Forecasting Sales and Inventory Demand by Country and Product Popularity

of a Toy Company’s Product Line

By Matthew Whitcombe

Student ID: 501074750

Supervisor: N/A

Data analytics is relied on in many organizations to compile and analyze the organization’s sales data in order to better predict the future needs of the organization. If a company does not properly prepare for the future, they may experience problems such as inventory shortages, inaccurate demand planning, having insufficient funds available to pay manufactures and/or distributors. This project will rely on a sales sample dataset from a toy company which was downloaded from Kaggle. The dataset provides sales information between January 2003 – May 2005. This project will integrate the Machine Learning techniques of Data Mining & Knowledge Discovery, and, Classification & Regression in order to forecast the potential future sales and inventory demands per country, and product popularity per country in the chosen dataset sample. The main questions that the project aims to answer are as follows: for the period 2006-2008, what will be the future sales and inventory demands per country, which products will be popular in the countries the company ships to, which country will contribute the most and the least to the organization’s overall sales.

The useful fields to help conduct the analysis are product types, product price, product purchase, shipped to location. R will be the programming environment where the analysis of the dataset will be conducted. The first part of the data analysis will be done using data mining. This will help clean the dataset and provide useful insights about the raw dataset. The Class Function and the Unique function will help give a high-level overview of the variables in the dataset. In the event there are some blank fields and inconsistencies in the dataset, part of the data cleaning will involve filling in the blanks, and making sure the data fields have consistent formatting. For the visualization portion of the data mining section the ggplot2 library, the geom\_histrogram, and geo\_boxplot graphs will be useful to visually display different variable combinations. I will refer to the ca.Tools library to split into a training set and a testing set. This will allow me to build the data model and then test it.

For the clustering portion of the analysis, I will use the following tools: Scatter Plots, Data Normalization, Euclidean Distance, Cluster Dendrogram, Non-Hierarchical k-means Clustering. These tools will help find patterns and links between the variables within the dataset. For the Regression part of the analysis, Pearson correlation is useful to find the correlation between different variables within the dataset. Also, for Regression, the Linear Model function is great at identifying Residuals, the Estimate, Std. Error, t-value, and p-value of the intercept.

It is anticipated that these techniques and tools will provide the level of analysis needed to answer the questions set out above. The analysis will forecast the Toy company’s future sales for 2006-2008, the company’s future sales by country in order to determine which country leads the company’s profitability, inventory demands per country, and product popularity by country.

**References**

Dataset source on Kaggle: <https://www.kaggle.com/datasets/kyanyoga/sample-sales-data?resource=download>

Roldán, María Carina, modified by Gus Segura (2014, June). Sample Sales Data; Version 1. [Retrieved Sept 13, 2022] from <https://www.kaggle.com/datasets/kyanyoga/sample-sales-data?resource=download>