Names: Hamida Paiman (hp602) & Rainie Li (rl864)
Professor Gunawardena
Data Management for Data Science
9 December 2024

Final Project Report

Introduction

For this project, we have used the dataset 'Advertising.csv', which includes the values for the advertising spend in relation to sales. The primary problem addressed by this project is the strategic allocation of advertising budgets to maximize sales. Businesses often face the challenge of finding which advertising channel, such as TV, radio, or newspapers, would yield the most effective advertising. Without a proper understanding of the data, it can lead to difficulties, such as misallocation, which can have a negative impact as the budget would not be put to good use. Businesses may even miss out on considerable opportunities to grow their business. This project aims to provide knowledgeable insights as to how advertising spending on different types of advertising could impact sales and develop a predictive model that would make good decisions based on the data. The strategic aspects involve understanding the relationship between the advertising platforms and the sales performance as well as finding the best and worst advertising channels to optimize the allocation of resources. By using data analysis and statistics, this project gives informed insights into strategic marketing decisions. This project references the concepts that were discussed in class, such as data analysis, predictive modeling, and implementations of linear regressions.

This project is important because businesses are always looking for ways to optimize their marketing, especially in this economy where it is filled with many competitors. These businesses must find a way to flourish and stand out in the market, and one of the main components of a successful business is advertising and marketing. Thus, much of the budget for business is spent on advertising, so the advertising money needs to be spent wisely and effectively. By finding clarity on which type of advertising contributes significantly to sales, this project will allow businesses to use their money in the most optimal way.

This project piqued our interest because the dataset relates to the real world. We see advertisements every day, whether that pertains to turning on the TV at home, opening the radio during our commutes to Rutgers, or reading the daily newspaper. We consume advertisements plentifully, and these advertisements, either consciously or subconsciously, affect our decisions on what type of business we want to invest in. Working with this dataset allows us to connect the topics we learned in class about datasets with real-world applications. The skill to predict sales outcomes can be helpful, especially if we want to start a business of our own in the future.

Some existing issues in current data management practice are close-mindedness and insufficient understanding of data. Some businesses do not try out multiple advertising techniques, thus missing out on good opportunities to expand their business sales. This close-mindedness would be a problem because their insights would be based on an isolated and narrow dataset. Additionally, without having an understanding of the data, some businesses would make decisions based on intuition rather than real and factual information. Thus, the decisions that would be made are not backed up by informative evidence. We collect data for a reason, and that reason is to provide proof and confirmation of our initial thought process. This data collection can also be used to develop predictive models, which can be helpful for future technological systems.

Data Collection

The data set that this project focuses on contains values for the advertising platforms, namely TV, radio, and newspaper, in reference to the sales values. The data contains independent variables, which are the records of the advertising spending from the channels, and the dependent variable, which is the corresponding sales values. The independent variables represent the advertising expenditure, denoted in thousands of dollars, while the dependent variable is the total product sales in thousands of units, which can be useful for predictive analysis. Using SQL, we have found the row with the minimum and maximum values for sales.

```
import sqlite3
import pandas as pd
conn = sqlite3.connect('database.db')
min_sales_row = pd.read_sql('SELECT * FROM advertise WHERE sales = (SELECT MIN(sales) FROM advertise)', conn)
print("Row with the minimum sales:")
print(min_sales_row)
conn.close()
```

```
Row with the minimum sales:
    TV  radio  newspaper  sales
0  0.7   39.6        8.7    1.6
```

```
import sqlite3
import pandas as pd
conn = sqlite3.connect('database.db')
max_sales_row = pd.read_sql('SELECT * FROM advertise WHERE sales = (SELECT MAX(sales) FROM advertise)', conn)
print("Row with the maximum sales:")
print(max_sales_row)
conn.close()
```

```
Row with the maximum sales:
      TV  radio  newspaper  sales
0  276.9   48.9       41.8   27.0
```

Data Cleaning

Before analyzing the data and making prediction models, one of the crucial steps is cleaning and transforming the data and preparing it for analysis. Our data set (Advertising.csv) consists of 200 rows and 4 columns. We have started by detecting missing values and outliers. After cleaning our data, we found that our data set does not contain any missing values or outliers.

