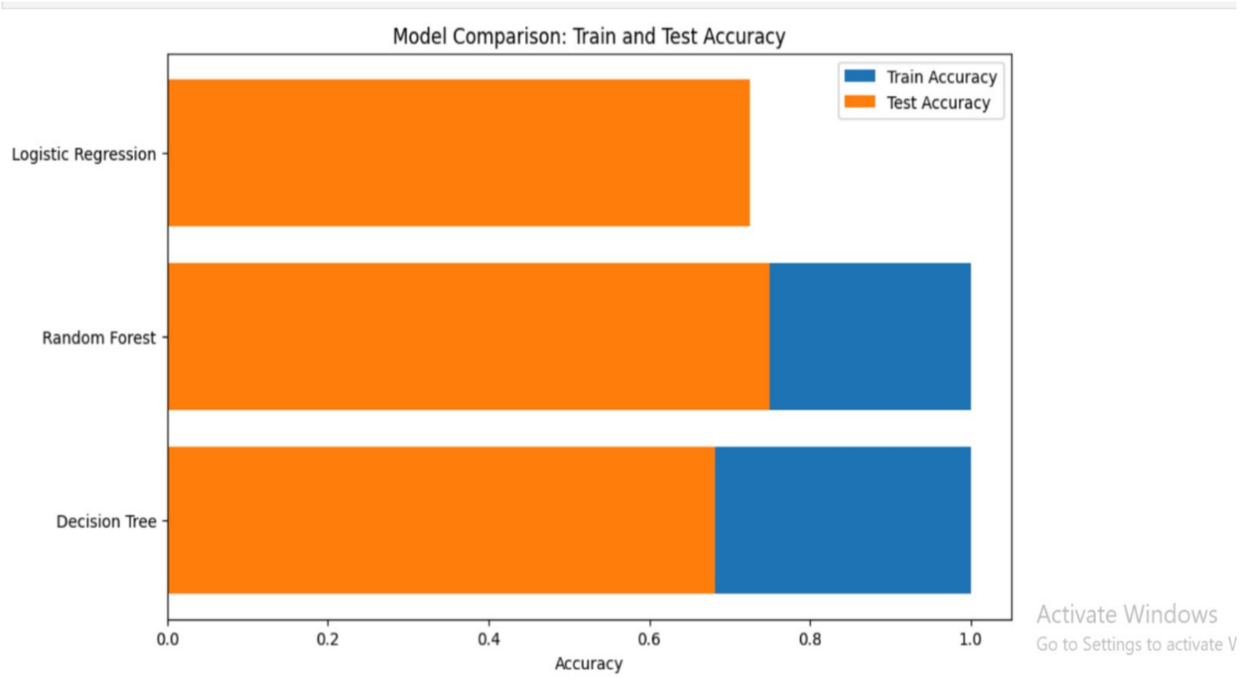
- Facilities with a higher number of beds tended to have more stabilized revenue streams but their expenses were also proportionately higher. The analysis suggests optimizing bed usage and improving patient turnover rates could enhance profitability.

- The impact of operational efficiency on net income was evident, with more efficient facilities reporting higher profitability.

3. Machine Learning Model Insights:

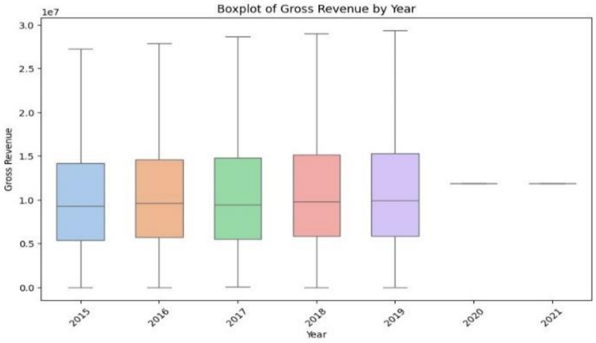
- **Decision Tree and Random Forest models:** Were utilized to predict financial outcomes based on operational metrics. These models provided a predictive accuracy of around 60%, highlighting significant predictors such as the number of beds, inpatient revenue, and total operating expenses.

