Data Analytics Capstone Project

Predicting Sales

Predictive Analysis

01

Problem Statement

Problem Statement

Activewear Inc., a leading online retailer of athletic apparel and accessories. Activewear Inc. is experiencing steady growth, but forecasting sales remains a challenge. Inaccurate sales forecasts lead to inefficiencies in marketing budget allocation across media mix channels.

Problem Statement

They need a more reliable method to predict future sales driven by their marketing tactics.

Methodology

Data Loading & EDA

Feature Engineering

Data Preprocessing

Model Selection & Training

Model Evaluation & Improvement

Conclusion

02

Data Loading, EDA & Data Preprocessing

Chart Explanation

The histogram plot of the TV shows there is sales of 8-12 after spending 50-250 on TV advert.

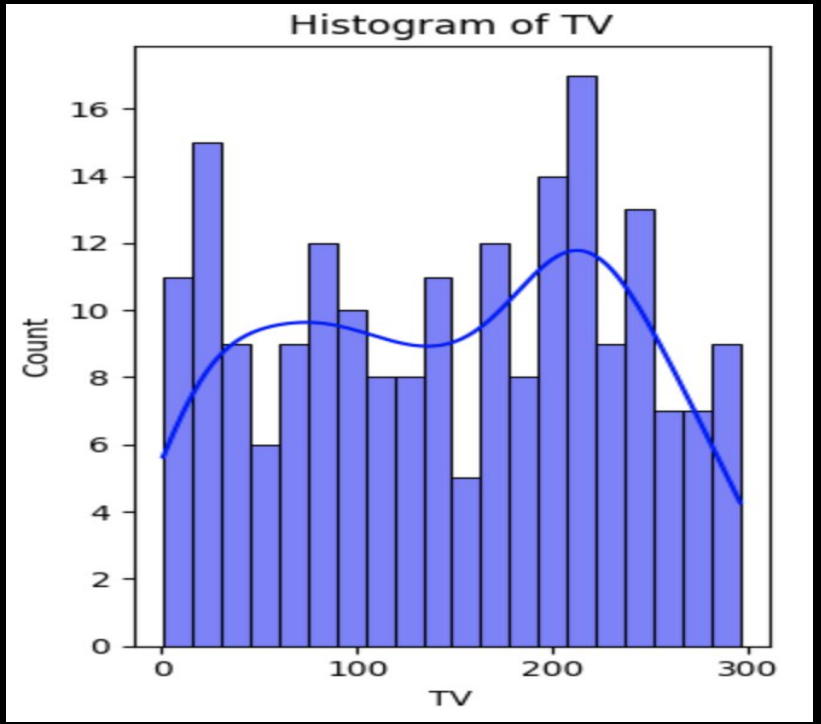


Chart Explanation

The histogram of the Radio shows there is sales of 8-10 after spending 5-42 on Radio advert.

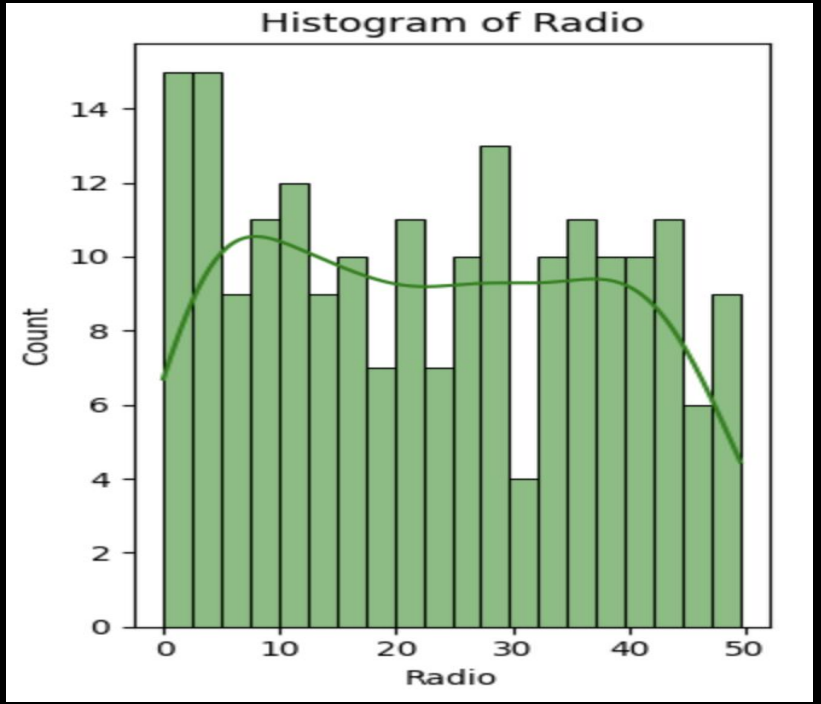


Chart Explanation

The histogram of the Newspaper shows there is sales of 10-21 after spending 5-50 on newspaper advert.

