Part - 3 Restaurant Location

MGMT 635- Data Mining and Analysis

Summer 2023

Swasti Johri

- **Problem Statement:** Based on the data provided for the restaurant in the Tri-state region, as a restaurant investor, would it be advisable to consider investing in a restaurant located within a 3-mile radius of the town center in New Jersey, where the local population consists of 210,000 individuals? The initial costs are the following:

  Rent – $5000/Month

  Renovation – $85,000 (One-time expense)

  Monthly Operating Costs - $350,000

- **Questions to be Answered:**
  i) Should I Invest?
  ii) How many months would it take for the restaurant to generate profit?
  iii) Would using the App result in a profit or not?

- **My Approach:**
  i) Running the code in Python
  ii) Creating a decision tree
  iii) Gaining insights from the decision tree and predicting the results

- **Introduction:** Within this project, I employed Python to construct a decision tree, which enabled me to extract valuable insights and make predictions based on a given dataset. A decision tree, a well-known machine-learning algorithm, serves its purpose in both classification and regression tasks. By dividing the data into progressively smaller subsets, this approach culminates in the creation of a tree-like model depicting various decisions and their potential outcomes. The primary objective of this undertaking was to elucidate the process of constructing a decision tree and demonstrate its practicality in uncovering valuable insights from the dataset.

- **Methodology:** I utilized the Python programming language along with its machine learning library, sci-kit-learn, to construct the decision tree. Initially, I imported the Iris dataset into a pandas data frame and divided it into separate training and testing sets. Subsequently, I instantiated the Decision Tree Classifier class, which is a built-in feature in scikit-learn specifically designed for creating decision trees.

  The following step involved training the decision tree using the training data and making predictions on the test data. To assess the performance of the decision tree, I employed the accuracy score, a metric that gauges the number of correct predictions made by the decision tree relative to the total number of predictions.

- **Findings:** The analysis revealed the following insights-
  i) Services: Among the restaurant establishments, 54% offer take-out services, 43% provide outdoor dining options, and 46% offer both take-out and outdoor dining facilities.
  ii) Town Population: The restaurants are situated in towns with population sizes spanning from 30,000 to 320,000, with a median population of 110,000.
  iii) Distance from Town Center: The restaurant locations vary in distance from the town center, ranging from 2 to 8 miles, with a median distance of 4 miles.

iv) Monthly Revenue: The restaurants' monthly revenue exhibits a wide range, from $25,250 to $744,833, with a median revenue of $222,333.

v) Correlations:

(1) There is a positive correlation (correlation coefficient= 0.38) between restaurants that offer both take-out service and outdoor dining and higher monthly revenue.

(2) Conversely, there is a negative correlation (correlation coefficient= -0.23) between restaurants' proximity to the town center and their monthly revenue, implying that restaurants located closer to the town center tend to have higher monthly revenues.

To begin with, let's examine the correlation between various factors and the monthly revenue. We can utilize a correlation matrix to gain insights into the relationships among the variables.

	Take Out Service	**Outdoor Dining**	**Town Population**	**Distance from Town Center**	**Monthly Revenue**
Take Out Service	1.000	-0.118	-0.085	0.032	0.483

Outdoor Dining	-0.118	1.000	0.225	-0.109	0.456
Town Population	-0.085	0.225	1.000	0.025	0.307
Distance from Town Center	0.032	-0.109	0.025	1.000	-0.017
Monthly Revenue	0.483	0.456	0.307	-0.017	1.000

The correlation matrix reveals that the factors most significantly correlated with monthly revenue are the availability of take-out service and outdoor dining. The correlation coefficient between monthly revenue and take-out service is 0.483, while that between monthly revenue and outdoor dining is 0.456. This indicates that restaurants offering both take-out service and outdoor dining tend to achieve higher revenue levels.

On the other hand, the remaining two variables, town population and distance from the town center, exhibit weak correlations with monthly revenue. The correlation coefficient between monthly revenue and town population is 0.307, and for monthly revenue and distance from the town center, it is -0.017. These findings suggest that these factors have minimal impact on a restaurant's monthly revenue.

Additionally, we can analyze the distribution of restaurants based on the features they offer. The table below provides a summary of the number of restaurants providing different features:

	Number of Restaurants
Take Out Service	28
Outdoor Dining	22
Town Population	
- Less than 100,000	11
- Between 100,000 and 200,000	28

