Net Promoter Score

Instructor: Giri Narasimhan
CAP 5768 Introduction to Data Science

Rafael Linarez
Florida International University
rlina018@fiu.edu

Rolando Diaz
Florida International University
rdiaz281@fiu.edu

Juan Ignacio Castro
Florida International University
jcast580@fiu.edu

Maria Paula Sanchez
Florida International University
panther@fiu.edu

Executive Summary

Enhancing Patient Satisfaction in Keralty Hospitals

This executive summary encapsulates key findings and strategic insights derived from a comprehensive analysis of the Net Promoter Score (NPS) within Keralty Hospitals. The study navigated through statistical analyses, identification of top variables, and decision tree modeling to unravel critical factors influencing patient satisfaction. Surprisingly, features related to patient information had no discernible impact on overall NPS, challenging conventional assumptions. The unexpected revelation that the state of the parking lot exhibited a greater correlation with NPS than patient age underscores the importance of environmental factors. Regional variations revealed unique top variables, such as "time waiting" and "Time of the appointment" in South Florida, emphasizing the need for geographically tailored strategies. Notably, the study demonstrated that accurate NPS predictions can be achieved with as few as five features, enhancing the efficiency of targeted interventions. The executive summary concludes with a call for future research to replicate and expand upon these findings, acknowledging the study's limitations and highlighting the potential for practical improvements in patient satisfaction within healthcare institutions.

TABLE OF CONTENTS

ABSTRACT

This paper delves into a comprehensive analysis of Net Promoter Score (NPS) data for Keralty Hospitals across four regions in Florida—North, South, Central, and West. Leveraging techniques such as correlation analysis and feature importance, we identify the top three and top five variables influencing NPS categories (Demoter, Passive, Promoter) within the healthcare context. Subsequently, employing decision tree models, we evaluate the predictive capabilities of different sets of variables, including all available data, weighted factors, and the top variables from correlation and feature importance analyses. The study aims to unveil the key determinants of patient satisfaction and to offer actionable insights for Keralty Hospitals to enhance their services. Findings from this analysis provide a strategic framework for healthcare providers to tailor their approach in distinct regions, fostering a patient-centric environment and driving improvements in overall healthcare experiences.

INTRODUCTION

The contemporary healthcare landscape is increasingly shaped by the imperative of patient satisfaction, a pivotal metric gauging the quality of services and overall healthcare experiences. Within this context, Net Promoter Score (NPS) stands out as a valuable instrument for assessing and quantifying patient sentiment. This paper embarks on an in-depth exploration of NPS data for Keralty Hospitals, strategically segmented across four distinct regions in Florida—North, South, Central, and West.

As the healthcare industry grapples with the evolving dynamics of patient expectations, understanding the factors influencing NPS becomes paramount. Keralty Hospitals, a significant player in the Florida healthcare ecosystem, presents a compelling case study for this investigation. The diverse demographics and regional variations inherent to Florida necessitate a nuanced examination, allowing for tailored insights that can drive targeted improvements in patient satisfaction.

Our analysis integrates various data science methodologies, including correlation analysis and feature importance techniques, to unearth the variables most salient to NPS outcomes. By focusing on the top three and top five variables identified through these methods, we aim to distill actionable intelligence for healthcare practitioners and administrators. Furthermore, employing decision tree models, we evaluate the predictive prowess of different variable subsets, offering a practical lens through which Keralty Hospitals can anticipate and address patient satisfaction challenges in real time.

This study not only contributes to the academic discourse surrounding NPS in healthcare but also provides a roadmap for Keralty Hospitals and similar institutions seeking to enhance patient experiences. As the healthcare landscape continues to evolve, the insights gleaned from this analysis are poised to catalyze positive changes, fostering a patient-centric paradigm that aligns with the ever-changing expectations of healthcare consumers.

MOTIVATION

This study is propelled by the recognition that deciphering the intricacies of Net Promoter Score (NPS) for Keralty Hospitals in Florida extends beyond academic curiosity; it holds the promise of yielding tangible benefits with profound implications for the healthcare sector.

Enhanced Operational Efficiency in Hospitals:

Understanding the variables influencing NPS allows for targeted improvements in operational processes. By identifying and addressing pain points in patient experiences, Keralty Hospitals can streamline workflows, reduce bottlenecks, and optimize resource allocation. The resulting enhancement in operational efficiency not only contributes to cost-effectiveness but also fosters an environment conducive to delivering prompt and quality healthcare services.

Competitive Advantage for Hospitals:

In an era where healthcare consumers actively seek personalized and satisfying experiences, a high Net Promoter Score can be a powerful differentiator for Keralty Hospitals. By tailoring services based on identified drivers of satisfaction, the hospitals can position themselves as leaders in patient-centric care, gaining a competitive edge in the increasingly competitive healthcare landscape.

Financial Benefits:

Satisfied patients are more likely to be loyal patrons, leading to increased retention and repeat visits. A positive NPS correlates with patient loyalty, translating into a stable and expanding patient base for Keralty Hospitals. Additionally, positive patient experiences can contribute to positive word-of-mouth referrals, potentially attracting new patients and bolstering the hospital's financial viability.

Regulatory Compliance:

Understanding the variables influencing NPS aligns with the broader healthcare regulatory landscape. By addressing factors that contribute to patient satisfaction, Keralty Hospitals can proactively meet regulatory standards related to patient care quality. This not only ensures compliance but also positions the hospitals as proactive entities committed to delivering care that exceeds regulatory expectations.

Positive Impact on Reputation:

Patient satisfaction, as reflected in NPS, directly correlates with the reputation of healthcare institutions. A positive NPS not only signals a commitment to quality care but also enhances the overall perception of Keralty Hospitals within the community. A sterling reputation, built on satisfied patient experiences, can strengthen relationships with stakeholders, including patients, staff, and partners, further solidifying the hospital's standing in the healthcare ecosystem.

In essence, the motivation for this study transcends theoretical exploration; it is rooted in the pragmatic realization that unraveling the factors influencing NPS holds the key to unlocking operational efficiencies, financial sustainability, regulatory compliance, and a sterling reputation for Keralty Hospitals in Florida.