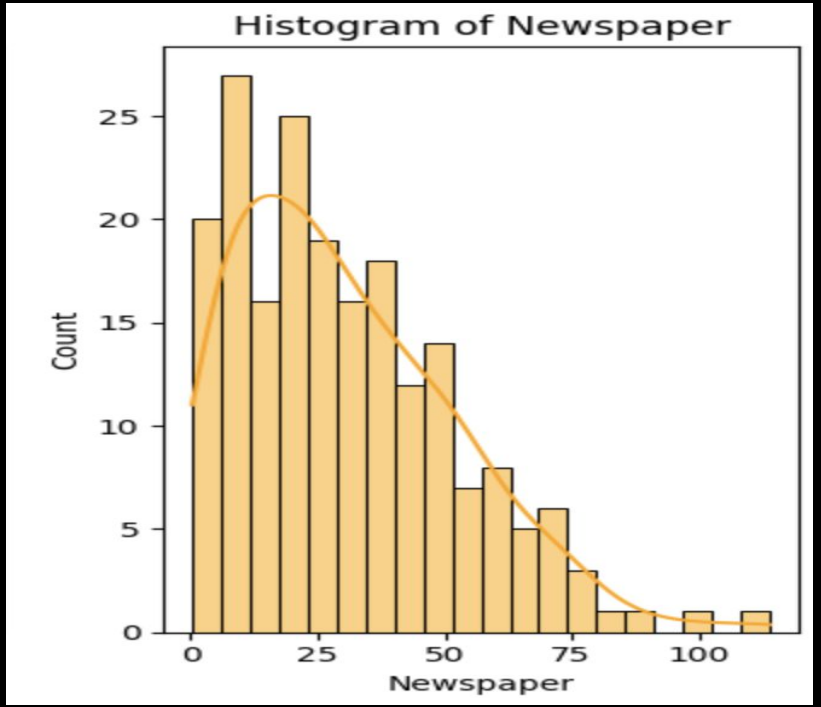


Chart Explanation

The boxplot of TV shows below 300 is being spent on TV advert but most of the times they spend between 75-210.

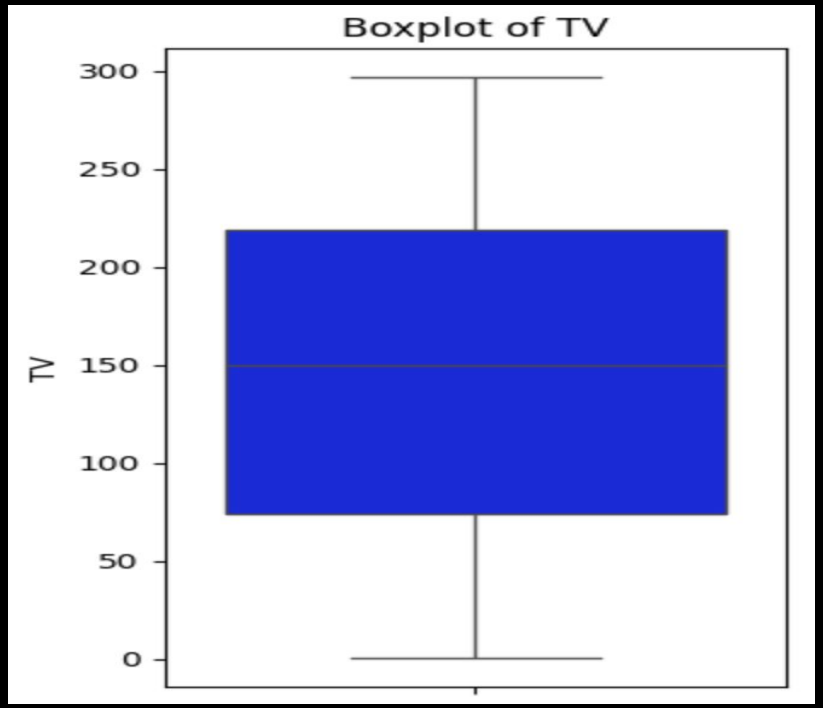


Chart Explanation

The boxplot of Radio shows below 50 is being spent on Radio advert and most of the time they spend between 10-35 on the advert.