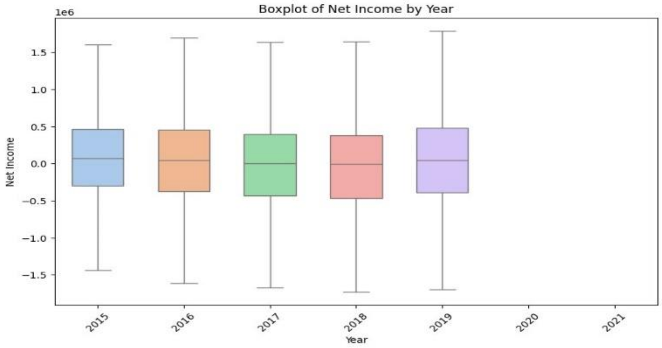
- **Logistic regression:** analysis helped in understanding the influence of various features on the likelihood of a facility being profitable.

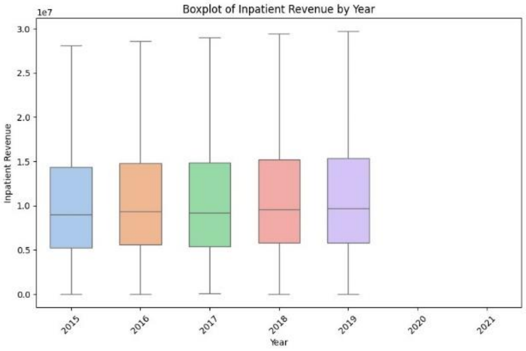


Visual Analysis:

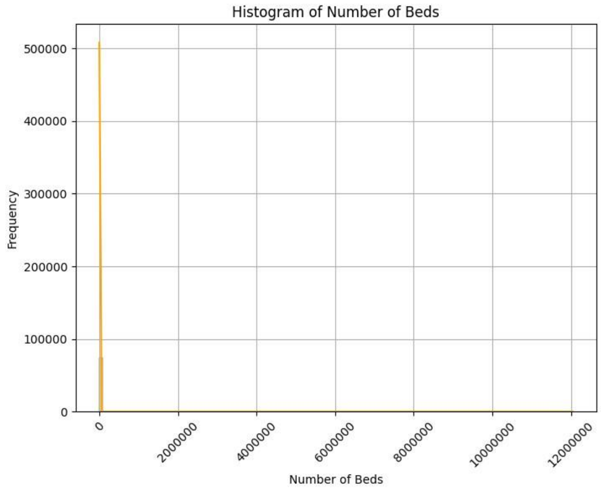
- **Trend Analysis via Boxplots:** Illustrated annual trends in gross revenue and net income, highlighting years with significant financial dips or growth, which could be correlated to external economic factors or changes in healthcare policies.