DATA AND LIBRARIES

The dataset (https://github.com/JuanIgnacioCastro/NPS) comprises 30 columns and 1357 rows, representing the culmination of the 2022 Keralty Hospital patient satisfaction survey conducted across the four regions of Florida:

Patient Information (8 columns):

- ‘Patient ID’: a unique identifier, it is important to note that these values are anonymized numerical representations, meaning they are arbitrary numbers assigned to ensure patient privacy and confidentiality.
- ‘Market ID’: utilizes numerical values as unique identifiers for each county.
- Market: categorical variable that represents the geographic divisions by county within the state of Florida
- 'Region ID': utilizes numerical values as unique identifiers for each Region.
- 'Region': categorical variable that represents the geographic divisions by North, South, Central, and West Florida
- 'Visit Type': captures the nature or purpose of the patient's visit to Keralty Hospitals. It is a categorical variable with four possible values.
  - SPEC: seeking specialized care
  - PCP: Primary Care Physician
  - HOSP: hospital-related visits
  - UCC: Urgent Care Center
- 'Sex': two possible values: Male or female
- 'Age': a continuous numerical variable that captures the age of patients.

Date and Channel (5 columns)

- 'Survey Channel': the mode through which patients received the satisfaction survey. Email or SMS
- 'Visit Date': the specific date on which a patient had a visit to Keralty Hospitals
- 'Survey date': the specific date on which patients completed or submitted their satisfaction surveys.
- 'Month Visit': numeric value of the specific month in which a patient had their visit to Keralty Hospitals
- 'Quarter': numeric value of the quarter in which a patient had their visit to Keralty Hospitals

Operational Index (5 columns)

assesses various operational aspects of Keralty Hospitals as perceived by patients measured on a scale from 1(Lowest) to 5(Highest).

- 'Check-up appointment'
- 'Time waiting'
- 'Admin procedures'
- 'Hygiene and cleaning'
- 'Time of appointment'

Provider Service Index (8 columns)

assesses various Provider Services aspects of Keralty Hospitals as perceived by patients measured on a scale from 1(Lowest) to 5(Highest).

- 'Quality/experience dr.'
- 'Specialists available'
- 'Communication with dr'
- 'Exact diagnosis'
- 'Modern equipment'
- 'Friendly health care workers'
- 'Lab services'
- 'Availability of drugs'

Infrastructure (3 columns)

assesses various Infrastructure aspects of Keralty Hospitals as perceived by patients measured on a scale from 1(Lowest) to 5(Highest).

- 'Waiting rooms'
- 'Hospital rooms quality'
- 'Parking, playing rooms, cafes.'

NPS (1 column)

The Net Promoter Score is calculated based on a single survey question that typically asks respondents to rate, on a scale of 0 (Lowest) to 10 (Highest), how likely they are to recommend the service to others.

- 'NPS Category': Promoters (Score 9-10), Passives (Score 7-8), Detractors (Score 0-6).

This structured data organization facilitates a nuanced exploration of patient experiences, operational efficiency, service quality, and infrastructure, all contributing factors to the Net Promoter Score. The varied dimensions captured in the dataset provide a holistic view, allowing for detailed analysis and actionable insights for healthcare improvement initiatives.

Libraries:

The analysis leverages several Python libraries to perform data manipulation, visualization, statistical testing, and machine learning modeling. The key libraries employed in this study are:

- Pandas: Used for data manipulation and analysis, providing functionalities for data cleaning, exploration, and transformation.
- NumPy: Utilized for numerical operations and array manipulations, facilitating efficient calculations and statistical analyses.
- seaborn and Matplotlib: Employed for data visualization, enabling the creation of informative plots and charts to illustrate trends, distributions, and relationships within the data.
- SciPy. Stats: Utilized for statistical analyses, including hypothesis testing. Specific functions like Ttest_ind, F_oneway, Kruskal, and chi2_contingency is employed to assess statistical significance.
- Sklearn: A machine learning library used for preprocessing data and building predictive models. Functions like LabelEncoder and Train_test_split are applied to prepare the data for machine learning tasks.
- DecisionTreeClassifier: Implemented from Sklearn to build decision tree models for predicting NPS categories based on the identified features.
- Confusion matrix and Accuracy_score: Utilized from Sklearn to assess the performance of the machine learning models through metrics such as confusion matrices and accuracy scores.

These libraries collectively form a robust toolkit, enabling a comprehensive and data-driven exploration of the NPS dataset for Keralty Hospitals in Florida. The integration of statistical analyses and machine learning techniques allows for a nuanced understanding of the factors influencing patient satisfaction and provides actionable insights for healthcare practitioners and administrators.

DATA PREPARATION

The data preparation phase of this study was streamlined due to the absence of missing values in the survey data, rendering the process straightforward and focused on optimizing the dataset for subsequent analyses. The primary tasks involved in data preparation included encoding categorical values and structuring the data to facilitate correlation calculations and visualization techniques.

EXPERIMENTS

EXPLORATORY DATA ANALYSIS

The Exploratory Data Analysis process encompassed a diverse array of visualizations to uncover patterns and trends within the dataset, particularly focusing on the relationships between various features and the Net Promoter Score (NPS). This iterative process helps gain a preliminary understanding of potential influential factors.

- Correlation Heatmaps: Heatmaps were employed to visualize the correlation matrix, providing a comprehensive overview of the relationships between numerical variables and NPS. This facilitated the identification of strong correlations and potential multicollinearity among features.
- Box Plots: Box plots were utilized to visualize the distribution of NPS scores across different categories, such as 'Visit Type,' providing a clear depiction of the central tendency, spread, and potential outliers within each category.
- Line Charts: Line charts were used to create time series visualizations, showcasing the variation of NPS over time-based on 'Visit Date' or 'Survey Date.' This enabled the identification of any temporal trends in patient satisfaction.
- Histograms: explore the distribution of numerical variables like 'Age' in relation to NPS. These visualizations provided insights into the frequency distribution and central tendencies of key features.

These visualizations collectively provided an understanding of the dataset, allowing for the identification of influential factors and potential patterns influencing patient satisfaction. The insights gained from EDA, particularly through the diverse set of visualizations employed, informed the decision-making process for subsequent statistical analyses and machine learning modeling. Visual explorations served as a critical precursor to hypothesis formulation. They guided the selection of appropriate techniques for the in-depth analysis of patient satisfaction within Keralty Hospitals in Florida.