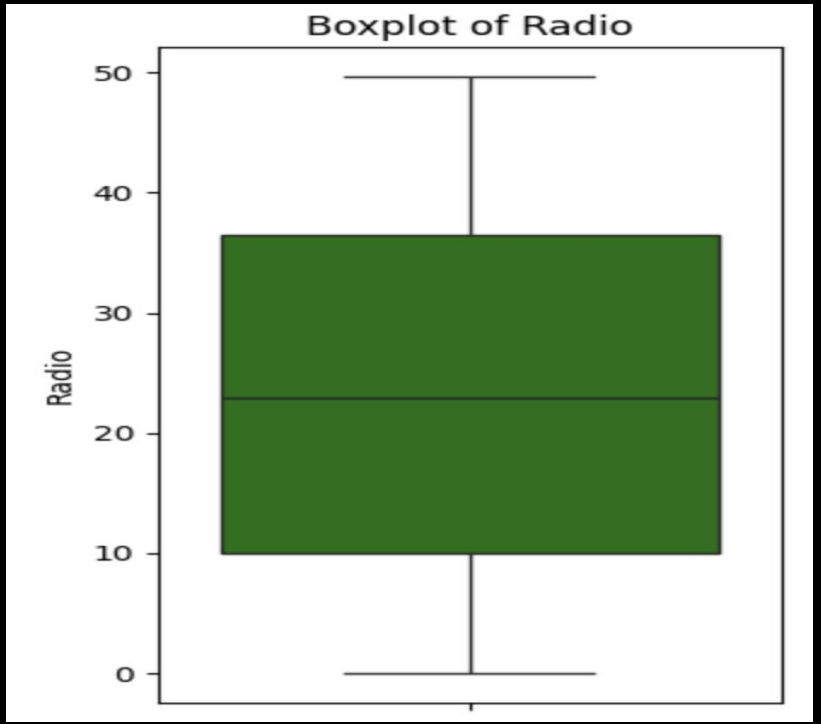
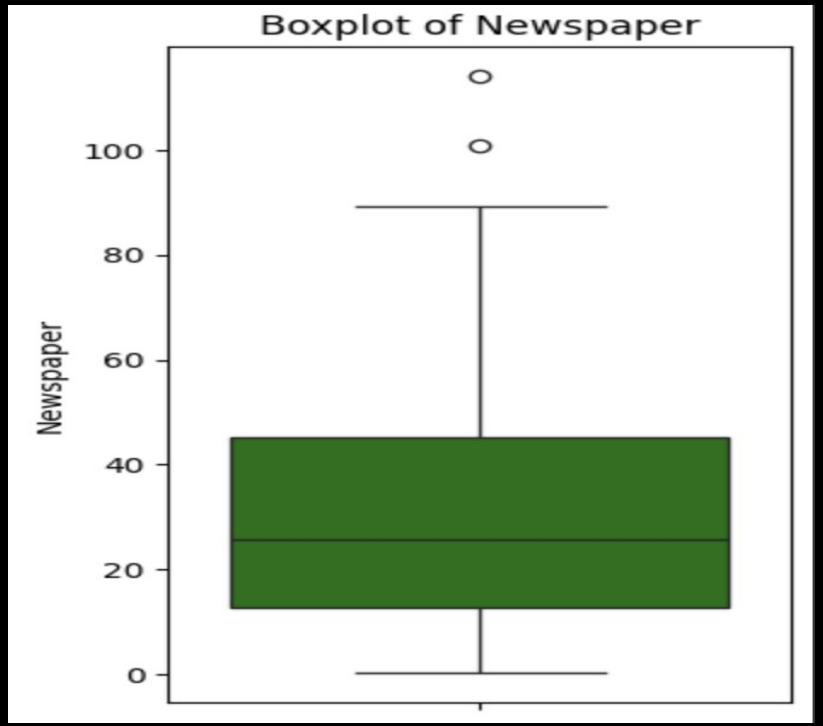


Chart Explanation

The boxplot of Newspaper shows mostly below 90 is being spent on Newspaper advert and most of the time they spend between 11-41 on the advert with some outliers between 100-120 which shows this has been spent on Newspaper advert on very few occasion.



03

Diagnostic Analysis

Chart Explanation

This scatter plot shows the higher the money spent on TV advert the higher the sales which shows more TV advert will give more sales.