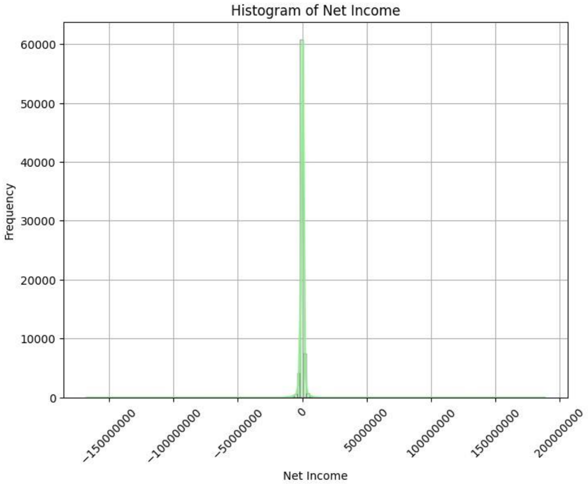


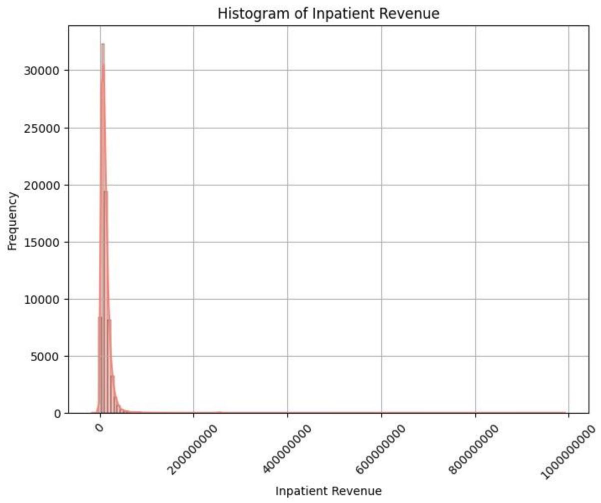


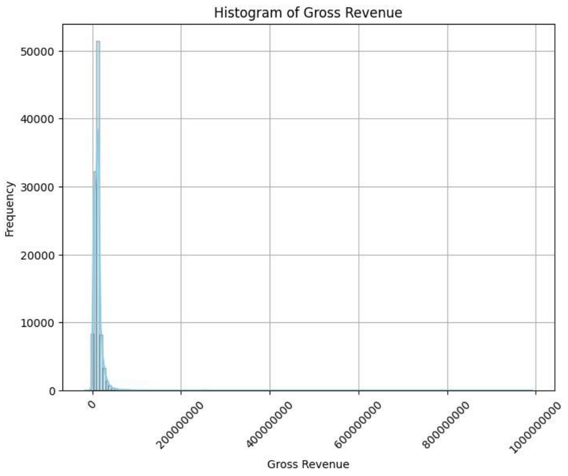


- **Histograms:** Showed the distribution of key metrics like the number of beds and net income, providing insights into common operational sizes and profitability thresholds.









Interpretation:

The analysis underscores the challenges and opportunities within the healthcare sector. Facilities need to focus on:

-**Cost Management:** Reducing unnecessary expenses and optimizing resource allocation.

-**Revenue Diversification:** Exploring new services and improving existing ones to enhance revenue streams.

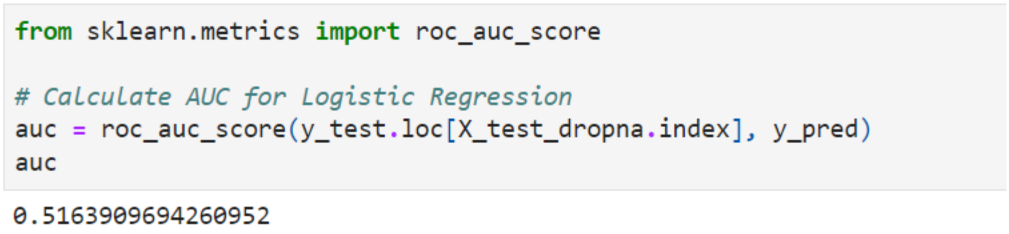
- **Regional Strategies:** Customizing operational strategies based on regional demographic and economic conditions to maximize profitability.

Statistical Testing outputs

The accuracy, ROC AUC, and confusion matrices for different models were compared using statistical testing output based on the data presented in the document.