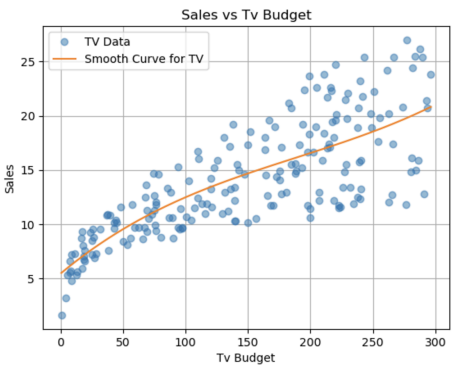
```
Data Quality Report:
      Column Data Type  Missing Values  Duplicates  Unique Values
0         tv   float64               0           0            190
1      radio   float64               0           0            167
2  newspaper   float64               0           0            172
3      sales   float64               0           0            121

Cleaned Data Preview:
      tv  radio  newspaper  sales
0  230.1   37.8       69.2   22.1
1   44.5   39.3       45.1   10.4
2   17.2   45.9       69.3    9.3
3  151.5   41.3       58.5   18.5
4  180.8   10.8       58.4   12.9
```

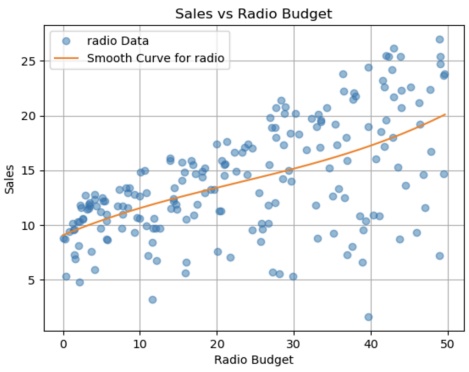
We can also clean the dataset in an alternative way using **SQL**:

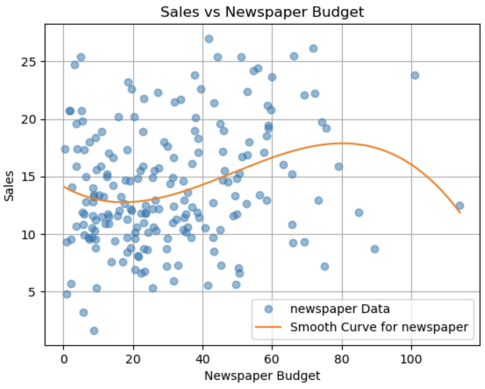
```
import sqlite3
import pandas as pd
conn = sqlite3.connect('database.db')
conn.execute('DELETE FROM advertise WHERE tv IS NULL OR radio IS NULL OR newspaper IS NULL OR sales IS NULL')
conn.commit()
cleaned_data = pd.read_sql('SELECT * FROM advertise', conn)
print("Table after removing rows with missing values:")
print(cleaned_data)
conn.close()
```

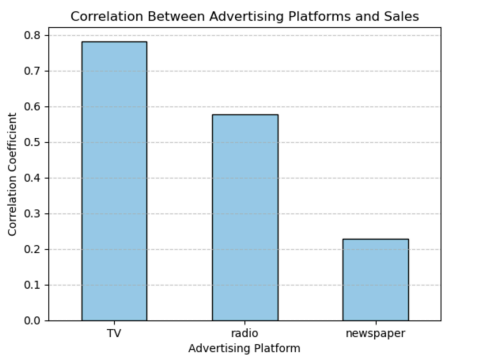
```
Table after removing rows with missing values:
        TV  radio  newspaper  sales
0    230.1   37.8       69.2   22.1
1     44.5   39.3       45.1   10.4
2     17.2   45.9       69.3    9.3
3    151.5   41.3       58.5   18.5
4    180.8   10.8       58.4   12.9
..     ...    ...        ...    ...
195   38.2    3.7       13.8    7.6
196   94.2    4.9        8.1    9.7
197  177.0    9.3        6.4   12.8
198  283.6   42.0       66.2   25.5
199  232.1    8.6        8.7   13.4

[200 rows x 4 columns]
```

Additionally, we found the correlation of the sales with TV, radio, and newspaper variables as a basic analysis of the dependent and independent variables.










```
correlation of sales with each advertising platform:
TV           0.782224
radio        0.576223
newspaper    0.228299
dtype: float64
```

The analysis shows that TV spend has a stronger correlation with sales, which is clearly visualized in a graph. However, if a company were to make a decision on which platform to allocate more of its budget, this analysis is not sufficient to justify focusing primarily on TV advertising. This is because advertising on TV requires a significantly higher investment compared to other platforms. As demonstrated below, you can see that TV has the highest average budget.

```
Averages:
TV: 147.0425
Radio: 23.264
Newspaper: 30.554000000000002
The highest average is in 'tv' with an average of 147.0425
The lowest average is in 'radio' with an average of 23.264
```

For more analysis, we have made a prediction model to select the best advertising platform.