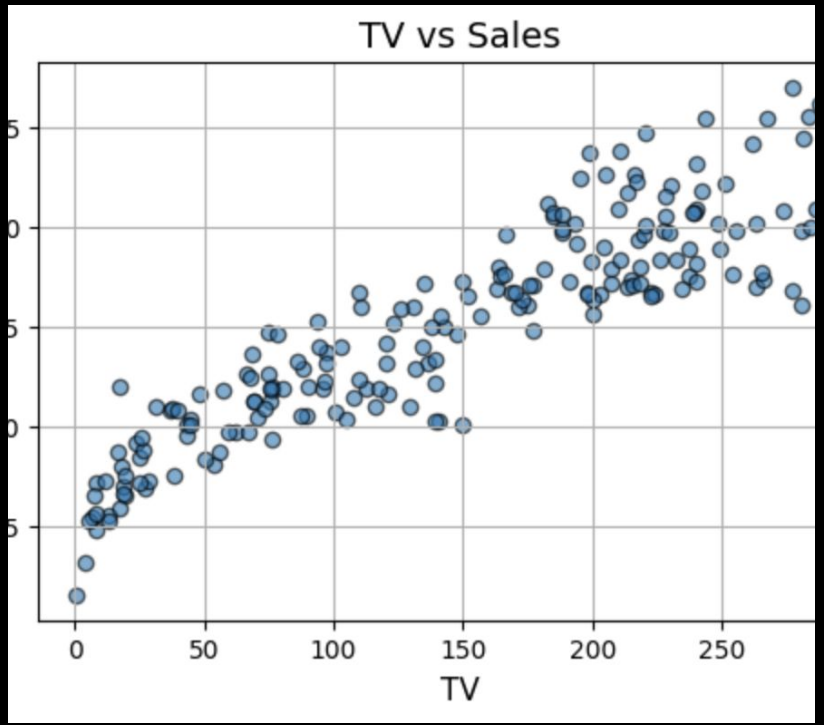


Chart Explanation

The scatter plot of the radio against sales shows the higher the money spent on radio advert does not guarantee higher sales though sometimes when between 40-50 is spent it increase sales but it is not frequent.

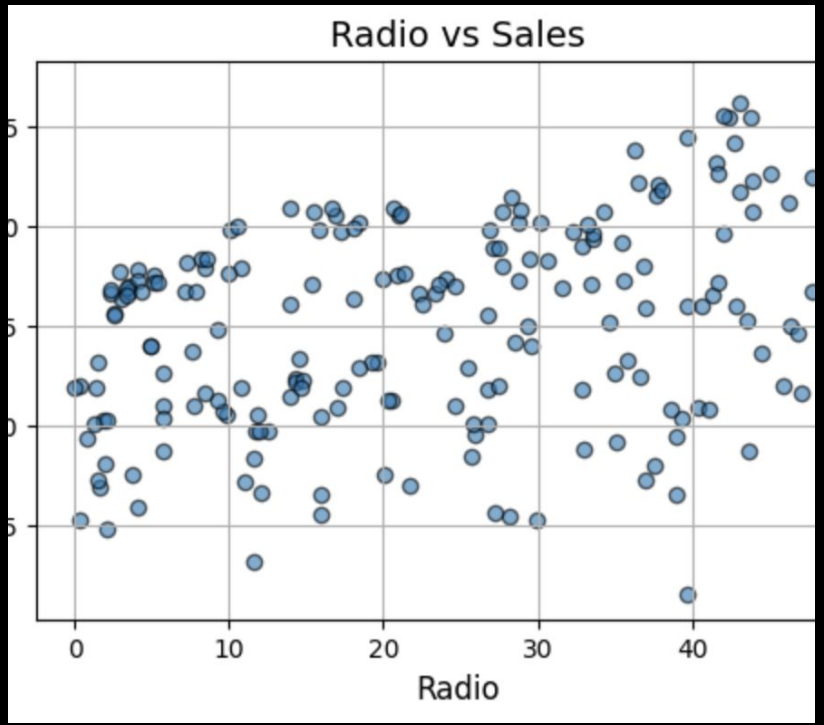


Chart Explanation

The scatter plot of the newspaper against sales shows the higher the money spent on newspaper advert does not guarantee higher sales and I will suggest the company should not spend more than 50 on newspaper advert as the higher money spent on the advert does not guarantee higher sales and where there have been higher sale is below 50 spent.