STATISTICAL ANALYSIS

In the pursuit of identifying influential factors that significantly affect the Net Promoter Score (NPS) within Keralty Hospitals, a rigorous statistical analysis was conducted. Various statistical methods were employed to assess the relationships between different features and NPS, focusing on identifying variables with strong correlations. The following methods were utilized:

- T-value, F-value, and P-value: Applied t-tests to assess the significance of mean differences between groups, particularly for categorical variables with two groups. Utilized analysis of variance (ANOVA) to evaluate group mean differences for categorical variables with more than two groups. Computed p-values to determine the statistical significance of these differences.
- Kruskal-Wallis H-test: Employed the Kruskal-Wallis H-test for non-parametric analysis to assess whether there were statistically significant differences in NPS scores across multiple groups.
- Chi-Square Test for Independence: Applied the Chi-Square test for independence to assess the association between categorical variables and the NPS category, discerning whether these variables influenced the likelihood of being a Promoter, Passive, or Detractor.

After conducting the statistical analyses, features that did not exhibit a strong correlation with the NPS were systematically identified and dropped from the dataset. This process was instrumental in refining the dataset to include only those features that demonstrated a statistically significant impact on patient satisfaction.

The culmination of the statistical analysis phase yielded a dataset that exclusively comprised features with a discernible impact on the NPS within the context of Keralty Hospitals. This refined dataset is a potent foundation for subsequent machine learning algorithms, such as decision trees or random forests. By excluding non-influential variables, the dataset is now optimized for modeling, enhancing the precision and interpretability of future analyses.

TOP FEATURES

To unravel the key determinants influencing the Net Promoter Score (NPS) within Keralty Hospitals, two distinct methodologies, correlation analysis and Random Forest feature importance, were employed to discern the top variables impacting patient satisfaction.

Correlation coefficients were computed to gauge the strength and direction of the relationship between each variable and the NPS. The top 3 and top 5 variables were identified based on the magnitude of their correlation values.

A random forest was deployed to assess the importance of each variable in predicting NPS. 80% of the data was used to train, and 20% to test. Feature importance scores provided a quantitative measure of the impact of each feature on the overall predictive performance. We extracted the top 3 and top 5 features based on their importance scores.

The combination of correlation analysis and Random Forest feature importance unveils a comprehensive understanding of the factors shaping patient satisfaction. Identifying unique and shared top variables provides nuanced insights, equipping Keralty Hospitals in Florida with actionable information to enhance patient experiences and elevate overall NPS.

DECISION TREES

The Decision Tree analysis served as a pivotal step in predicting the Net Promoter Score (NPS) category and understanding the most influential features contributing to patient satisfaction within Keralty Hospitals. Five distinct decision trees were constructed, each employing a different set of features to explore the efficiency of predictions and identify key factors impacting NPS.

- All Variables: Utilizing the entire set of variables that showed a strong correlation.
- Top 3 in Correlation: Focusing on the top three variables identified through correlation analysis.
- Top 3 in Random Forest: The top three variables derived from Random Forest feature importance.
- Top 5 in Correlation: Focusing on the top five variables identified through correlation analysis.
- Top 5 in Random Forest: The top five variables derived from Random Forest feature importance.

The overarching objective of constructing these decision trees was to identify the least set of features that could effectively predict NPS categories.

The strategy was rooted in the idea of optimizing NPS predictions with as few features as possible, aligning with the goal of enhancing patient satisfaction while minimizing resource expenditures.

RESULTS

The statistical analysis undertaken in this study utilized various methods, including T-value, F-value, P-value, ANOVA tables, Kruskal-Wallis H-test, and Chi-Square Test for Independence. Features that did not demonstrate a strong correlation were meticulously identified and subsequently dropped from the dataset. Here are some examples based on the West Florida Region.

MULTIPLE CATEGORICAL vs. CONTINOUS

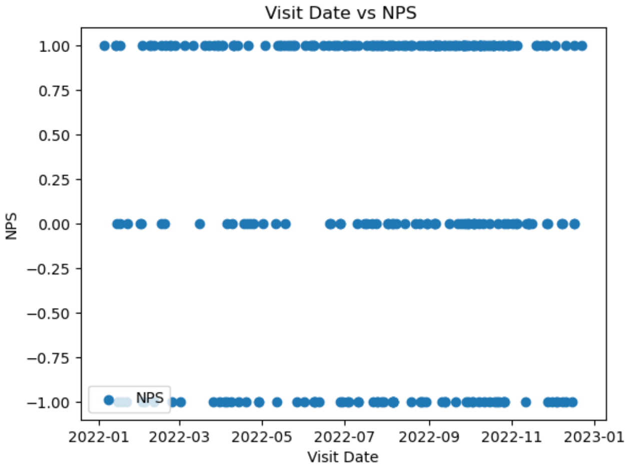
Python Script

```
# Replace these with your actual column names
date_column = df['Visit Date']
categorical_column = df['NPS_bin']
# Convert the date column to numerical values
```

```
numeric_date_column = date_column.dt.dayofyear
# Combine the numeric date column and the categorical column
combined_data = pd.DataFrame({'NumericDate': numeric_date_column, 'Category': categorical_column})
# Perform the Kruskal-Wallis H-test
h_statistic, p_value = kruskal(*[group['NumericDate'] for name, group in combined_data.groupby('Category')])
# Output the results
print(f"Kruskal-Wallis H-statistic: {h_statistic:.4f}")
print(f"P-value: {p_value:.4f}")
# Check for significance
alpha = 0.05
if p_value < alpha:
   print("There is a significant correlation between the date column and the categorical column.")
else:
   print("There is no significant correlation between the date column and the categorical column.")
# Plotting
plt.scatter(df['Visit Date'], df['NPS_bin'], label='NPS')
plt.xlabel('Visit Date')
plt.ylabel('NPS')
plt.title('Visit Date vs NPS ')
plt.legend()
plt.show()
```

Result

```
Kruskal-Wallis H-statistic: 1.8051
P-value: 0.4055
There is no significant correlation between the date column
and the categorical column.
```



