SUMMARY

The retail sales regression project is a data analysis project that aims to determine the factors that influence retail sales in a given region. The project involves collecting and analyzing data on various factors such as store location, advertising expenditure, product prices, and promotions. The objective is to identify the relationship between the various factors and retail sales, and use this information to make informed business decisions.

The first step in the project is to collect data on the various factors that could influence retail sales. This data is typically gathered through surveys, interviews, and other data collection methods. Once the data is collected, it is cleaned and organized for analysis.

The next step is to perform a regression analysis on the data. Regression analysis is a statistical method that is used to identify the relationship between two or more variables. In this case, the goal is to identify the relationship between the various factors and retail sales.

The regression analysis involves fitting a regression model to the data. A regression model is a mathematical equation that describes the relationship between the dependent variable (in this case, retail sales) and one or more independent variables (such as store location, advertising expenditure, product prices, and promotions). The regression model is then used to make predictions about the relationship between the variables.

Linear modeling was used to fit the regression model to the data Lasso regression analysis is a regularization technique that can help prevent overfitting by penalizing the model for having too many features. The results of the Lasso regression analysis showed that the R2 and adjusted R2 values for this technique are higher than those for linear modeling, indicating that the model may be more accurate.

Decision tree analysis is a non-linear technique that can handle complex relationships between features and the target variable. The low MSE, MAE, and RMSE values for this technique indicate that the model's predictions are generally very close to the actual values. The high R2 and adjusted R2 values indicate that the model can explain a significant amount of the variability in the target variable.

Random forest analysis is an ensemble technique that combines multiple decision trees to improve the accuracy of predictions. The results of the random forest analysis showed that the high R2 value and lower MSE, RMSE, and MAE values for this technique indicate that the model's predictions are more accurate than those from the other techniques. The adjusted R2 value is also high, indicating that the model is not overfitting.

Overall, the results of the modeling techniques suggest that the random forest analysis is the most effective at improving the accuracy of our predictions, followed by decision tree analysis and lasso regression. Linear modeling provided some insights but did not result in as accurate predictions as the other techniques.

The results of this project have important implications for businesses that want to improve their sales performance. By collecting and analyzing data on factors such as store location, advertising expenditure, product prices, and promotions, businesses can make informed decisions about how to increase their sales and improve their profitability.

In addition to the regression analysis, other statistical techniques can also be used to analyze the data. Descriptive statistics can be used to summarize the data, such as the mean and standard deviation of the variables. Correlation analysis can also be used to identify the relationship between two or more variables.

It is important to note that the results of this project are specific to the dataset that was analyzed. The relationship between the various factors and retail sales may be different in other regions or for different types of products. Therefore, it It is important to note that these modeling techniques are only as good as the data that is used to train them. Therefore, it is important to collect high-quality data

that is representative of the factors that influence retail sales in the region. Additionally, businesses should continue to collect data and update their models over time to ensure that they remain accurate and relevant.