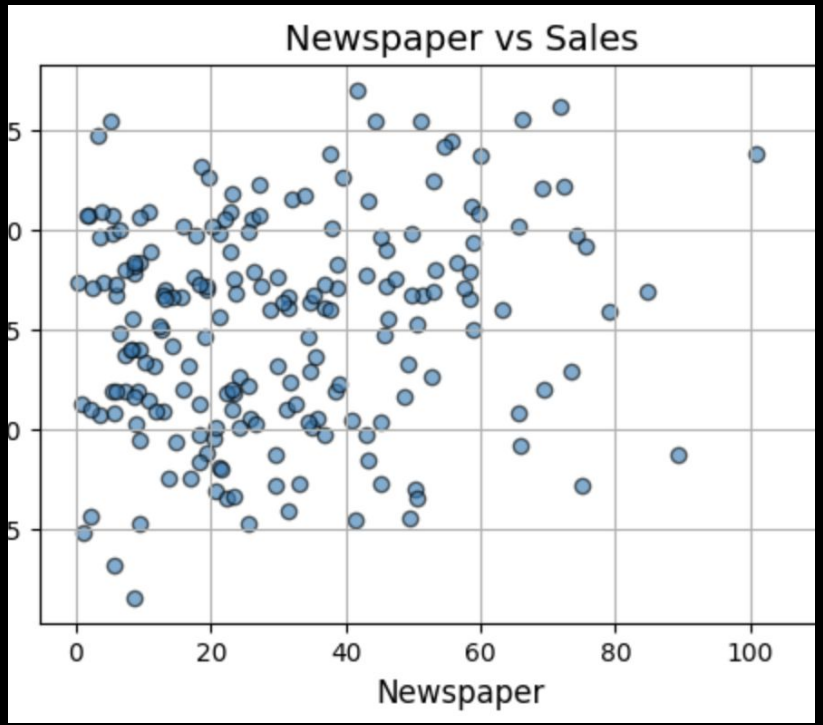
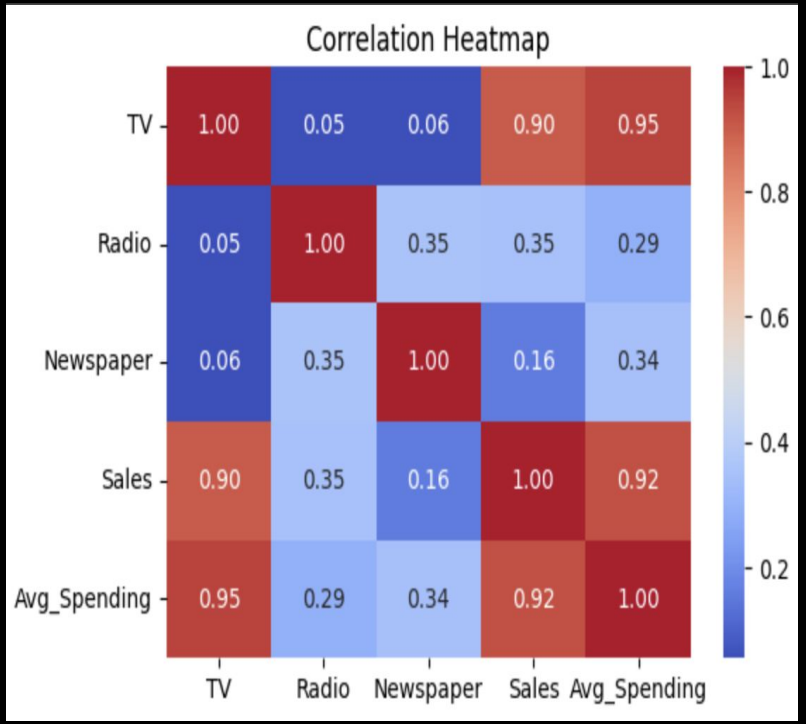


Chart Explanation

The TV and Average spending is positively correlated with sales as they both increase sales also increase. Radio also show a slightly positive correlation with sales and newspaper show a very weak positive correlation with sales.



04

Model Selection & Evaluation

Chart Explanation

The evaluation of the linear regression model showing the predicted values against the actual values for the parity plot show the data points closer to the red line show the model provides a decent fit of predicting the actual values but might not be perfect for all the data points.