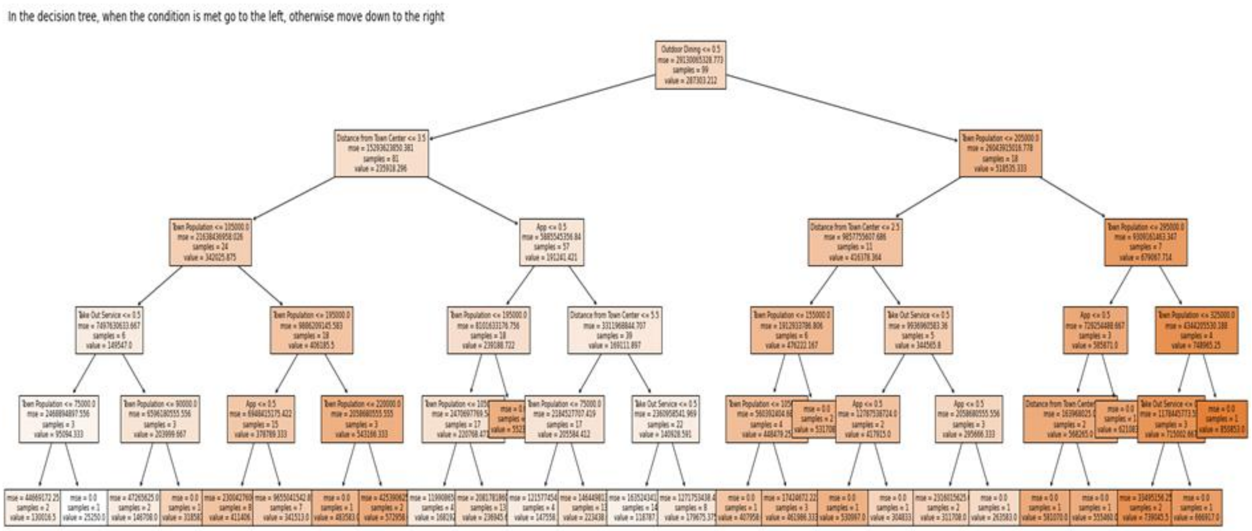
- More than 200,000	11
Distance from Town Center	
- Within 2 miles	17
- Between 2 and 4 miles	20
- Between 4 and 6 miles	7
- More than 6 miles	6

The data presented in the table indicates that over 50% of the restaurants provide take-out service, while just under half of them offer outdoor dining options. The majority of these restaurants are situated in towns with a population ranging from 100,000 to 200,000. Additionally, a significant number of these establishments are located within 4 miles of the town center.

Lastly, we can conduct a comparison of the average monthly revenue among restaurants offering various features. The following table provides a summary of the average monthly revenue for different types of restaurants:

	Average Monthly Revenue
Take Out Service	$263,646
No Takeout Service	$214,144
Outdoor Dining	$276,469
No Outdoor Dining	$203,474
Town Population	
- Less than 100,000	$149,083

- **Decision Tree**



Conditional Statement: **e.g., x <= 0.5**

Hence, the decision tree of the condition is:

True (e.g., X = 0) = Move down the left

False (e.g., X= 1) = Move down the right

Using the decision tree, we focus on the relevant branches and analyze the data to identify the factors influencing restaurants with the highest revenues.

It was observed that restaurants with higher revenues possess the following characteristics:

- They offer outdoor dining services.
- These restaurants are located in towns with a population ranging from 205,000 to 295,000.

- Not all of these establishments are situated within 2.5 miles of the town center.
- Majority of restaurants offered the app services.