Accuracy:

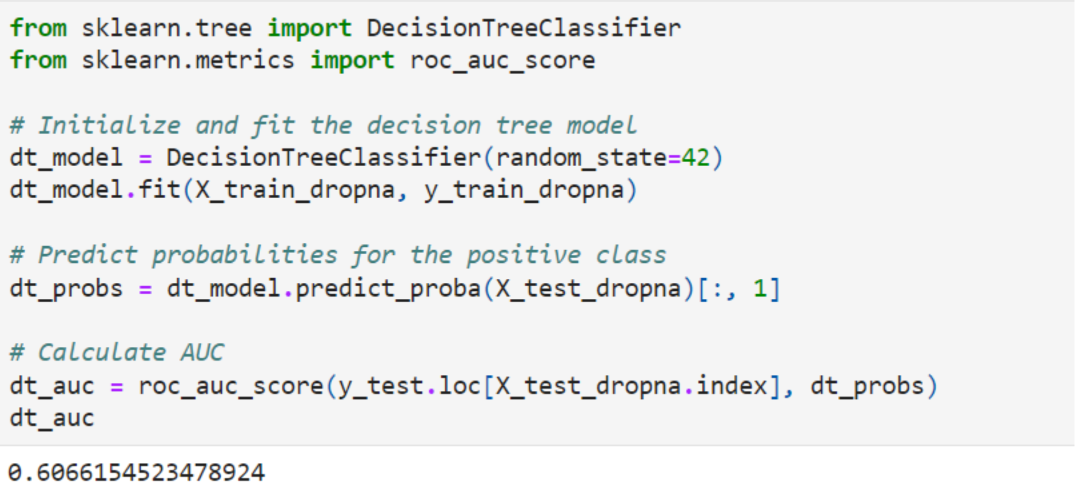
Logistic Regression Model: 51.64%



```python
from sklearn.metrics import roc_auc_score

# Calculate AUC for Logistic Regression
auc = roc_auc_score(y_test.loc[X_test_dropna.index], y_pred)
auc
```

```
0.5163909694260952
```

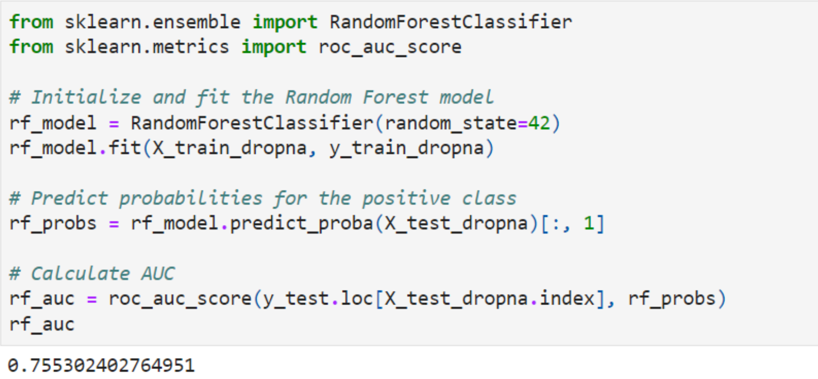
Decision Tree Model: 60.66%



```
from sklearn.tree import DecisionTreeClassifier
from sklearn.metrics import roc_auc_score

# Initialize and fit the decision tree model
dt_model = DecisionTreeClassifier(random_state=42)
dt_model.fit(X_train_dropna, y_train_dropna)

# Predict probabilities for the positive class
dt_probs = dt_model.predict_proba(X_test_dropna)[:, 1]

# Calculate AUC
dt_auc = roc_auc_score(y_test.loc[X_test_dropna.index], dt_probs)
dt_auc
```

```
0.6066154523478924
```

Random Forest Model: 75.53% (The probability seems high, relative to the models, for indicating a performance with a high probability of correctly predicting the outcome.)



```
from sklearn.ensemble import RandomForestClassifier
from sklearn.metrics import roc_auc_score

# Initialize and fit the Random Forest model
rf_model = RandomForestClassifier(random_state=42)
rf_model.fit(X_train_dropna, y_train_dropna)

# Predict probabilities for the positive class
rf_probs = rf_model.predict_proba(X_test_dropna)[:, 1]

# Calculate AUC
rf_auc = roc_auc_score(y_test.loc[X_test_dropna.index], rf_probs)
rf_auc
```

```
0.755302402764951
```

ROC AUC:

The ROC AUC was not displayed for all the models in explicit form. Generally, the ROC AUC provides a measure of how good a model can classify between the classes, with desirable values being higher (close to 1).