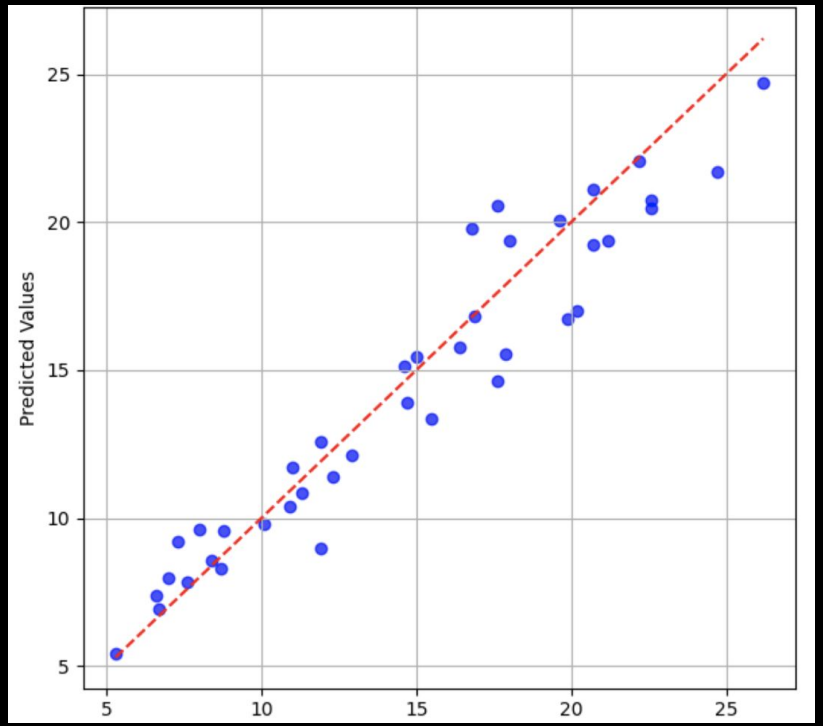


Chart Explanation

The residuals of the linear regression model appear randomly distributed around the red dashed line, suggesting the model captures the dataset well. There are some deviations, which might indicate areas where the model's predictions are less accurate.

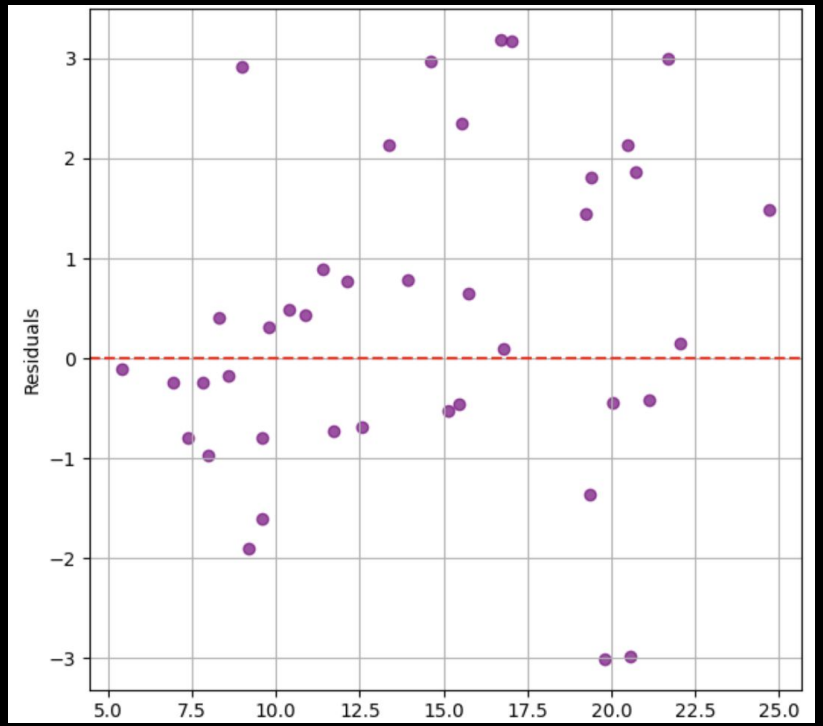
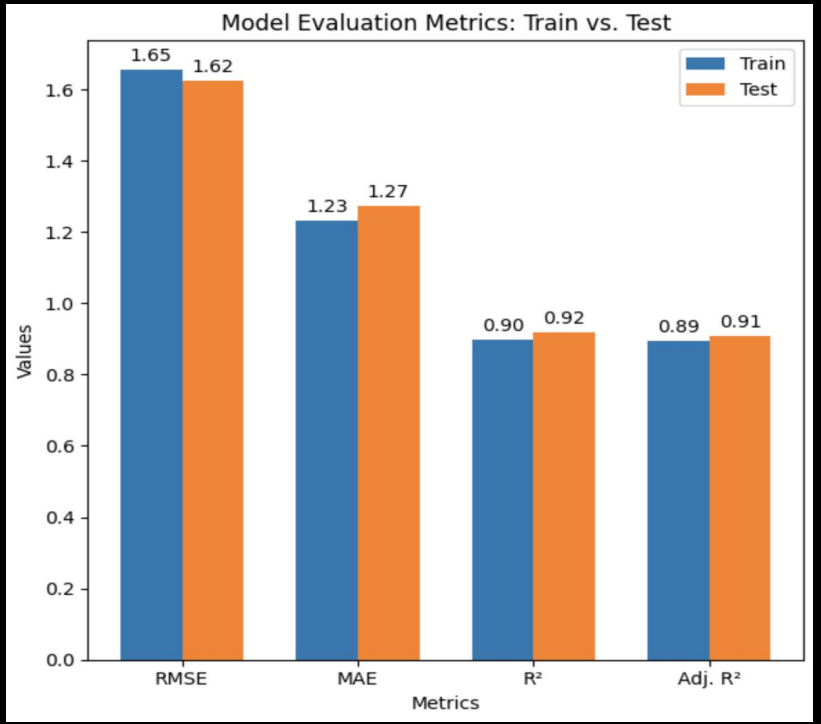


Chart Explanation

The metrics between the train and test sets suggest that the model is not overfitting or underfitting. The model can generalize well on unseen data with a low prediction of error in the linear regression test and train model performance.



