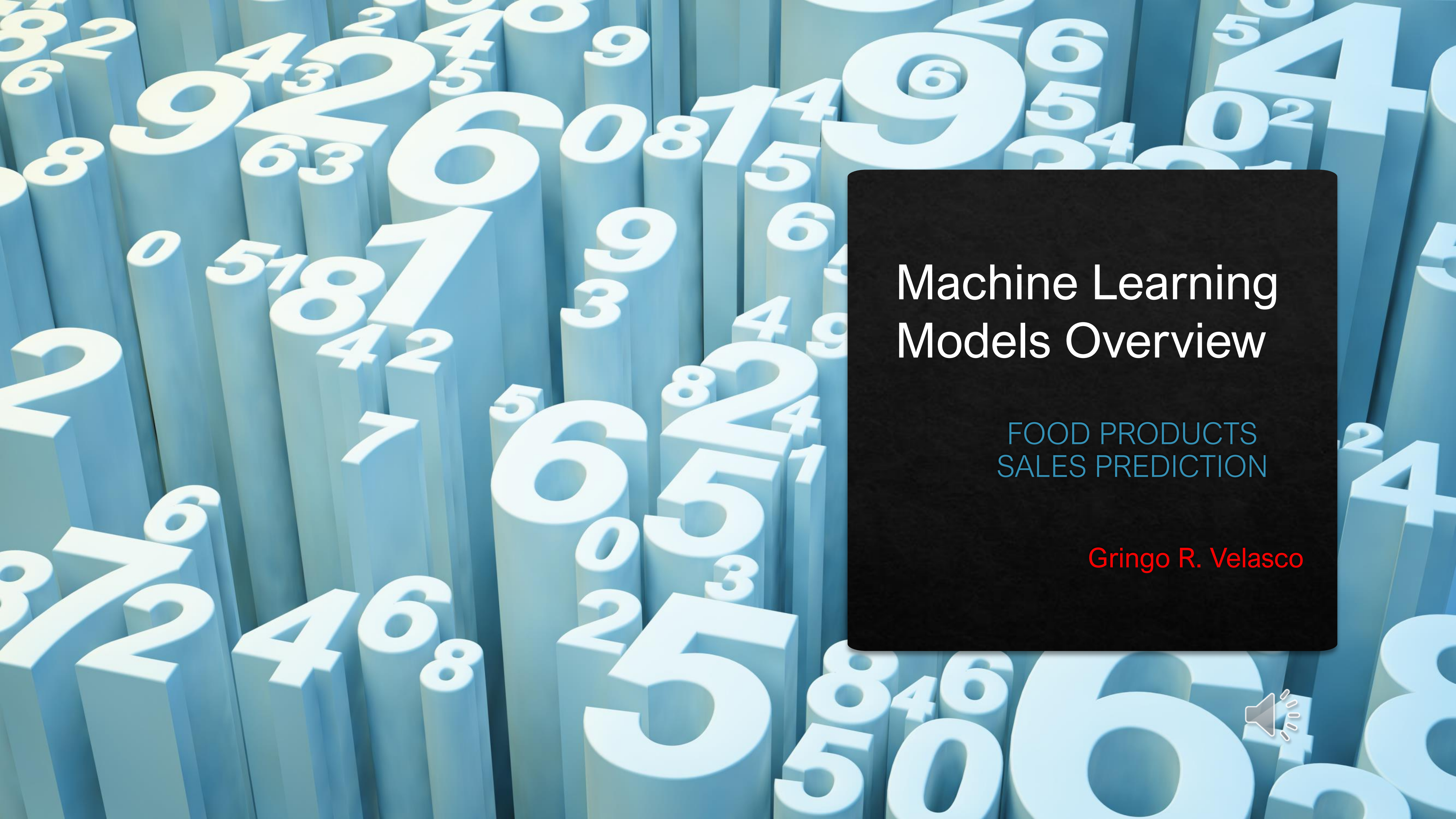


Machine Learning Models Overview

FOOD PRODUCTS SALES PREDICTION

Gringo R. Velasco

BACKGROUND

As with all running businesses around the globe, sales prediction or forecasting is one of the most important areas in order to successfully run a business. It determines how the company invests and grows and can have a massive impact on company valuation.

This presentation will be a sales prediction for food items sold at various stores. The goal of this is to help the retailer understand the properties of products and outlets that play crucial roles in increasing sales.

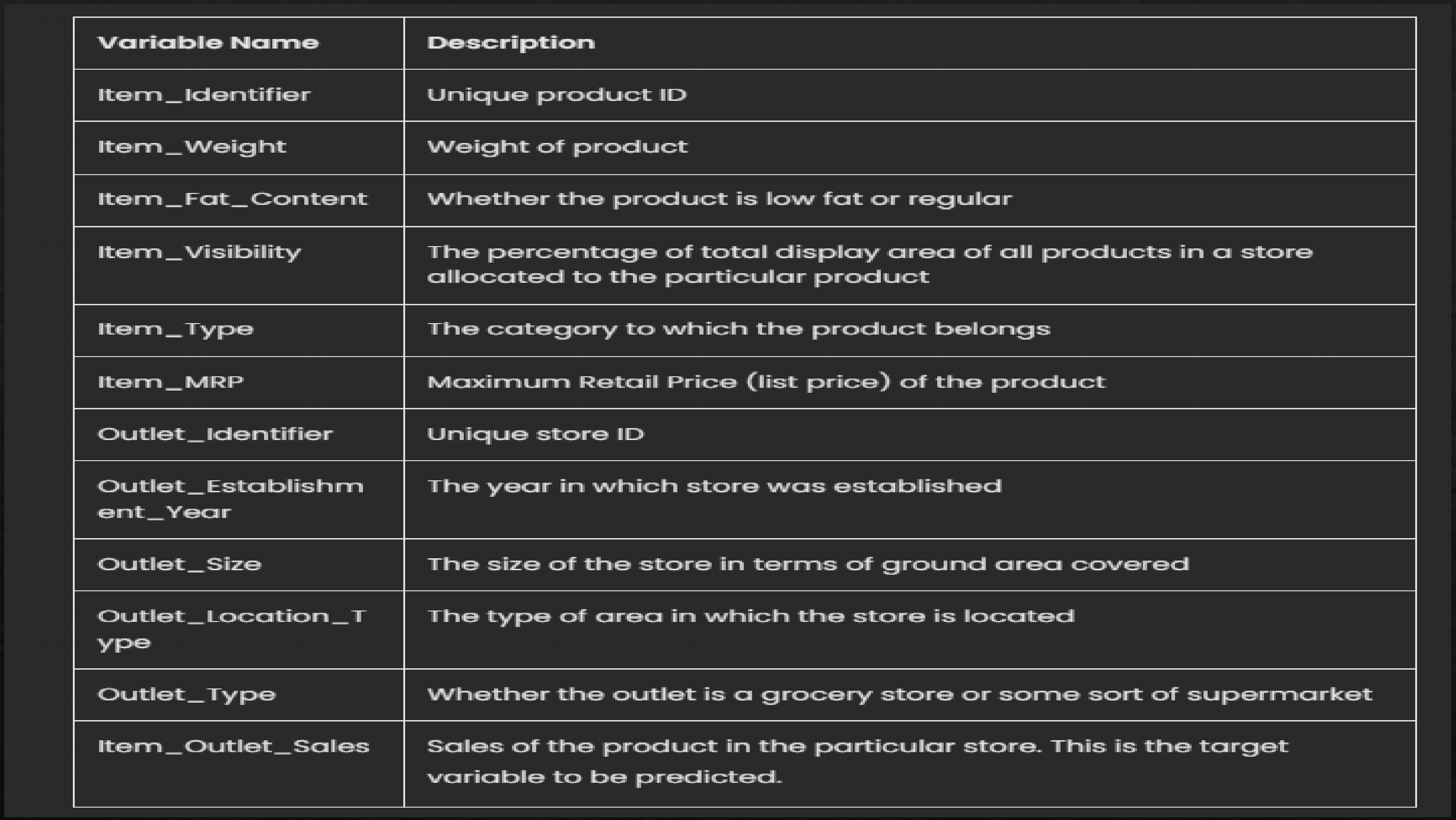
WHY USED MACHINE LEARNING IN PREDICTING SALES?

