Abelardo Garcia, Sai Lakshman Ethakatla, Ramyanth Nakshathram

04/19/2022

CIS 5364 Data Warehousing

Homework 5

SAC username: Ramyanth Nakshathram

**Part 1**

* 1. Acquired the Data.

Dataset name: Global Bike Sales

A screenshot of a computer

Description automatically generated

* 1. Changed the OrderNumber, OrderItem, Year, Month, Day, Customer, Latitude and Longitude to Dimensions.

Graphical user interface, application

Description automatically generated

* 1. Validated the dataset and saved.

**Part 2 – Creating the Regression Analysis Model**

Data Model name: GB Sales Regression Analysis

1. Created a predictive scenario with the data given and selected Predictive target as SalesQuantity.

Graphical user interface, text, application, email

Description automatically generated

1. Excluded all influencers except for the following: Month, CustDescr, City, SalesOrg, ProdDescr, CatDescr, SalesQuantity

Graphical user interface, table

Description automatically generated

1. Trained the data.

**Part 3 – Examine the results**

**Question 1 on page 4**

**Question 3a:** Which variable contributes the most to the regression model?

**Answer:** CatDescr contributes the most to the regression model.

Graphical user interface

Description automatically generated with low confidence

Chart, waterfall chart

Description automatically generated

**Part 4 – Continue the Analysis**

* 1. Created another predictive model on the predictive scenario with new settings

Graphical user interface, text, application

Description automatically generated

* 1. Trained the data. Model 2 analysis below.

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

* 1. Model Comparison is shown as follows

Graphical user interface, application

Description automatically generated

**Question 3 on page 6**

**Question 3b:** Now which variable(s) contribute the most to the regression model? Does this make sense? Are there variables that should be eliminated from the model? Explain.

**Answer:** In the Model 2, CostsUSD variable contributes the most to the regression model. Yes, this makes sense because in the general business scenario, the sales quantity is majorly influenced by the cost of the product. Looking at the above model comparison, it is suggestible not to eliminate any influencers from the model, since the Root Mean Square Error (RMSE) is low as compared to the scenario where influencers were eliminated. Also the Prediction Confidence in Model 2 is higher than that of Model 1.