While there are minor variances indicating no significant bias in the model's forecast, the random forest model evaluation show data points surrounding the red line demonstrate that the model's predictions are accurate. The model can make accurate predictions and has high generalisation skills.

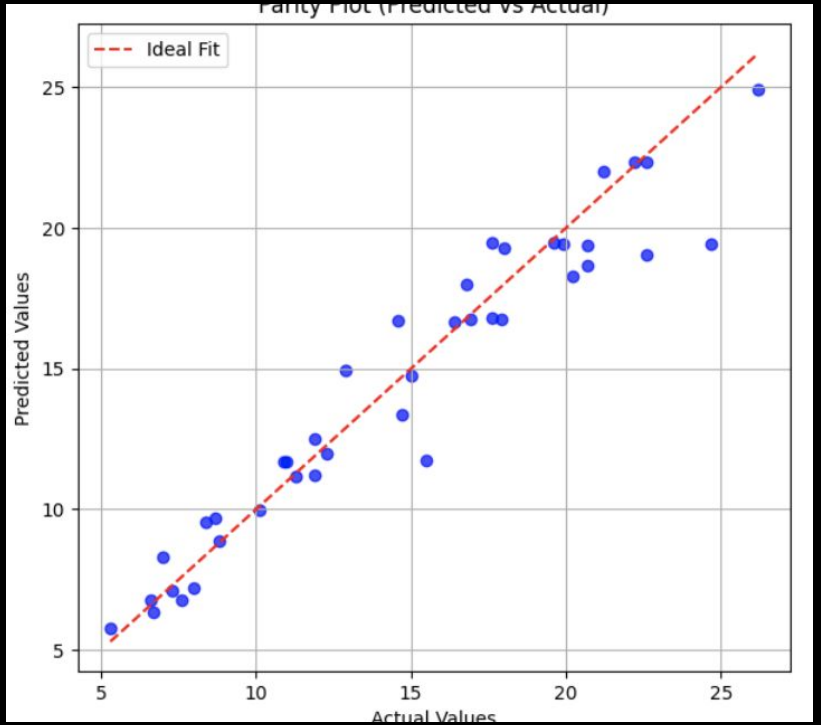
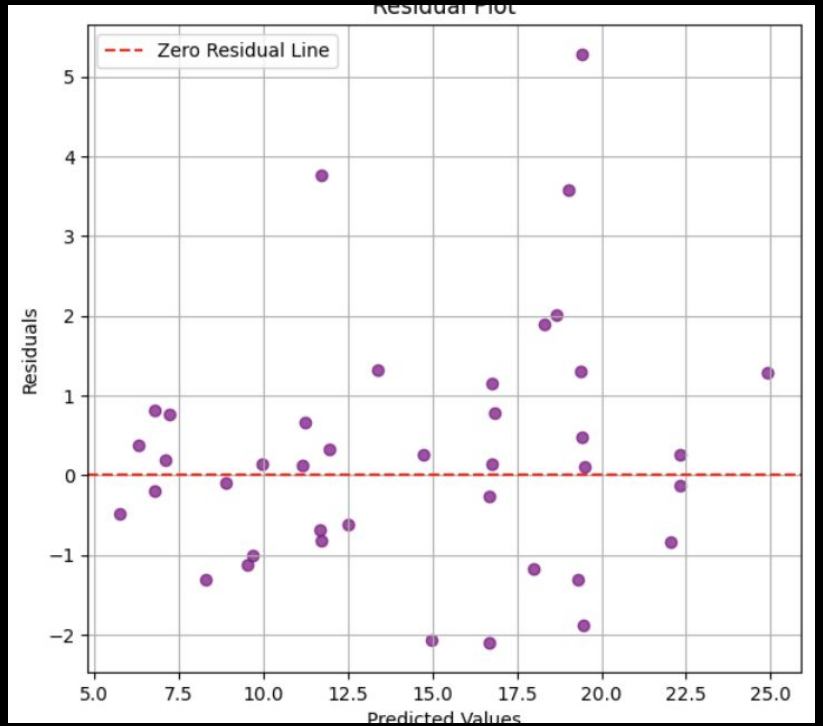


Chart Explanation

There are no clear patterns indicating the absence of systematic errors, and the residuals, which display the difference between the actual and predicted values, appear randomly distributed around the zero line. This suggests that the model captures the relationship in the data fairly well. Larger prediction errors for those data are suggested by a few spots with residuals farther from the zero line.