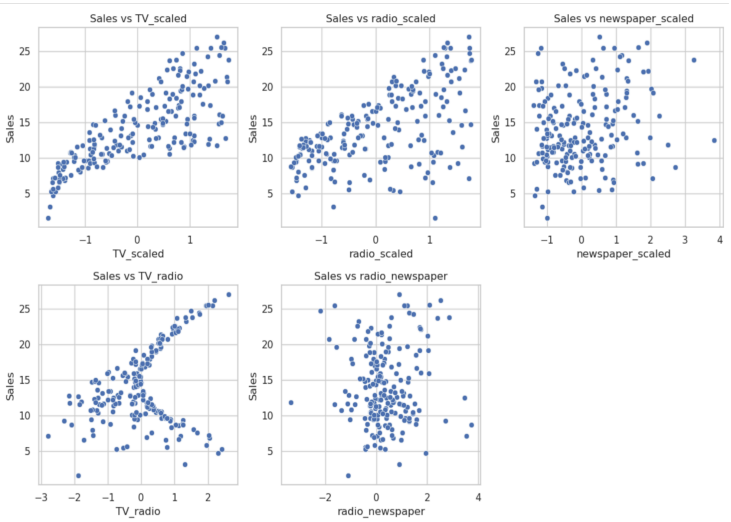
Machine Learning and Prediction Model Using Linear Regression

First, we have defined the features and target variable, split the data into the training and testing sets, and evaluated the model using parameters like Mean Squared Error, Root Mean Squared Error, Mean Absolute Error, and R-squared.

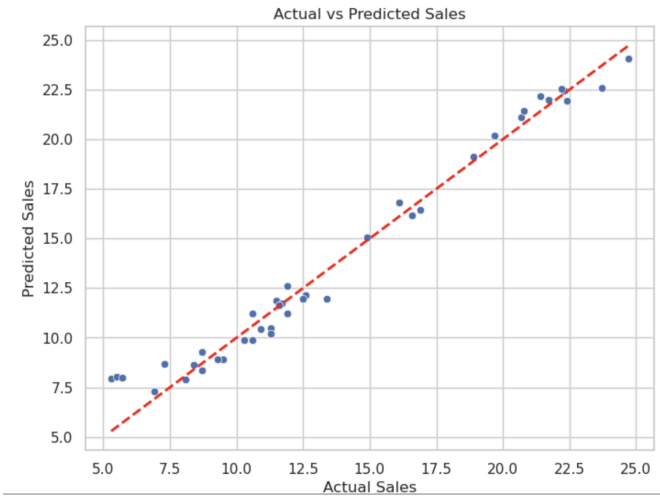
```
Mean Squared Error: 3.1740973539761046
   Actual  Predicted
0    16.9  16.408024
1    22.4  20.889882
2    21.4  21.553843
3     7.3  10.608503
4    24.7  22.112373
R-squared: 0.899438024100912
```

```
Root Mean Squared Error (RMSE): 1.7815996615334502
Mean Absolute Error (MAE): 1.4607567168117606

Model for TV:
Mean Squared Error: 10.204654118800956
R-squared: 0.6766954295627077

Model for radio:
Mean Squared Error: 23.248766588129108
R-squared: 0.2634309396999791

Model for newspaper:
Mean Squared Error: 30.620733995242567
R-squared: 0.029871749149522175
```

We can achieve better and higher accuracy through **data transformation**, **scaling**, and **feature engineering.**




```
Mean Squared Error: 0.8270625515249478
R-squared: 0.9737969459981106

Sample Predictions:
     Actual  Predicted
95     16.9  16.447745
15     22.4  21.933118
30     21.4  22.158128
158     7.3   8.664689
128    24.7  24.070416
```



Now, we have a higher R-squared for the model.