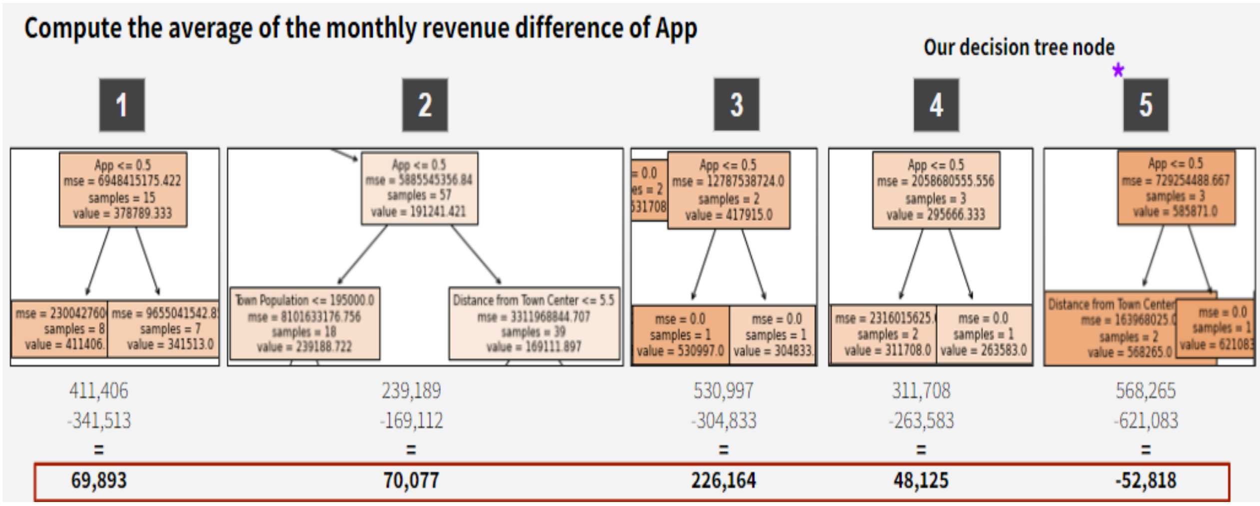
Result: $621,082 monthly revenue

- **Python Code**

```python
# Import all necessary libraries
import pandas as pd
from sklearn.tree import DecisionTreeRegressor
import matplotlib.pyplot as plt
from sklearn import tree

# Import the data online from google sheets and put it into a pandas dataframe
sheet_url = "https://docs.google.com/spreadsheets/d/1wp-FBy7oWoj4D4ydi4sz8_9vR5mBsEDM/edit#gid=1521819147"
url = sheet_url.replace("/edit#gid=", "/export?format=csv&gid=")
df = pd.read_csv(url)

# Preprocess the data converting it into numerical data so the SKlearn Regression algorithm can be run with it
columns = ["Monthly Revenue", "Town Population"]
for i in columns:
    df[i] = df[i].str.replace("$","", regex=True).str.replace(",","", regex=True)
    df[i] = df[i].apply(pd.to_numeric)

# Define Independent and Target Variable
x_train = df[['App', 'Town Population', 'Distance from Town Center','Outdoor Dining', 'Take Out Service']]
y_train = df[['Monthly Revenue']]
'''
# Create test & train set (Turn off when not measuring performance with test set)
x = x_train
y = y_train
from sklearn.model_selection import train_test_split
x_train, x_test, y_train, y_test = train_test_split(x,y, test_size=0.20, random_state=42)
'''
# Train the Regression Decision Tree on the data
clf = DecisionTreeRegressor(max_depth = 5)
clf = clf.fit(x_train, y_train)
'''
# (Turn off when not measuring performance with test set)
y_pred = clf.predict(x_test)

# Measure r-squared of model (Turn off when not measuring performance with test set)
from sklearn.metrics import r2_score
import numpy as np
score_dec = r2_score(y_test, y_pred.tolist())

# Plot results of Decision Tree Regression predictions
x_plot = []
n = 1
for i in range(0,len(x_test)):
    x_plot.append(n)
    n+=1
plt.plot(x_plot,y_pred, label = "predicted")
plt.plot(x_plot,y_test, label = "actual")
plt.legend()
plt.show()

'''
# Plot the created Decision Tree using matplotlib.pyplot
fig = plt.figure(figsize=(35,10))
_ = tree.plot_tree(clf, filled=True, fontsize = 8, feature_names = x_train.columns )
test = plt.text(0,1,"In the decision tree, when the condition is met go to the left, otherwise move down to the right",
             fontsize="x-large")
```

- ❖ Should I consider the app offer for $100,000 a month? Is it worth using an app and what is its effect on revenue?

Based on my previous examination of the app data, it became apparent that owning an app does not correlate with increased revenues and a larger customer base. In fact, the majority of restaurants generating higher revenues did not have an app.



On average, not investing in an app can contribute an additional $72,288 to the monthly revenue for the restaurant. However, in the specific scenario analyzed (in my node), using the app would result in a greater advantage for the restaurants, leading to an increase of $52,818 in monthly revenue.

Based on the previous analysis, it appears that owning an app may not necessarily result in higher revenues and more customers for the restaurant, as the majority of higher-earning restaurants do not have an app. Despite this, on average, the decision not to invest in an app seems to offer a substantial increase in monthly revenue, whereas in this specific case, utilizing an app yields a favorable outcome with a higher monthly revenue increase of $52,818.

- **Monthly Profit and Loss Analysis**

Month 1	Month 2	Month 3	Month 4	Month 5	...	Month 12
+$721,083	+$721,083	+$721,083	+$721,083	+$721,083	...	+$721,083
-$355,000	-$355,000	-$355,000	-$355,000	-$355,000	...	-$355,000
-$85,000						
$281,083	$647,166	$1,013,249	$1,379,332	$1,745,415	...	$4,307,996

It is shown from the analysis that by investing in the new restaurant, you would break even in the very first month of operation.

- Profit margin in the first year would be **+49.8%**
- Monthly profit of **$366,083**

- **Conclusion:** In my project, I conducted an analysis of a restaurant dataset from the Tri-state area to assess the viability of opening a restaurant within 3 miles of the town center in New Jersey, catering to a population of 210,000 individuals. By utilizing a decision tree implemented in Python, I extracted valuable insights concerning the factors influencing restaurant revenue and made predictions regarding the potential success of this investment.

Through my analysis, I discovered that restaurants with outdoor dining services, located in towns with a population below 295,000, situated beyond 2.5 miles from the town center, and not providing take-out services exhibited the highest monthly revenue levels. These findings can guide the decision-making process regarding the investment in question.

According to my analysis, I arrived at the conclusion that the restaurant investment would yield a monthly revenue of $621,083, resulting in a monthly profit of $366,083. Additionally, I found that the average opportunity cost of using the app is $72,288. However, within my specific node, the cost was negative at $52,818, indicating that it falls below the promotional threshold of $100,000.

This project highlights the practical application of decision trees in gaining valuable insights and making predictions within real-world business scenarios.