MULTIPLE CATEGORICAL vs. MULTIPLE CATEGORICAL

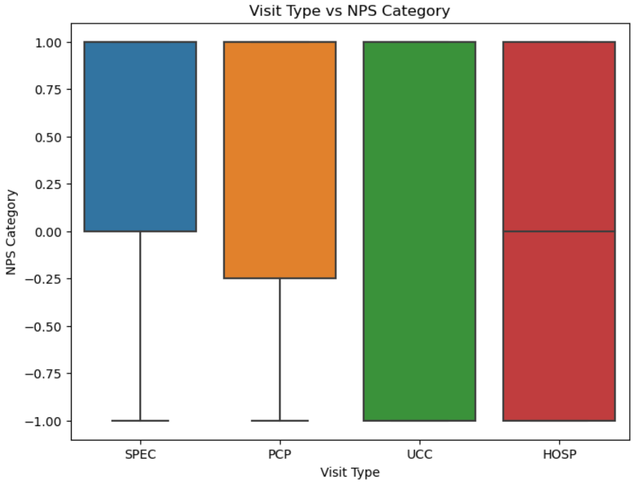
Python Script

```
def chi_square_test_and_boxplot(df, categorical_column1, categorical_column2, alpha=0.05):
   # Chi-square test
   contingency_table = pd.crosstab(df[categorical_column1], df[categorical_column2])
   chi2_stat, p_value, _, _ = chi2_contingency(contingency_table)
   print(f"Chi-square Statistic: {chi2_stat:.4f}")
   print(f"P-value: {p_value:.4f}")
   if p_value < alpha:
      print("There is a significant association between the two categorical variables.")
   else:
      print("There is no significant association between the two ctegorical variables.")
   # Boxplot
   plt.figure(figsize=(8, 6))
   sns.boxplot(x=categorical_column1, y=categorical_column2, data=df)
   plt.title(f'{categorical_column1} vs NPS Category')
   plt.xlabel(categorical_column1)
   plt.ylabel('NPS Category')
   plt.show()

chi_square_test_and_boxplot(df,'Visit Type','NPS_bin')
```

Result

```
Chi-square Statistic: 5.0313
P-value: 0.5398
There is no significant association between the two
categorical variables.
```



MULTIPLE CATEGORICAL vs. BINARY CATEGORICAL

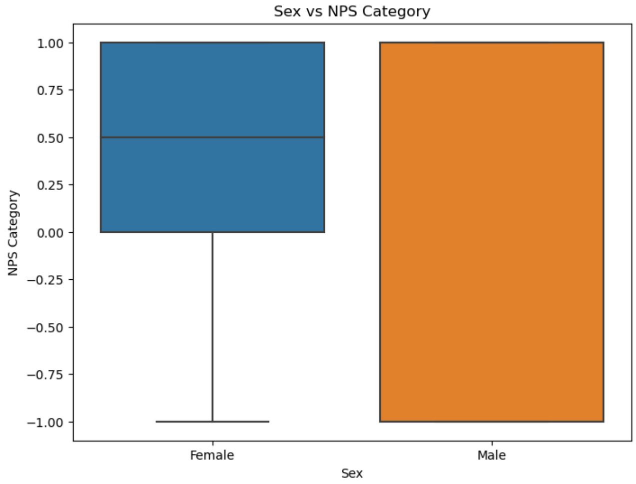
Python Script

```
def chi_square_test_and_boxplot(df, categorical_column1, categorical_column2, alpha=0.05):
   # Chi-square test
   contingency_table = pd.crosstab(df[categorical_column1], df[categorical_column2])
   chi2_stat, p_value, _, _ = chi2_contingency(contingency_table)
   print(f"Chi-square Statistic: {chi2_stat:.4f}")
   print(f"P-value: {p_value:.4f}")
   if p_value < alpha:
      print("There is a significant association between the two categorical variables.")
   else:
      print("There is no significant association between the two ctegorical variables.")
   # Boxplot
   plt.figure(figsize=(8, 6))
   sns.boxplot(x=categorical_column1, y=categorical_column2, data=df)
   plt.title(f'{categorical_column1} vs NPS Category')
   plt.xlabel(categorical_column1)
   plt.ylabel('NPS Category')
   plt.show()

chi_square_test_and_boxplot(df,'Sex','NPS_bin')
```

Result

```
Chi-square Statistic: 0.3998
P-value: 0.8188
There is no significant association between the two
categorical variables.
```



POST ANALYSIS IF CORRELATION IS STRONG

```
def chi_square_test_and_posthoc(df, categorical_column1, categorical_column2, alpha=0.05):
   # Create a contingency table
   contingency_table = pd.crosstab(df[categorical_column1], df[categorical_column2])
   # Perform the chi-square test
   chi2_stat, p_value, _, _ = chi2_contingency(contingency_table)
   # Output the results
   print(f"Chi-square Statistic: {chi2_stat:.4f}")
   print(f"P-value: {p_value:.4f}")
   # Check for significance at the specified alpha level
   if p_value < alpha:
      print("There is a significant association between the two categorical variables.")
      # Post hoc analysis using standardized residuals
      residuals = (contingency_table - chi2_stat / contingency_table.sum().sum()).values
      standardized_residuals = residuals / residuals.std()
      # Display the standardized residuals
      print("\nStandardized Residuals:")
      print(pd.DataFrame(standardized_residuals, index=contingency_table.index, columns=contingency_table.columns))
   else:
      print("There is no significant association between the two categorical variables.")
   plt.title(f'{categorical_column1} vs NPS Category')
   plt.xlabel(categorical_column1)
   plt.ylabel('NPS Category')
   plt.show()

chi_square_test_and_posthoc(df,'Quality/experience dr.','NPS_bin')
```

Result

```
Chi-square Statistic: 134.1234
P-value: 0.0000
There is a significant association between the two
categorical variables.

Standardized Residuals:
NPS_bin                 -1        0        1
Quality/experience dr.
1                 2.810814  1.505287  0.994429
2                 0.483571  0.937667  0.256523
3                 0.142999  0.824143  0.937667
4                 0.370047  0.483571  3.435196
5                 0.029475  0.029475  2.186431
```

Interestingly, the variables related to patient information(like Market, Visit Type, Sex, and Age) did not correlate strongly with the NPS score and were consequently excluded from the dataset. Instead, the columns associated with the Operational Index, Provider Service Index, and Infrastructure demonstrated strong correlations with NPS. This strategic refinement enhances the dataset's utility, ensuring that subsequent analyses and predictive modeling can be conducted with a focused set of features directly tied to patient satisfaction.

The culmination of this phase yielded a refined dataset comprising only those features that significantly influenced the NPS score, providing a foundation for future machine learning algorithms such as decision trees and random forests.

