Retail Sales Prediction

Project Summary

SAI NARASIMHA CHARAN THULASI

The Retail Sales prediction data is provided with two csv files that is Rossman and Stores which contains different variables of information.It consists combinly about 19 variables which contain different kind of information

As the first step of the project we have performed data cleaning as well as data wrangling by merging both of tables in the next step we have performed Exploratory Data Analysis(EDA) in which we created different visualization charts to analyze the data we found that some interesting facts like sales are highly correlated with customers,there were more sales on Monday,probably because shops are closed on Sundays,it could be seen that the promo leads to more sales,more stores open on school holidays than on state holiday and hence had more sales than state holidays,on an average store type B had the highest sales,Highest average sales were seen with Assortment levels-b which is ‘extra’,82.1% sales are not affected and only 17.9% sales is affected because of school holiday etc.

In the next step I have done hypothetical testing
Hypothesis: Stores located closer to competition have significantly lower sales than stores located further away.Null hypothesis: There is no significant difference in sales between stores located closer to competition and stores located further away.
Alternative hypothesis: Stores located closer to competition have significantly lower s ales than stores located further away.To test this hypothesis, we can perform a twosample t-test between the sales of stores located within 10 km of competition and stores located further away. We can set a significance level of 0.05

After that I have performed feature engineering like filling missing values,handling of null values,handling columns,deleting unnecessary columns,feature processing,feature extracting,Outliers handling,feature selection.

In the last step most important step of my project that is model deployment. I have deployed two models first one is the linear regression and second one is the lasso regression final conclusion of both of the models -
The MSE and R2 score are commonly used evaluation metrics for regression models. In this case, the Linear Regression and Lasso Regression models have very similar per formance, with the Lasso Regression model having a slightly lower MSE and a slightl y higher R2 score.The mean squared error (MSE) measures the average squared differ ence between the predicted and actual values, where a lower MSE indicates better per formance.The Rsquared (R2) score measures the proportion of the variance in the dep endent variable that is predictable from the independent variables, where a higher R2 score indicates better performance.

Contributors Roles:

SAI NARASIMHA CHARAN THULASI

1.Data Loading:
2.Data handling
3.Handling missing values
4.Data exploration/Visualization
5.Outliers detection
6.Hypothesis Testing
7.Feature engineering
8.Model deployment

Github Repo link:

https://github.com/TSainarasimhacharan/Regression_Retail-Sales-Prediction.git