The model may be overfitted to the training data, capturing noise or certain patterns that do not transfer well to fresh data, as indicated by the significant difference between the training and testing RMSE/MAE values. The test R2 and Adjusted R2 values (0.93 and 0.92) show that the model still accounts for a significant amount of the variance in the test data, even in the face of overfitting.

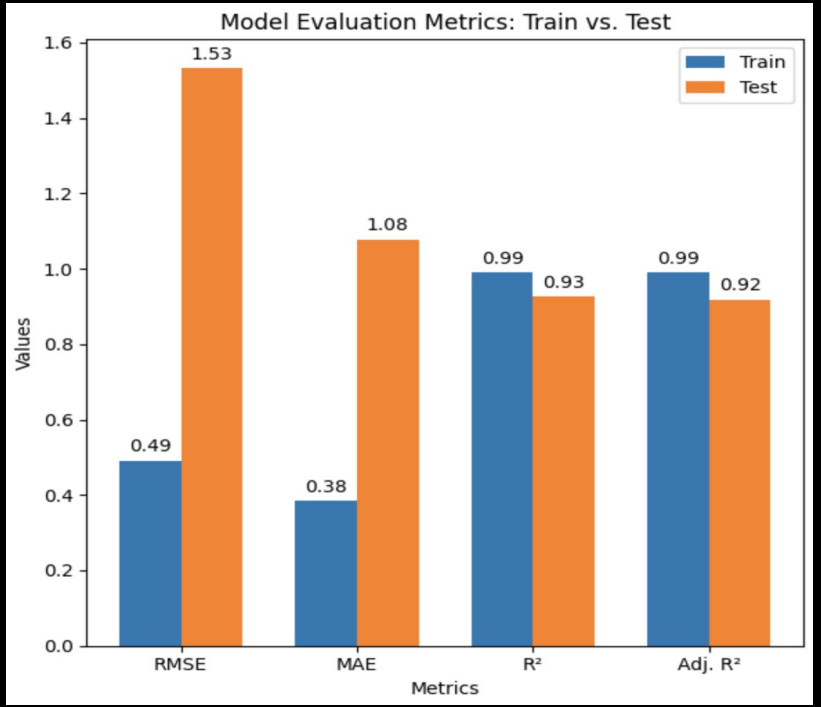
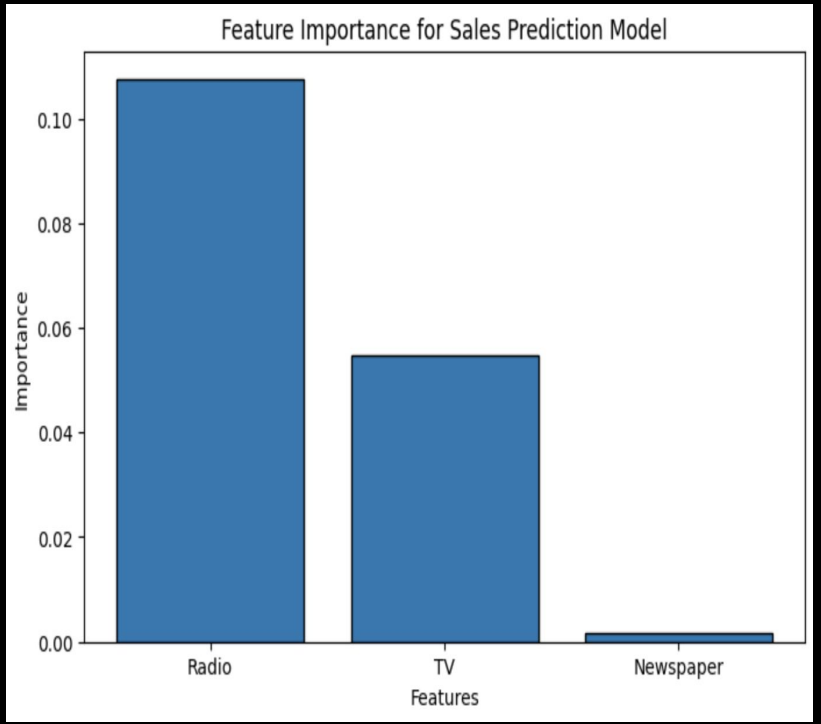


Chart Explanation

The importance of the feature in sales prediction shows radio have the highest impact on the model followed by TV and lastly newspaper which have the least importance in the model.



05

Conclusion

Conclusion

Tho the Random Forest regression model is overfitting on testing set, it demonstrated high accuracy on the training dataset while the Linear regression model does not overfit or underfit.

Conclusion.

The Linear regression model will be the best to predict sales as it will do well and predict on unseen data than the Random Forest regression that have the high accuracy on the training set.

THANK YOU

Project