- Machine learning models are created from machine learning algorithms, which are trained using either labeled, unlabeled, or mixed data that are used to recognize patterns in data or make predictions.
- This project will cover 2 machine learning models that are commonly used in making predictions
  1. Linear Regression
  2. Decision Tree Regressor

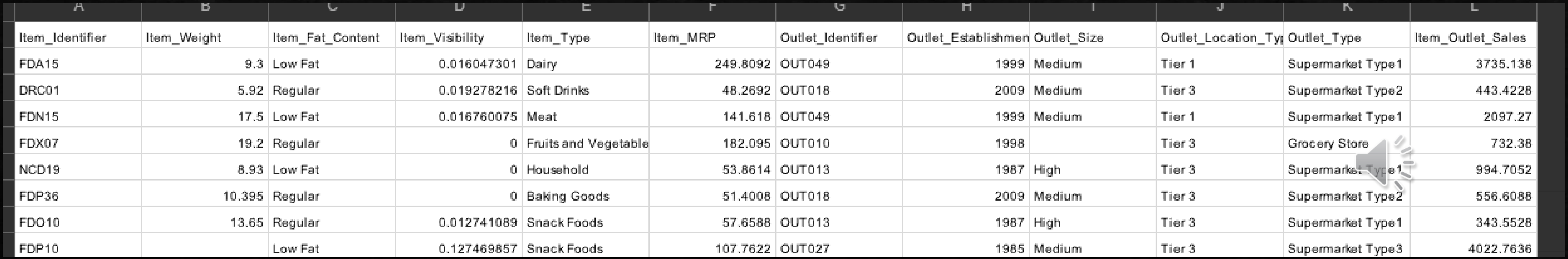
1. Linear Regression

- Linear regression is one of the easiest and most popular Machine Learning algorithms that is used for predictive analysis.
- Linear regression makes predictions for continuous/real or numeric variables such as **sales**, salary, age, product price, etc.

Dataset to be used:



Variable Name	Description
Item_Identifier	Unique product ID
Item_Weight	Weight of product
Item_Fat_Content	Whether the product is low fat or regular
Item_Visibility	The percentage of total display area of all products in a store allocated to the particular product
Item_Type	The category to which the product belongs
Item_MRP	Maximum Retail Price (list price) of the product
Outlet_Identifier	Unique store ID
Outlet_Establishment_Year	The year in which store was established
Outlet_Size	The size of the store in terms of ground area covered
Outlet_Location_Type	The type of area in which the store is located
Outlet_Type	Whether the outlet is a grocery store or some sort of supermarket
Item_Outlet_Sales	Sales of the product in the particular store. This is the target variable to be predicted.