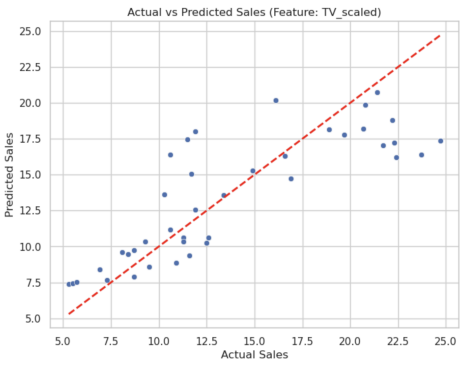
Furthermore, we used **Random Forest** and **Logistic Regression** to evaluate which one does better on this dataset.

```
Logistic Regression Mean Squared Error (CV): 0.9996709919221534
Random Forest Mean Squared Error (CV): 0.6833565799999993

Model Comparison:
                 Model  Mean Squared Error
0  Logistic Regression            0.999671
1        Random Forest            0.683357
```

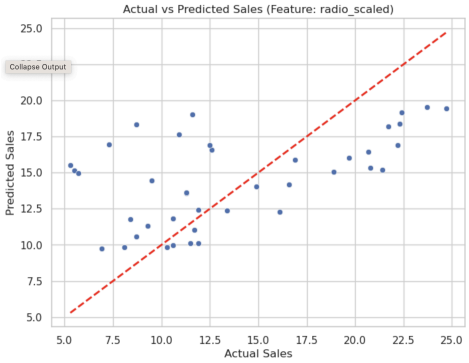
By the output above we can conclude that Random Forest performs better on this dataset as it gives us a lower mean squared error.

Final Evaluation

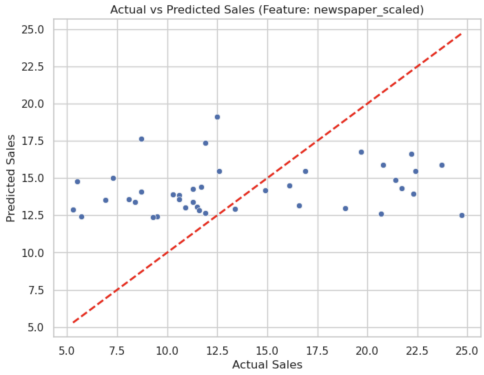
```
Model using feature: TV_scaled
Mean Squared Error: 10.204654118800956
R-squared: 0.6766954295627077

Sample Predictions:
     Actual  Predicted
95     16.9  14.717944
15     22.4  16.211548
30     21.4  20.748197
158     7.3   7.664036
128    24.7  17.370139
```




```
Model using feature: radio_scaled
Mean Squared Error: 23.248766588129108
R-squared: 0.2634309396999791

Sample Predictions:
     Actual  Predicted
95     16.9  15.883322
15     22.4  19.174272
30     21.4  15.208779
158     7.3  16.966678
128    24.7  19.440001
```




```
Model using feature: newspaper_scaled
Mean Squared Error: 30.620733995242567
R-squared: 0.029871749149522175

Sample Predictions:
     Actual  Predicted
95     16.9  15.471678
15     22.4  15.471678
30     21.4  14.892038
158     7.3  15.011551
128    24.7  12.501770
```



By our output, we can conclude that TV is the best platform of advertising to spend on in order to achieve higher sales. The worst platform for advertising are newspapers according to our analysis. This project is a sample and is important for a company to reference for future use to make their business decisions and budget allocations on advertising platforms.

In addition, this project is a very helpful reference for learning and practicing data analysis such as collecting data, SQL, handling missing values, data transformation, feature engineering, linear regression and prediction model using Python.

The advantages of our approach is that we can use the data analysis resulting from this project in a broader sense. We can implement it into future projects in regards to advertising sales. Thus, if we were to establish a business that requires advertising, we can use the results from this project and keep it in mind on what is the best platform to advertise the business. However, one limitation of our approach is that we focused solely on sales in reference to TV, newspaper, and radio. There are many more existing platforms that are used for advertising. One example includes advertising from social media platforms, such as Youtube Ads, Instagram Ads, and more. Thus, the restrictions of the project should be taken into account when deciding the best platform, since platforms are not limited to just TV, radio, and newspapers.