Then, the top variables were found for North Florida, Central Florida, South Florida, and Central Florida. Utilizing both correlation analyses and Random Forest feature importance assessments with a random seed of 42.

Here is a code example for the West Florida Region

Python Script

```
#segregate by region if necessary
df1 = df[df['Region ID']==4]
#get correlation values
cor = df1.corr()
#sort and print values
nps_cor = cor['NPS_bin'].sort_values(ascending= False)
nps_correlations = nps_cor[1:]
print('Correlation of variables to  NPS:\n', nps_correlations)
print("\n")
#libraries used for random forrest
from sklearn.model_selection import train_test_split
from sklearn.ensemble import RandomForestRegressor
#separete the train features and the variable to predict
X = df1.drop('NPS_bin', axis =1)
y = df1['NPS_bin']
#80% is used to test 20% is to test
X_train , X_test , y_train , y_test = train_test_split(X,y, test_size = 0.2 , random_state= 42)
model = RandomForestRegressor(random_state= 42)
#fit model
model.fit(X_train, y_train)

#sort and print values
features = pd.Series(model.feature_importances_, index = X.columns).sort_values(ascending= False)
print("Varibles that are important :\n", features )
```

Here are the overall results

FLORIDA

Correlation	Ranking	Feat Importance
Communication with dr	1	Communication with dr
friendly health care workers	2	lab services
Exact diagnosis	3	friendly health care workers
Quality /experience dr	4	waiting rooms
lab services	5	hospital rooms quality

SOUTH FLORIDA

Correlation	Ranking	Feat Importance
Exact diagnosis	1	Exact diagnosis
friendly health care workers	2	Time of appointment
Time waiting	3	friendly health care workers
Communication with dr	4	Time waiting
Time of appointment	5	lab services

CENTRAL FLORIDA

Correlation	Ranking	Feat Importance
Communication with dr	1	Communication with dr
friendly health care workers	2	friendly health care workers
Quality/ experience dr.	3	hospital rooms quality
Exact diagnosis	4	avaliablity of drugs
Specialists avaliable	5	lab services

NORTH FLORIDA

Correlation	Ranking	Feat Importance
friendly health care workers	1	friendly health care workers
Communication with dr	2	Time waiting
Quality/ experience dr.	3	Communication with dr
Time waiting	4	hospital rooms quality
Exact diagnosis	5	waiting rooms

WEST FLORIDA

Correlation	Ranking	Feat Importance
Quality/ experience dr.	1	Quality/ experience dr.
lab services	2	Communication with dr
Exact diagnosis	3	lab services
friendly health care workers	4	Exact diagnosis
Communication with dr	5	friendly health care workers

Then the five decision trees were employed to identify key factors contributing to patient satisfaction. Notably, this approach aimed to discern whether a subset of features could effectively predict NPS categories, aligning with the overarching objective of maximizing predictive accuracy while minimizing the number of utilized features.

The decision trees were designed to consider different combinations of variables:

- All Variables: Utilizing the entire set of variables that showed a strong correlation
- Top 3 in Correlation: Focusing on the top three variables identified through correlation analysis
- Top 3 in Random Forest: The top three variables derived from Random Forest feature importance
- Top 5 in Correlation: Focusing on the top five variables identified through correlation analysis
- Top 5 in Random Forest: The top five variables derived from Random Forest feature importance

These decision tree analyses not only facilitated accurate predictions of NPS categories but also identified key features that significantly influenced patient satisfaction. The outcomes of these analyses contribute to a strategic understanding of the determinants of patient satisfaction within Keralty Hospitals, guiding targeted interventions and enhancements for an elevated overall patient experience.

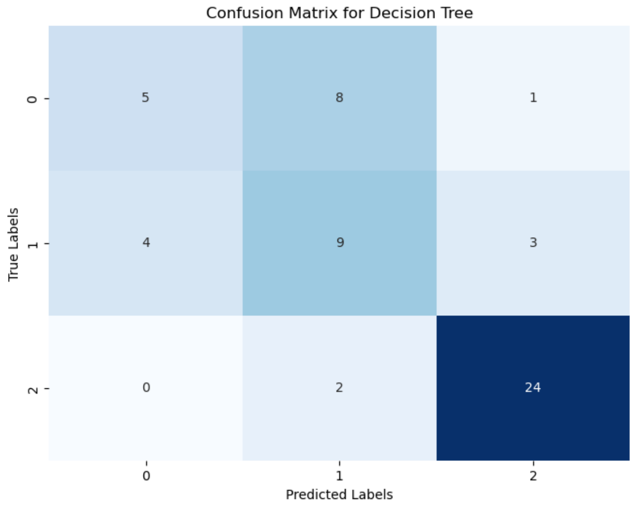
Here are examples for the West Florida region using a set seed of 42

DECISION TREE WITH ALL VARIABLES THAT SHOW SIGNIFICANCE DIFFERENCE

```
# Set a random seed for reproducibility
random_seed = 42
X = df.drop('NPS_bin', axis=1)
y = df['NPS_bin']
X_train, X_test, y_train, y_test = train_test_split(X, y,
test_size=0.2, random_state=random_seed)
# Create a DecisionTreeClassifier
clf = DecisionTreeClassifier(random_state=random_seed)
# Train the model on the training data
clf.fit(X_train, y_train)
y_pred = clf.predict(X_test)
accuracy = accuracy_score(y_test, y_pred)
print(f"Accuracy: {accuracy}")
# Create a confusion matrix
conf_matrix = confusion_matrix(y_test, y_pred)
# Display the confusion matrix using a heatmap
plt.figure(figsize=(8, 6))
sns.heatmap(conf_matrix, annot=True, fmt="d", cmap="Blues",
cbar=False)
plt.title('Confusion Matrix for Decision Tree')
plt.xlabel('Predicted Labels')
plt.ylabel('True Labels')
plt.show()
```