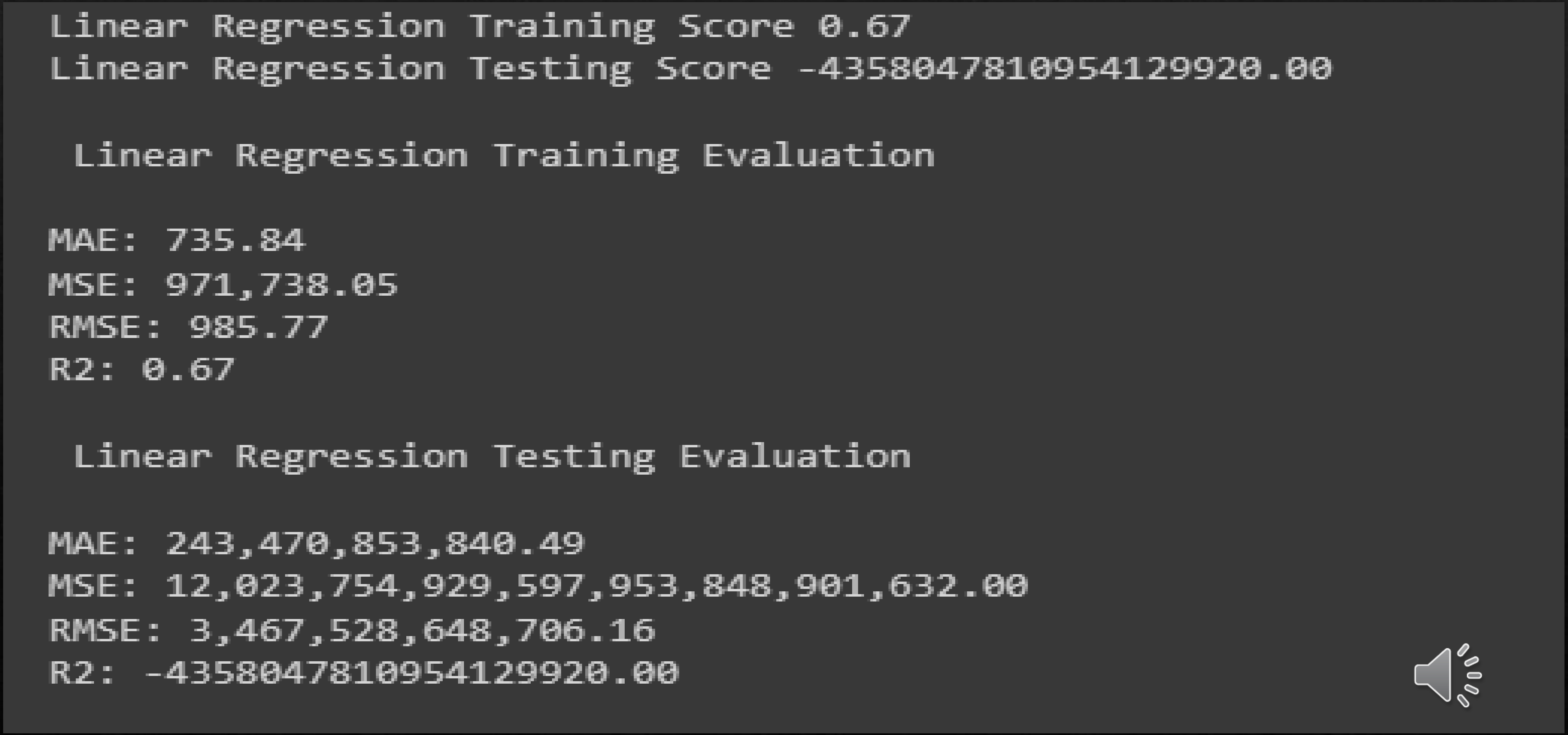


A	B	C	D	E	F	G	H	I	J	K	L
Item_Identifier	Item_Weight	Item_Fat_Content	Item_Visibility	Item_Type	Item_MRP	Outlet_Identifier	Outlet_Establishmen	Outlet_Size	Outlet_Location_Ty	Outlet_Type	Item_Outlet_Sales
FDA15	9.3	Low Fat	0.016047301	Dairy	249.8092	OUT049	1999	Medium	Tier 1	Supermarket Type1	3735.138
DRC01	5.92	Regular	0.019278216	Soft Drinks	48.2692	OUT018	2009	Medium	Tier 3	Supermarket Type2	443.4228
FDN15	17.5	Low Fat	0.016760075	Meat	141.618	OUT049	1999	Medium	Tier 1	Supermarket Type1	2097.27
FDX07	19.2	Regular	0	Fruits and Vegetable	182.095	OUT010	1998		Tier 3	Grocery Store	732.38
NCD19	8.93	Low Fat	0	Household	53.8614	OUT013	1987	High	Tier 3	Supermarket Type1	994.7052
FDP36	10.395	Regular	0	Baking Goods	51.4008	OUT018	2009	Medium	Tier 3	Supermarket Type2	556.6088
FDO10	13.65	Regular	0.012741089	Snack Foods	57.6588	OUT013	1987	High	Tier 3	Supermarket Type1	343.5528
FDP10		Low Fat	0.127469857	Snack Foods	107.7622	OUT027	1985	Medium	Tier 3	Supermarket Type3	4022.7636

BUILDING THE MODEL

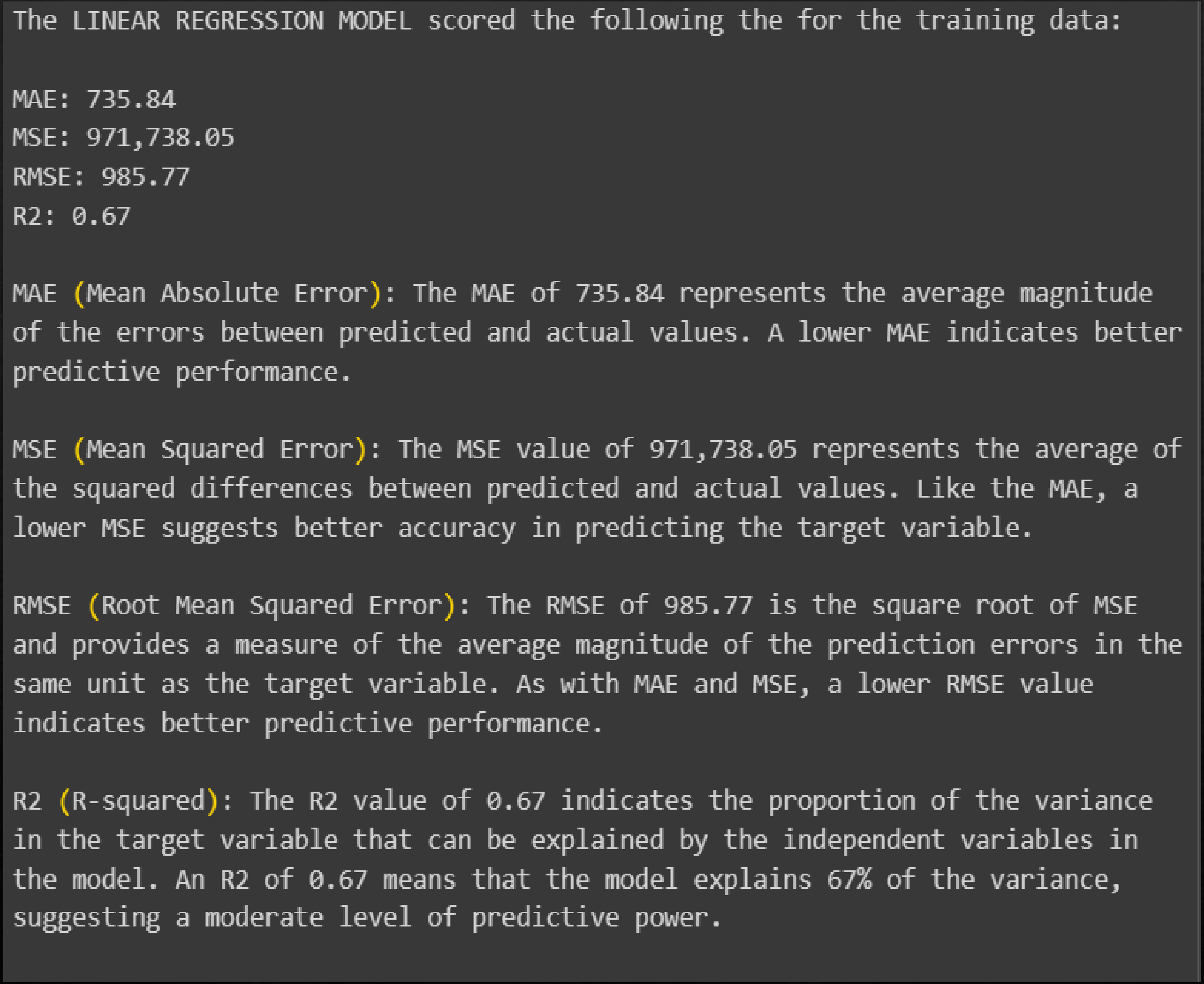
- STEPS
  1. Load the data (training data)
  2. Fit the model in the training data
  3. Evaluate the model performance
  4. Find the best model

LINEAR REGRESSION MODEL PERFORMANCE



```
Linear Regression Training Score 0.67
Linear Regression Testing Score -4358047810954129920.00

 Linear Regression Training Evaluation

MAE: 735.84
MSE: 971,738.05
RMSE: 985.77
R2: 0.67

 Linear Regression Testing Evaluation

MAE: 243,470,853,840.49
MSE: 12,023,754,929,597,953,848,901,632.00
RMSE: 3,467,528,648,706.16
R2: -4358047810954129920.00
```

LINEAR REGRESSION MODEL PERFORMANCE



The LINEAR REGRESSION MODEL scored the following the for the training data:

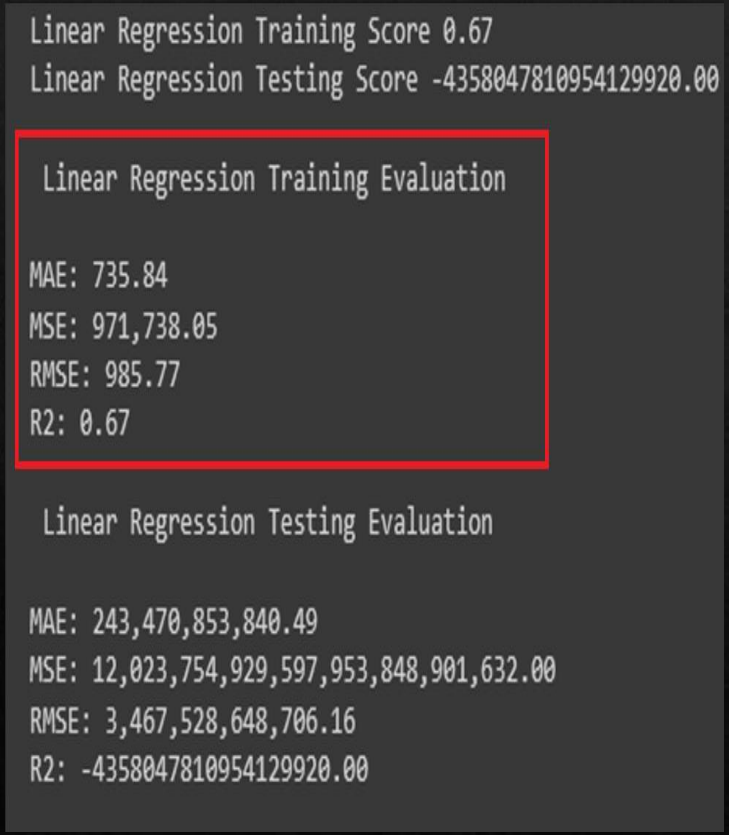
MAE: 735.84
MSE: 971,738.05
RMSE: 985.77
R2: 0.67

MAE (Mean Absolute Error): The MAE of 735.84 represents the average magnitude of the errors between predicted and actual values. A lower MAE indicates better predictive performance.

MSE (Mean Squared Error): The MSE value of 971,738.05 represents the average of the squared differences between predicted and actual values. Like the MAE, a lower MSE suggests better accuracy in predicting the target variable.

RMSE (Root Mean Squared Error): The RMSE of 985.77 is the square root of MSE and provides a measure of the average magnitude of the prediction errors in the same unit as the target variable. As with MAE and MSE, a lower RMSE value indicates better predictive performance.

R2 (R-squared): The R2 value of 0.67 indicates the proportion of the variance in the target variable that can be explained by the independent variables in the model. An R2 of 0.67 means that the model explains 67% of the variance, suggesting a moderate level of predictive power.



```
Linear Regression Training Score 0.67
Linear Regression Testing Score -4358047810954129920.00

 Linear Regression Training Evaluation

MAE: 735.84
MSE: 971,738.05
RMSE: 985.77
R2: 0.67

 Linear Regression Testing Evaluation

MAE: 243,470,853,840.49
MSE: 12,023,754,929,597,953,848,901,632.00
RMSE: 3,467,528,648,706.16
R2: -4358047810954129920.00
```

After training the data, the below shows the Testing scores for the model Linear Regression

MAE: 243,470,853,840.49
MSE: 12,023,754,929,597,953,848,901,632.00
RMSE: 3,467,528,648,706.16
R2: -4,358,047,810,954,129,920.00
These scores indicate the following:

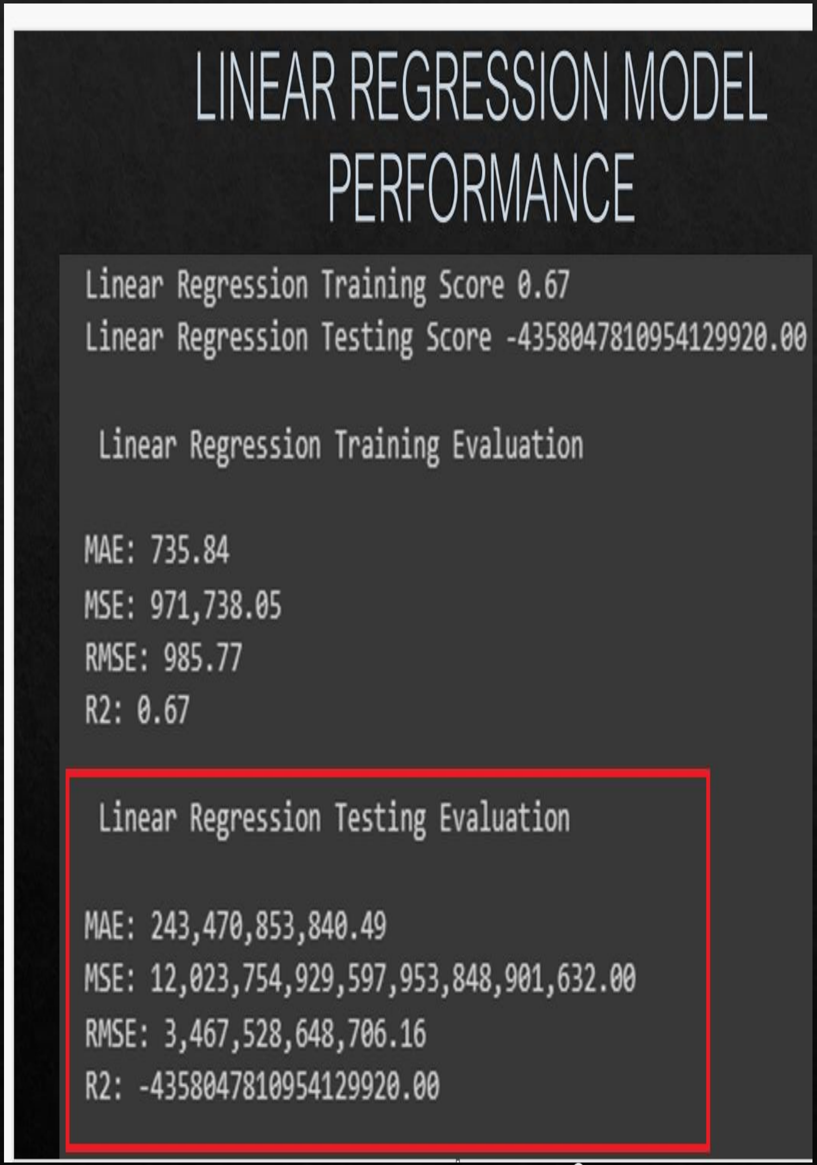
MAE: The MAE value of 243,470,853,840.49 is exceptionally high. It suggests that there is a significant difference between the predicted and actual values, indicating poor predictive performance.

MSE: The MSE value of 12,023,754,929,597,953,848,901,632.00 is extremely large. It implies a high degree of variation between predicted and actual values, further indicating poor predictive performance.

RMSE: The RMSE value of 3,467,528,648,706.16, which is the square root of MSE, is also extremely large. It confirms the significant magnitude of the prediction errors.

R2: The R2 value of -4,358,047,810,954,129,920.00 is negative, which is highly unusual and suggests a severe problem with the model. R2 should typically range between 0 and 1, with negative values indicating that the model performs worse than a simple mean prediction.