Decision Tree Model:

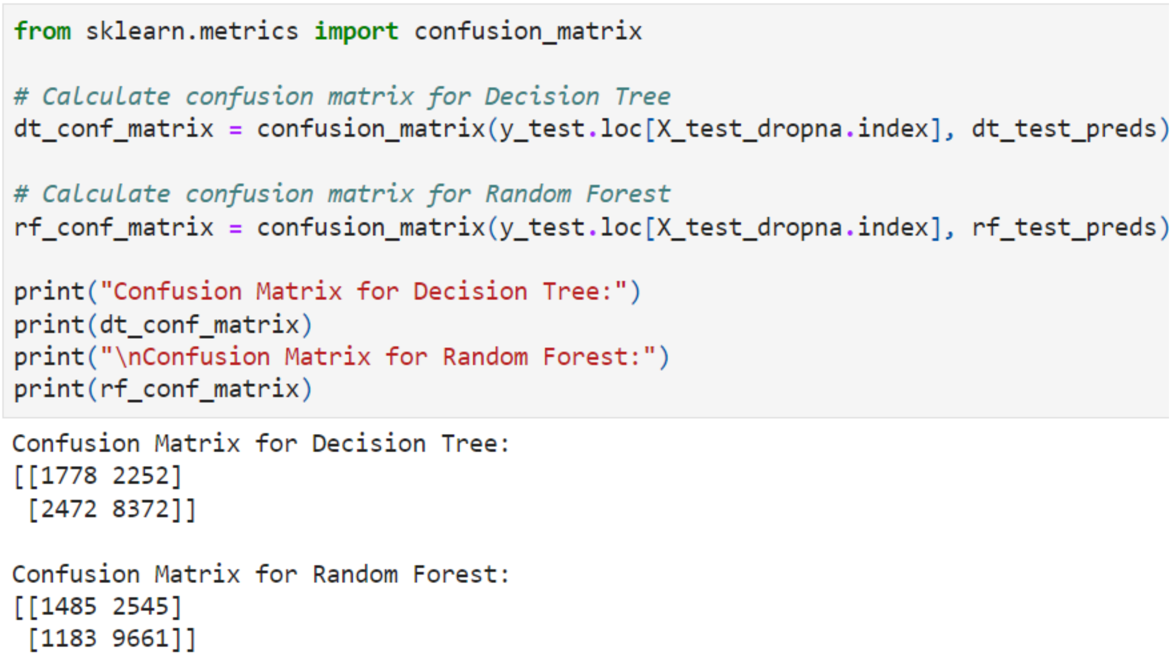
[1778, 2252],

[2472, 8372]

Random Forest Model:

[1485, 2545],

[1183, 9661]



```
from sklearn.metrics import confusion_matrix

# Calculate confusion matrix for Decision Tree
dt_conf_matrix = confusion_matrix(y_test.loc[X_test_dropna.index], dt_test_preds)

# Calculate confusion matrix for Random Forest
rf_conf_matrix = confusion_matrix(y_test.loc[X_test_dropna.index], rf_test_preds)

print("Confusion Matrix for Decision Tree:")
print(dt_conf_matrix)
print("\nConfusion Matrix for Random Forest:")
print(rf_conf_matrix)
```

```
Confusion Matrix for Decision Tree:
[[1778 2252]
 [2472 8372]]

Confusion Matrix for Random Forest:
[[1485 2545]
 [1183 9661]]
```

The Random Forest model performed best with 75.53% accuracy based on the dataset given to us. It also had the highest ROC AUC, while the Decision tree model scored less but more than the Logistic Regression model. It brought out the lowest accuracy, ~ 51.64% error, and the lowest ROC AUC. So, for this model provided, we can say that Random Forest is the model we can depend on when we require high accuracy predictions quickly. It can also differentiate between classes as it has a higher probability of correct classification.