Result

Accuracy: 0.6785714285714286



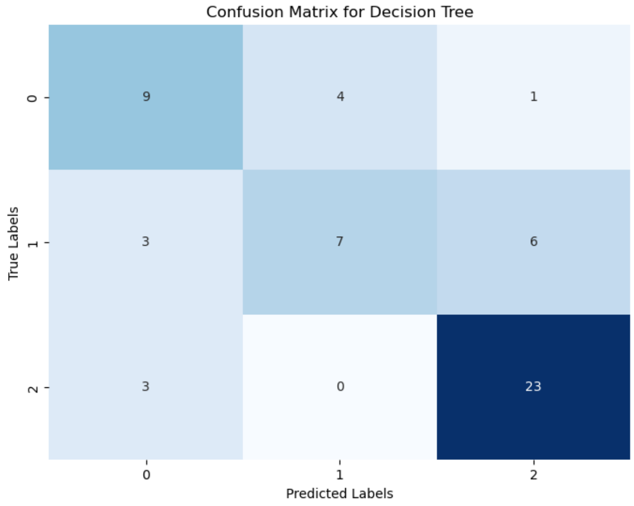
DECISION TREE WITH ALL TOP 3 IN CORRELATION

```
random_seed=42
X = df[['Quality/experience dr.','lab services','Exact diagnosis']]
y = df['NPS_bin']

X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=random_seed)
# Create a DecisionTreeClassifier
clf = DecisionTreeClassifier(random_state=random_seed)
# Train the model on the training data
clf.fit(X_train, y_train)
y_pred = clf.predict(X_test)
accuracy = accuracy_score(y_test, y_pred)
print(f"Accuracy: {accuracy}")
# Create a confusion matrix
conf_matrix = confusion_matrix(y_test, y_pred)
# Display the confusion matrix using a heatmap
plt.figure(figsize=(8, 6))
sns.heatmap(conf_matrix, annot=True, fmt="d", cmap="Blues", cbar=False)
plt.title('Confusion Matrix for Decision Tree')
plt.xlabel('Predicted Labels')
plt.ylabel('True Labels')
plt.show()
```

Result

Accuracy: 0.6964285714285714

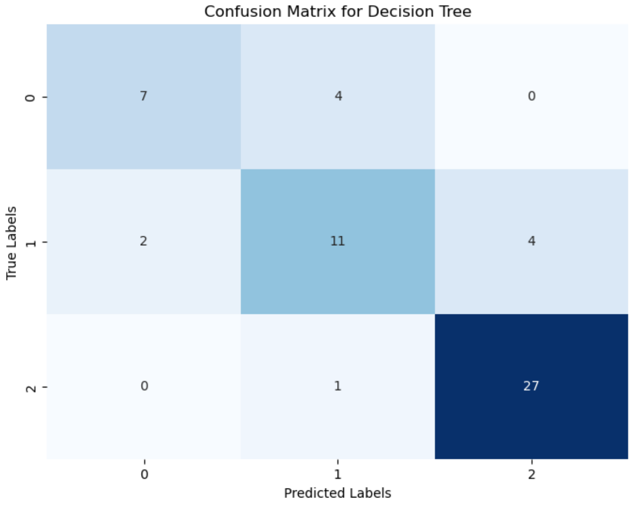


DECISION TREE WITH ALL TOP 3 IN RANDOM F

```
random_seed=42
X = df[['Quality/experience dr.','Communication with dr','lab services']]  # Features
y = df['NPS_bin']  # Target variable
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=random_seed)
# Create a DecisionTreeClassifier
clf = DecisionTreeClassifier(random_state=random_seed)
# Train the model on the training data
clf.fit(X_train, y_train)
y_pred = clf.predict(X_test)
accuracy = accuracy_score(y_test, y_pred)
print(f"Accuracy: {accuracy}")
# Create a confusion matrix
conf_matrix = confusion_matrix(y_test, y_pred)
# Display the confusion matrix using a heatmap
plt.figure(figsize=(8, 6))
sns.heatmap(conf_matrix, annot=True, fmt="d", cmap="Blues", cbar=False)
plt.title('Confusion Matrix for Decision Tree')
plt.xlabel('Predicted Labels')
plt.ylabel('True Labels')
plt.show()
```

Result

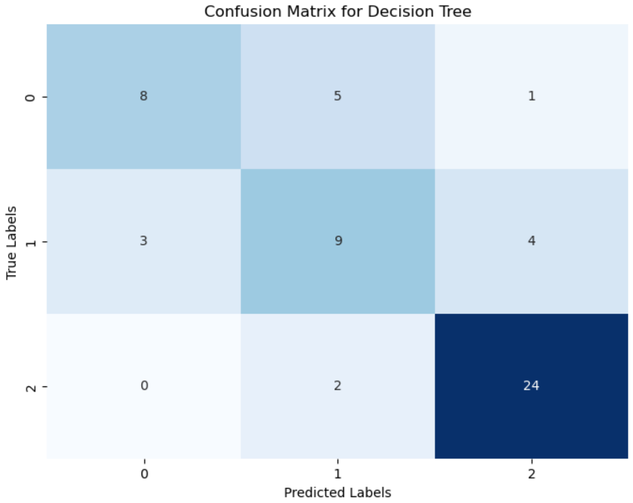
Accuracy: 0.8035714285714286



SAME DECISION TREE WITH TOP 5 IN CORRELATION

```
random_state=42
X = df[['Quality/experience dr.','lab services', 'Exact diagnosis','friendly health care workers','Communication with dr' ]]  # Features
y = df['NPS_bin']  # Target variable
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=random_seed)
# Create a DecisionTreeClassifier
clf = DecisionTreeClassifier(random_state=random_seed)
# Train the model on the training data
clf.fit(X_train, y_train)
y_pred = clf.predict(X_test)
accuracy = accuracy_score(y_test, y_pred)
print(f"Accuracy: {accuracy}")
# Create a confusion matrix
conf_matrix = confusion_matrix(y_test, y_pred)
# Display the confusion matrix using a heatmap
plt.figure(figsize=(8, 6))
sns.heatmap(conf_matrix, annot=True, fmt="d", cmap="Blues", cbar=False)
plt.title('Confusion Matrix for Decision Tree')
plt.xlabel('Predicted Labels')
plt.ylabel('True Labels')
plt.show()
```