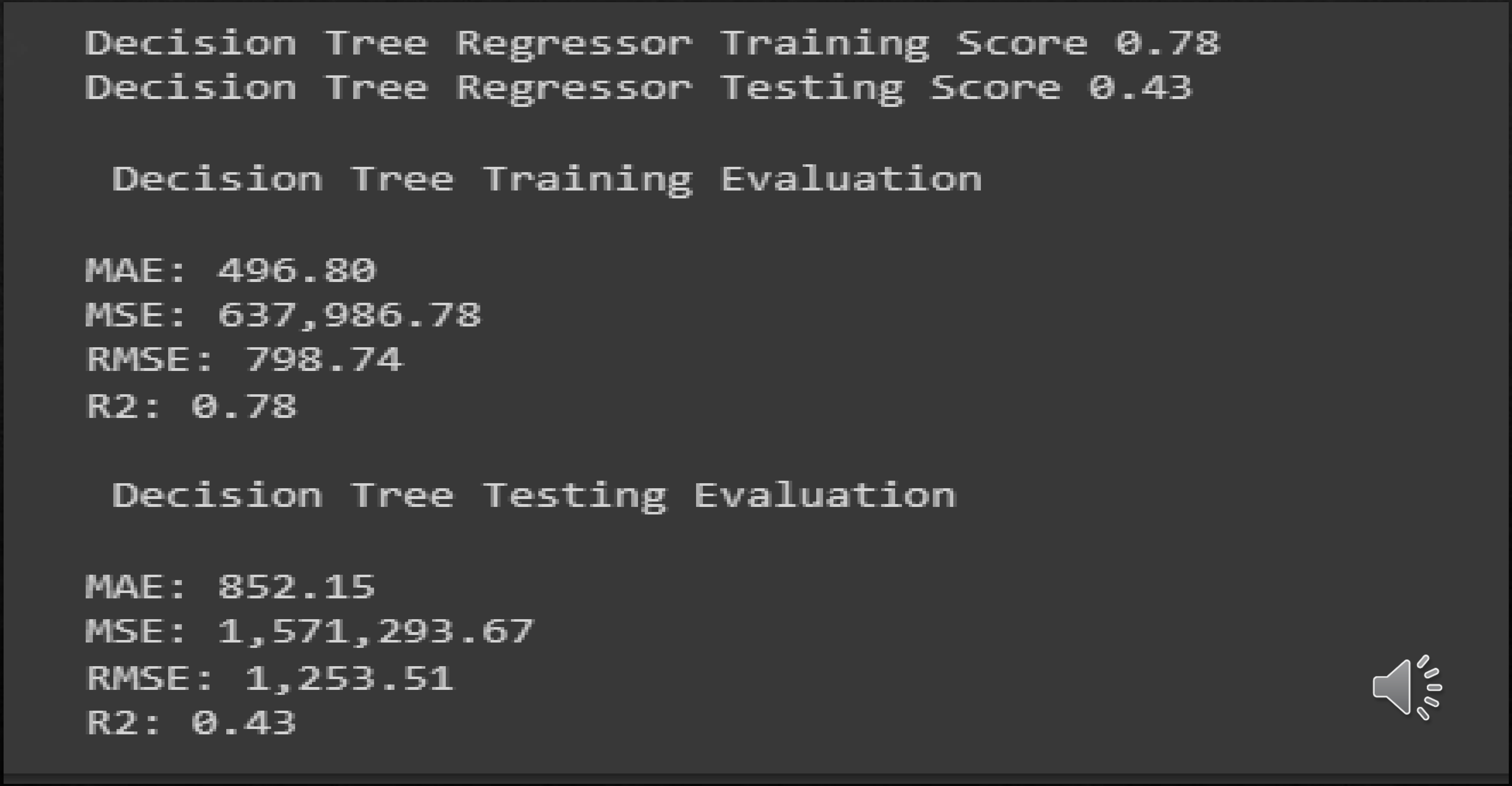
In summary, the first set of scores indicates relatively good predictive performance with moderate values for MAE, MSE, RMSE, and R2. On the other hand, the second set of scores shows extremely poor predictive performance, with extremely high values for MAE, MSE, and RMSE, as well as a negative R2 value. It is important to assess the context and potential issues in the model to understand the reasons behind such extreme scores.



LINEAR REGRESSION MODEL PERFORMANCE

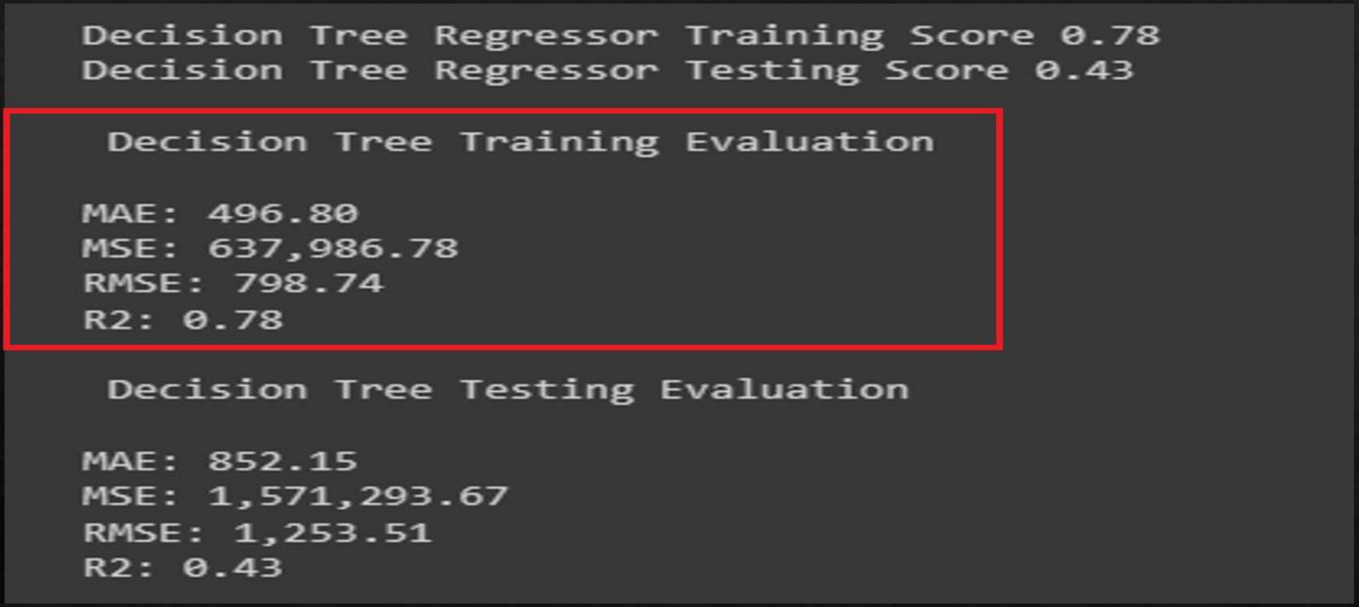
```
Linear Regression Training Score 0.67
Linear Regression Testing Score -4358047810954129920.00

 Linear Regression Training Evaluation

MAE: 735.84
MSE: 971,738.05
RMSE: 985.77
R2: 0.67

 Linear Regression Testing Evaluation

MAE: 243,470,853,840.49
MSE: 12,023,754,929,597,953,848,901,632.00
RMSE: 3,467,528,648,706.16
R2: -4358047810954129920.00
```