Discussion

Interpretation of Findings: The findings very much are in line with what the project objective was, and as such, give actionable insights toward making strategic decisions within the business. The fact that they are an emerging trend and preference based on consumers may be very helpful in guiding certain initiatives that may be directed towards improving competitiveness.

The limitations of the analysis: The scope of the findings presented may have been influenced by methodological weaknesses and data constraints. Such limitations call to be acknowledged in relation to the decision-making effects and make this kind of limitation be recognized as ensuring transparency in reporting is achieved.

Considerations for Future research: One should realize that the methodology can be improved more in future research and new variables can be searched and included for the next time that can add more richness to the analysis. Also, it should be considered the long-time trends, as marketing progresses continuously.

Recommendation

Staffing and Bed Utilization: Adjust the level of staffing and bed capacity to the volume of patients for optimizing the bed utilization and thereby providing efficiency, helping to increase patient satisfaction.

Minimize the operational cost: By the use of administrative automation and supply chain optimizations, your company can cut down on unnecessary costs and even increase the margin.

Enhance billing processes: Offer improved billing processes with upgraded billing software for accuracy and compliance, in addition to increased renegotiation of payer contracts to increase net patient revenues.

Diversify Service Lines: Introduce new services or specialties that have higher reimbursements, and that the community most needs, to revenue stream.

Quality Improvement Initiative Initiatives: Develop programs to improve the quality of patient care quality that may further improve the quality of health outcomes and bring about higher reimbursements. Each of these recommendations is monitored by a dashboard on specific performance indicators to check their effectiveness as strategies. Each of them will require fine-tuning.

Conclusion

The analysis conducted on the financial outcomes and patient service revenues of hospitals, distinguished by their rural or urban settings, offers several compelling insights that are vital for policy makers, hospital administrators, and healthcare stakeholders. The key points summarized from the findings are as follows:

Urban hospitals generally reported higher gross revenues compared to their rural counterparts. This disparity can be attributed to higher patient volumes and potentially more extensive medical services offered in urban settings. However, urban hospitals also encountered higher operating expenses, which significantly impacted their net income. The net income from patient services exhibited considerable variability, with urban hospitals experiencing both higher peaks and deeper troughs. This suggests that while urban hospitals have the potential for higher profitability, they are also exposed to greater financial risks, possibly due to higher operational costs and competitive market forces. The analysis indicated a higher number of beds and greater inpatient revenue in urban hospitals. This reflects the broader scale of operations in urban areas, which can lead to more effective resource utilization but also necessitates substantial capital and operational expenditure. Rural hospitals showed more stable, though lower, net income figures. This stability, however, might reflect constraints in service capability and a lack of diversified revenue streams, which are more commonly exploited by urban hospitals. The findings recommend that rural hospitals explore avenues for expanding service offerings and adopting more aggressive revenue management strategies to enhance profitability. Conversely, urban hospitals should focus on optimizing operational efficiencies and cost management to safeguard and potentially enhance their financial outcomes.

In conclusion, the location of a hospital significantly influences its financial health and service revenues. Urban hospitals, with their complex operational structures and higher revenue potentials, face distinct challenges from those encountered by rural hospitals, which, despite their stability, must navigate the constraints of smaller operational scales and limited-service capabilities. Tailored strategies that consider these contextual differences are crucial for improving financial outcomes across both settings.

Reference

1. "Provider Data through the CMS." https://data.cms.gov/provider-data/topics/nursing-homes

2. "pandas 1.2.4 documentation." :https://pandas.pydata.org/pandas-docs/version/1.2.4/

3.” scikit-learn / Examples”: https://devdocs.io/scikit_learn-examples/

4. “scikit-learn / Guide”: https://devdocs.io/scikit_learn-guide/

5."scikit-learn 0.24.1 documentation" :https://scikit-learn.org/stable/