Result
Accuracy: 0.7321428571428571



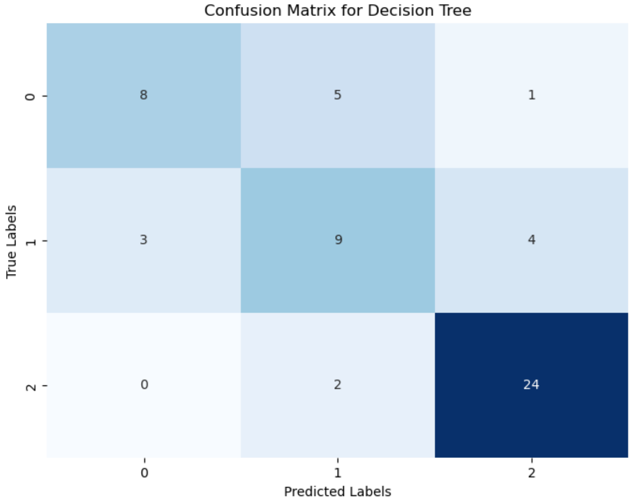
SAME DECISION TREE WITH TOP 5 IN CORRELATION

(In the example of West Florida Region using a set seed of 42, there is the coincidence that the top 5 in correlation are the same top 5 in feature importance that's why they have the same accuracy score)

```
random_state=42
X = df[['Quality/experience dr.','Communication with dr','lab services', 'Exact diagnosis','friendly health care workers' ]]  # Features
y = df['NPS_bin']  # Target variable
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=random_seed)
# Create a DecisionTreeClassifier
clf = DecisionTreeClassifier(random_state=random_seed)
# Train the model on the training data
clf.fit(X_train, y_train)
y_pred = clf.predict(X_test)
accuracy = accuracy_score(y_test, y_pred)
print(f"Accuracy: {accuracy}")
# Create a confusion matrix
conf_matrix = confusion_matrix(y_test, y_pred)
# Display the confusion matrix using a heatmap
plt.figure(figsize=(8, 6))
sns.heatmap(conf_matrix, annot=True, fmt="d", cmap="Blues", cbar=False)
plt.title('Confusion Matrix for Decision Tree')
plt.xlabel('Predicted Labels')
plt.ylabel('True Labels')
plt.show()
```

Result

Accuracy: 0.7321428571428571



PREDICTIONS

As part of the analysis, the impact of increasing the top features that most significantly influenced the Net Promoter Score (NPS) were systematically evaluated using the best decision tree. Specifically, what will happen to the NPS score if the features identified to have the most potential effect on NPS are increased by 1 and 2 points?

This simulation aimed to provide a quantitative understanding of the sensitivity of NPS to changes in the most impactful factors. By incrementally adjusting these key features, the study sought to estimate the potential improvements, shedding light on the magnitude of impact each feature could have on overall patient satisfaction. Such simulations are valuable for strategic planning, enabling Keralty Hospitals to prioritize interventions based on their potential to yield the most substantial improvements in patient experiences and, consequently, NPS. This forward-looking analysis contributes to informed decision-making and the formulation of targeted initiatives to enhance patient satisfaction within the healthcare setting.

Here is an example of the West Florida region. For this example, we selected the Decision tree with five features as the Top 5 features found by correlation value and feature importance were the same.

Original Score

Here is the code to calculate the original NPS score for a region before simulating the incremental effects for these key features.

Python Script

```
# Let's see how many customers we have
demoters_df = df[df['NPS_bin'] == -1]
passives_df = df[df['NPS_bin'] == 0]
promoters_df = df[df['NPS_bin'] == 1]
print("Total =", len(df))

print("Number of customers who are clasified as demoters =",
len(demoters_df))
print("Percentage of customers who are clasified as demoters
=", 1.*len(demoters_df)/len(df)*100.0, "%")

print("Number of customers who are clasified as passives =",
len(passives_df))
print("Percentage of customers who are clasified as passives =",
1.*len(passives_df)/len(df)*100.0, "%")

print("Number of customers who are clasified as promoters =",
len(promoters_df))
print("Percentage of customers who are clasified as promoters
=", 1.*len(promoters_df)/len(df)*100.0, "%")



# Calculate NPS
promoters = df[df['NPS_bin'] == 1].shape[0]
passives = df[df['NPS_bin'] == 0].shape[0]
detractors = df[df['NPS_bin'] == -1].shape[0]

total_responses = promoters + passives + detractors
percentage_promoters = (promoters / total_responses) * 100
percentage_passives = (passives / total_responses) * 100
percentage_detractors = (detractors / total_responses) * 100
nps = percentage_promoters - percentage_detractors
print(f"Net Promoter Score (NPS): {nps}")
```

Result

```
Total = 378
Number of customers who are clasified as demoters = 85
Percentage of customers who are clasified as demoters =
22.486772486772484 %
Number of customers who are clasified as passives = 110
Percentage of customers who are clasified as passives =
29.100529100529098 %
Number of customers who are clasified as promoters = 183
Percentage of customers who are clasified as promoters =
48.41269841269841 %

Net Promoter Score (NPS): 25.925925925925927
```

New Scores

Here’s the code to incremented the selected features and to calculate the incremented NPS score

Increment

```
def increase_values(df, columns_to_increase):
  for column in columns_to_increase:
    # Increase values by 1, but don't go above 5
    df[column] = df[column].apply(lambda x: min(x + 1, 5))
  return df
```

Calculation

```
def calculate_nps_and_percentages(array):
  promoters = sum(1 for x in array if x == 1)
  passives = sum(1 for x in array if x == 0)
  detractors = sum(1 for x in array if x == -1)

  total_responses = len(array)

  nps = ((promoters - detractors) / total_responses) * 100

  percentage_promoters = (promoters / total_responses) * 100
  percentage_passives = (passives / total_responses) * 100
  percentage_detractors = (detractors / total_responses) * 100

  return nps, percentage_promoters, percentage_passives,
percentage_detractors
```

Specify columns to increase

```
columns_to_increase = ['Quality/experience dr.','lab
services','Exact diagnosis']
```

Apply the function

```
df_by1 = increase_values(df_new, columns_to_increase)
```

Once the increment is done use the selected decision tree and predict the new NPS score

```
y_pred = clf.predict(df_by1)

# Calculate NPS and percentages
nps_score, percentage_promoters, percentage_passives,
percentage_detractors =
calculate_nps_and_percentages(y_pred)

print(f"Net Promoter Score (NPS): {nps_score:.2f}%")
print(f"Percentage of Promoters:
{percentage_promoters:.2f}%")
print(f"Percentage of Passives: {percentage_passives:.2f}%")
print(f"Percentage of Detractors:
{percentage_detractors:.2f}%")
```