2. DECISION TREE REGRESSOR MODEL PERFORMANCE



```
Decision Tree Regressor Training Score 0.78
Decision Tree Regressor Testing Score 0.43

 Decision Tree Training Evaluation

MAE: 496.80
MSE: 637,986.78
RMSE: 798.74
R2: 0.78

 Decision Tree Testing Evaluation

MAE: 852.15
MSE: 1,571,293.67
RMSE: 1,253.51
R2: 0.43
```

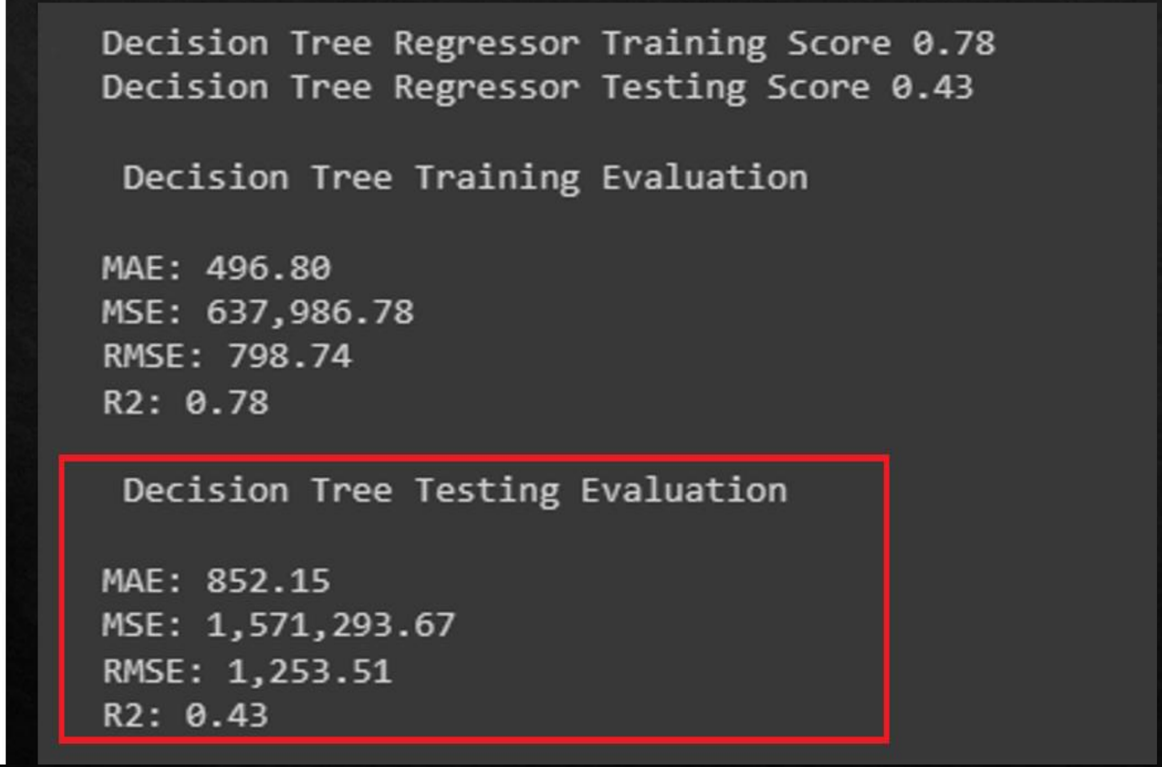
2. DECISION TREE REGRESSOR MODEL PERFORMANCE



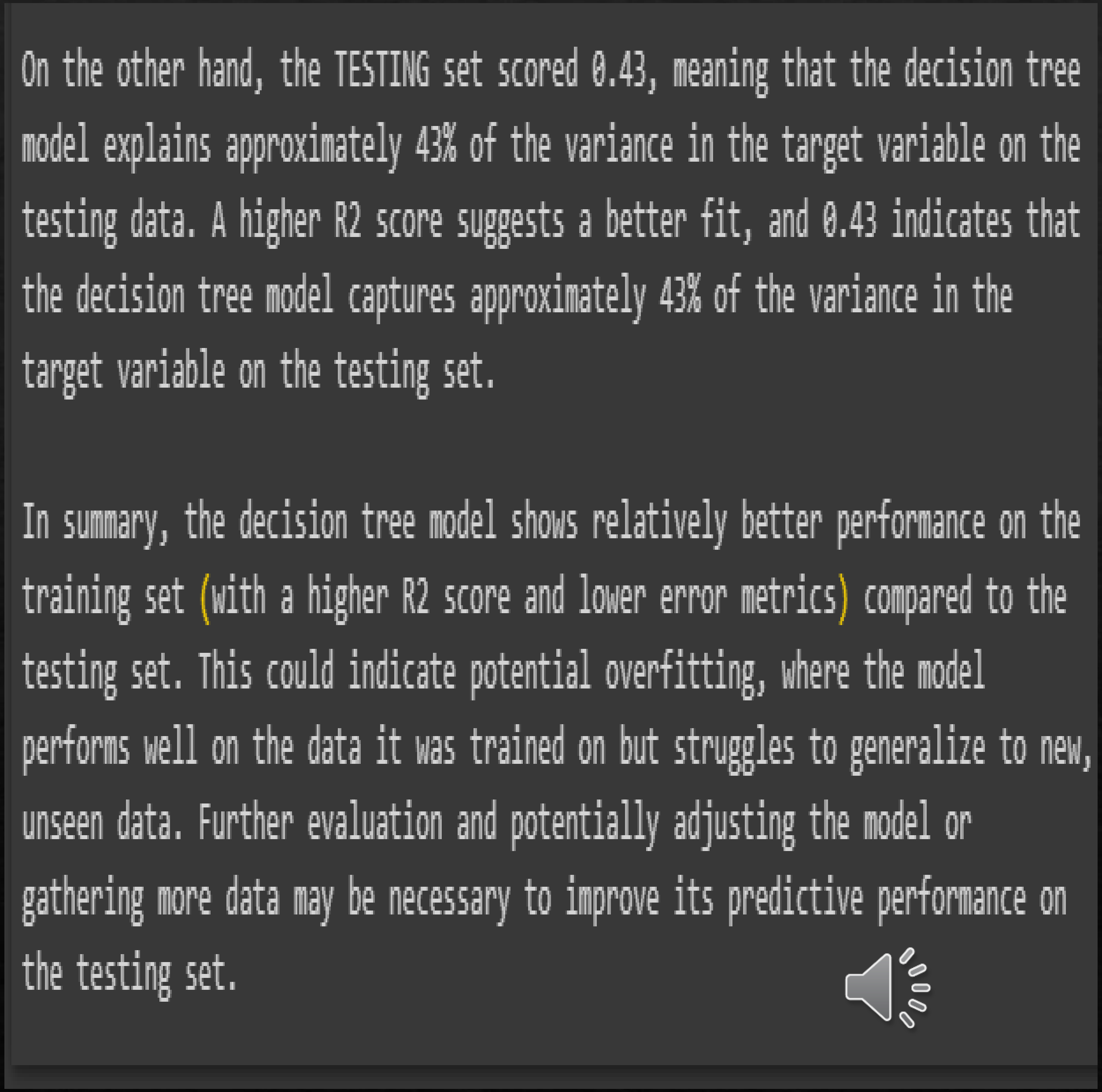
```
Decision Tree Regressor Training Score 0.78
Decision Tree Regressor Testing Score 0.43

 Decision Tree Training Evaluation

MAE: 496.80
MSE: 637,986.78
RMSE: 798.74
R2: 0.78

 Decision Tree Testing Evaluation

MAE: 852.15
MSE: 1,571,293.67
RMSE: 1,253.51
R2: 0.43
```

The DECISION TREE REGRESSOR MODEL scored .78 coefficient of determination (R2 score) on the TRAINING data indicating the model explains approximately 78% of the variance in the target variable. The model also scored 798.74 RMSE suggesting better predictive performance and a moderate error level.