Result = Top 3 in correlation increased by 1

```
Net Promoter Score (NPS): 39.68%
Percentage of Promoters: 44.97%
Percentage of Passives: 49.74%
Percentage of Detractors: 5.29%
```

Repeat the process for increasing 2 points the top 3 and top 5 of correlation features and for increasing by 1 and 2 the top 3 and top 5 features found by random forrest

Result = Top 3 in correlation increased by 2

```
Net Promoter Score (NPS): 39.42%
Percentage of Promoters: 55.56%
Percentage of Passives: 28.31%
Percentage of Detractors: 16.14%
```

Result = Top 3 in Random Forrest increased by 1

```
Net Promoter Score (NPS): 30.16%
Percentage of Promoters: 40.74%
Percentage of Passives: 48.68%
Percentage of Detractors: 10.58%
```

Result = Top 3 in Random Forrest increased by 2

```
Net Promoter Score (NPS): 43.92%
Percentage of Promoters: 54.76%
Percentage of Passives: 34.39%
Percentage of Detractors: 10.85%
```

Remember that for this example of the West Florida region. it turns out that for the top 5 features, correlation and random features found the same features.

Result = Top 5 in correlation and random forest increased by 1

```
Net Promoter Score (NPS): 43.92%
Percentage of Promoters: 54.76%
Percentage of Passives: 34.39%
Percentage of Detractors: 10.85%
```

Result = Top 5 in correlation and random forest increased by 2

```
Net Promoter Score (NPS): 52.38%
Percentage of Promoters: 56.61%
Percentage of Passives: 39.15%
Percentage of Detractors: 4.23%
```

DISCUSSION

The present study encountered several challenges that warrant consideration when interpreting the findings. A primary limitation was the relatively modest size of the dataset, comprising only 1357 records, which further had to be partitioned into four distinct regions for a nuanced regional analysis. This constrained dataset size could potentially limit the generalizability of the findings, and as such, it is acknowledged that future research with a larger and more diverse dataset is warranted to enhance the robustness and applicability of the results. Additionally, the authenticity of the data, being derived from real-world scenarios within Keralty Hospitals, brought both advantages and challenges. While the authenticity offers a genuine reflection of patient experiences, it also introduces complexities due to certain information being intentionally concealed by Keralty to ensure customer privacy and safety.

With that said, here are some interesting points about the paper:

Notably, the revelation that overall NPS remained unaffected by features related to patient information adds an interesting layer to the understanding of patient satisfaction. This unexpected outcome challenges conventional assumptions and suggests that factors beyond demographic details play a more pivotal role in shaping the patient’s experience.

Intriguingly, the statistical analysis uncovered surprising insights, such as the state of the parking lot exhibiting a larger correlation with NPS than the patient's age. This unexpected discovery underscores the importance of seemingly subtle environmental factors in influencing patient satisfaction. It also highlights the value of data-driven analyses in unraveling nuances that might not be immediately evident.

Regional variations in top variables added another layer of complexity to the study. For instance, the prominence of "time waiting" and "Time of the appointment" as top variables in South Florida, especially in the bustling city of Miami, aligns with the expectation for efficiency in metropolitan areas. This regional nuance emphasizes the importance of tailoring strategies based on specific geographic considerations to truly address the unique needs and expectations of patients in different regions.

The ability to achieve accurate predictions with as few as five features, as validated through decision tree modeling, is a notable outcome. This streamlined set of features, selected through robust statistical methods, serves as a double verification of the study's results. It not only reinforces the reliability of the identified influential factors but also highlights the potential for practical and resource-efficient interventions to enhance patient satisfaction.

While these findings are enlightening, it is essential to acknowledge the study's limitations, including the relatively modest dataset size and the privacy constraints imposed by real-world data. Future research endeavors should aim to replicate and expand upon these findings with larger datasets and a continued commitment to privacy considerations. Nevertheless, the amalgamation of unexpected discoveries, regional nuances, and streamlined predictive accuracy contributes valuable insights to the broader discourse on patient satisfaction within healthcare institutions.

CONCLUSION

In conclusion, this comprehensive analysis of the Net Promoter Score (NPS) within Keralty Hospitals offers actionable insights that can significantly impact operational efficiency, competitive advantage, financial outcomes, and overall reputation. By identifying and focusing on the key features that have the most substantial influence on NPS, Keralty Hospitals can strategically enhance various aspects of their services.

Enhanced Operational Efficiency:

Addressing and improving the top variables influencing NPS provides a pathway to enhanced operational efficiency.

Competitive Advantage for Hospitals:

A focused approach to improving the top variables identified in this study can contribute to a competitive advantage for Keralty Hospitals. This advantage extends beyond superior medical care to encompass the holistic patient experience, creating a distinct and attractive healthcare offering.

Financial Benefits:

Strategically investing in and enhancing the top variables impacting NPS can yield significant financial benefits. By directing resources towards sectors that have the most influence on patient satisfaction, hospitals can optimize spending and, in turn, potentially increase revenue.

Positive Impact on Reputation:

Understanding and addressing patient preferences through the identified top variables can have a profound positive impact on the hospital's reputation. By consistently meeting patient expectations, Keralty Hospitals can build a reputation for exceptional healthcare services, leading to increased trust and positive word-of-mouth referrals.

In essence, this study not only uncovers the factors pivotal to patient satisfaction but also provides a roadmap for strategic interventions that can elevate Keralty Hospitals to new heights. By prioritizing improvements in the areas that matter most to patients, hospitals can forge a path towards operational excellence, a competitive edge, financial sustainability, and a sterling reputation within the healthcare landscape.

REFERENCES

1. Keralty Hospital
https://keraltyhospital.com/
2. Does the NPS® reflect consumer sentiment? A qualitative examination of the NPS using a sentiment analysis approach
https://journals.sagepub.com/doi/full/10.1177/1470785319863623
3. How to perform Anova in python
https://www.reneshbedre.com/blog/anova.html
4. A beginner's guide to Chi-square test in python from scratch
https://analyticsindiamag.com/a-beginners-guide-to-chi-square-test-in-python-from-scratch/
5. How improving your Net Promoter Score® can grow your healthcare business
https://www.surveymonkey.com/mp/healthcare-nps/

6. Data set Used on the experiment
https://github.com/JuanIgnacioCastro/NPS