2. DECISION TREE REGRESSOR MODEL PERFORMANCE



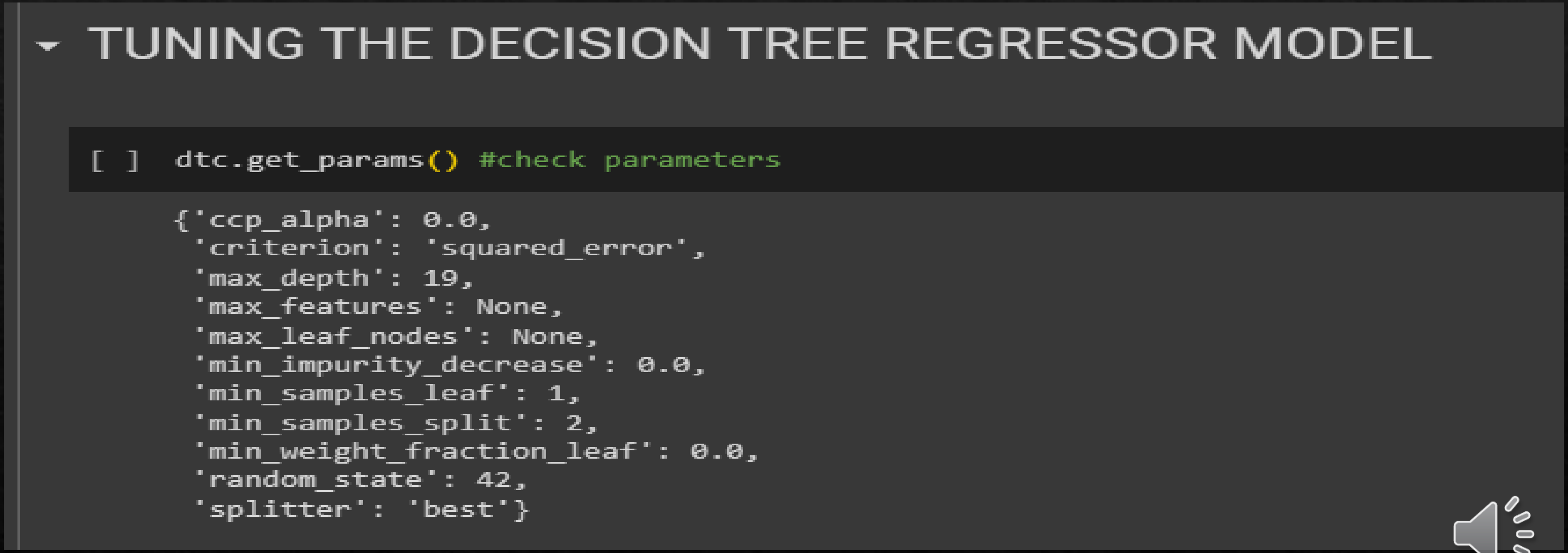
```
Decision Tree Regressor Training Score 0.78
Decision Tree Regressor Testing Score 0.43

 Decision Tree Training Evaluation

MAE: 496.80
MSE: 637,986.78
RMSE: 798.74
R2: 0.78

 Decision Tree Testing Evaluation

MAE: 852.15
MSE: 1,571,293.67
RMSE: 1,253.51
R2: 0.43
```



On the other hand, the TESTING set scored 0.43, meaning that the decision tree model explains approximately 43% of the variance in the target variable on the testing data. A higher R2 score suggests a better fit, and 0.43 indicates that the decision tree model captures approximately 43% of the variance in the target variable on the testing set.

In summary, the decision tree model shows relatively better performance on the training set (with a higher R2 score and lower error metrics) compared to the testing set. This could indicate potential overfitting, where the model performs well on the data it was trained on but struggles to generalize to new, unseen data. Further evaluation and potentially adjusting the model or gathering more data may be necessary to improve its predictive performance on the testing set.

LET'S IMPROVE MODEL PERFORMANCE THRU HYPERPARAMETER TUNING

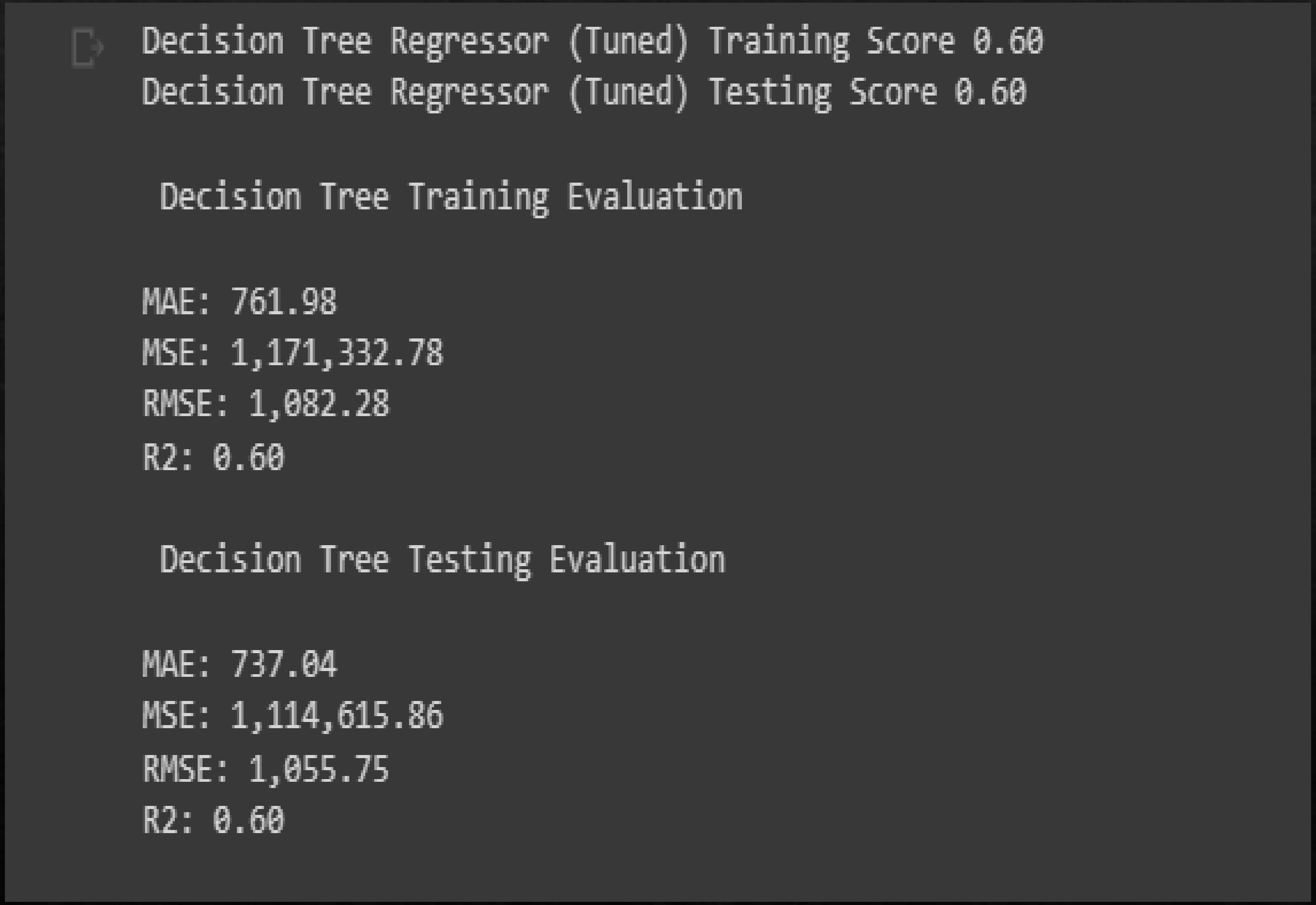


TUNING THE DECISION TREE REGRESSOR MODEL

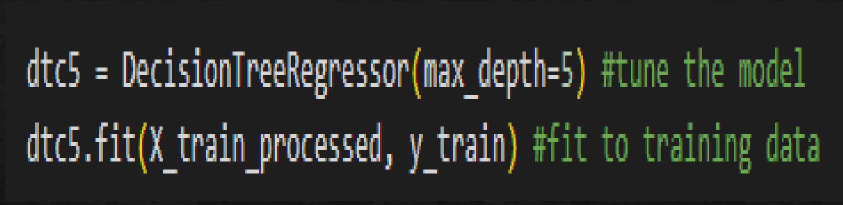
```
[ ]  dtc.get_params() #check parameters
```

```
{'ccp_alpha': 0.0,
 'criterion': 'squared_error',
 'max_depth': 19,
 'max_features': None,
 'max_leaf_nodes': None,
 'min_impurity_decrease': 0.0,
 'min_samples_leaf': 1,
 'min_samples_split': 2,
 'min_weight_fraction_leaf': 0.0,
 'random_state': 42,
 'splitter': 'best'}
```

DECISION TREE REGRESSOR (TUNED) MODEL PERFORMANCE



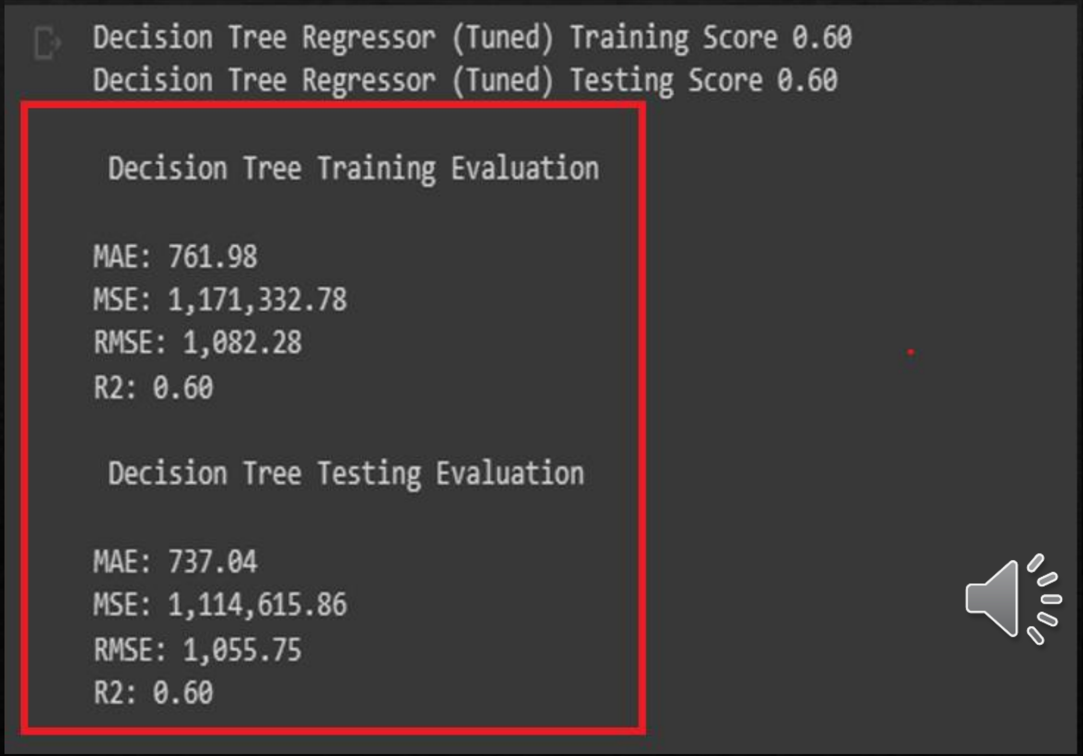
```
Decision Tree Regressor (Tuned) Training Score 0.60
Decision Tree Regressor (Tuned) Testing Score 0.60

 Decision Tree Training Evaluation

MAE: 761.98
MSE: 1,171,332.78
RMSE: 1,082.28
R2: 0.60

 Decision Tree Testing Evaluation

MAE: 737.04
MSE: 1,114,615.86
RMSE: 1,055.75
R2: 0.60
```



```
dtc5 = DecisionTreeRegressor(max_depth=5) #tune the model
dtc5.fit(X_train_processed, y_train) #fit to training data
```

After tuning the DECISION TREE REGRESSOR model, the R2 scores improve dramatically. R2 represents the proportion of the variance in the target variable that can be explained by the independent variables in the model. It ranges from 0 to 1, with 1 being a perfect fit. In both sets of scores, R2 is 0.60, which means that the model explains 60% of the variance in the target variable. It suggests that the model captures a moderate amount of information and has some predictive power, but there is still room for improvement. Compare to the pre-tuned model, this would be a better fit and provides a more predictive power and I would chose and recommend this model based on the evaluation performance and metrics.

DECISION TREE REGRESSOR (TUNED) MODEL PERFORMANCE



```
Decision Tree Regressor (Tuned) Training Score 0.60
Decision Tree Regressor (Tuned) Testing Score 0.60

 Decision Tree Training Evaluation

MAE: 761.98
MSE: 1,171,332.78
RMSE: 1,082.28
R2: 0.60

 Decision Tree Testing Evaluation

MAE: 737.04
MSE: 1,114,615.86
RMSE: 1,055.75
R2: 0.60
```

END