COS10022 PROJECT 1: INTRODUCTION TO DATA SCIENCE

HUY NGOC NGUYEN - 103582239
HAI NGUYEN PHAM - 102953559
LE ANH MINH TRINH - 102953766

February 28, 2021

1. ABSTRACT

This report aims to establish a sales forecast for several items based on their characteristics and the selling location of those items. By outlining independent and dependent variables and classifying each of them into categorical or numerical data, this study pinpoints the necessary factors that impact the sales of a product. The use of the KNIME Analytics Platform helps clean out missing values from the dataset as well as splitting the dataset into training and test data. KNIME provides several statistical models to display product sales or location features of the dataset. In this report, a histogram and a pie chart are two options to display some characteristics of products and location features. However, these models are not actual prediction models. This is followed by a discussion regarding whether the output variable should be continuous or categorical as well as its advantages and disadvantages. Finally, from the ability to predict the sales of a product, businesses can consider a few business values that can guide them with a sense of security to make better decisions and help achieve their company goals.

2. INTRODUCTION

In today's world, every record of financial sales is monitored by worksheets and databases. Retail businesses such as supermarkets use statistical models and other data management tools to analyze their sales trends. From there, they can predict customer demands and organize better inventory management, especially when online shopping is promoting its way into the market. In this task, a dataset from BigMart containing 8,523 sales records for 1,559 items is categorized under 12 different attributes, with each concerning the product or selling location features. Although there are some missing values in the dataset, it remains possible to build a prediction model for outlet sales. The algorithm used in this report involves the KNIME Analytics Platform to generate statistical models as well as the training and testing datasets.

3. HYPOTHESES AND TESTING

Before forming hypotheses to predict the sales of a product, the classification of data variables identifies the Item_Outlet_Sales as the dependent variable because the variable contains the expected outcome of product sales. Other independent variables include Item_MRP, Item_Type, Item_Visibility, Outlet_Type, Outlet_Location_Type, and Outlet_Size. Out of the twelve variables, Item_Outlet_Sales, Item_MRP, Item_Visibility, and Item_Weight are continuous data types. Categorical data types are split into nominal data types, which include Item_Identifier, Outlet_Identifier, Item_Type, and Outlet_Establishment_Year, and ordinal data types, which include Outlet_Type, Outlet_Location_Type, Outlet_Size, and Item_Fat_Content.

When taking a closer look into each independent variable, there exists a connection between each variable and the expected outlet sales. For example, the maximum retail price of an item is the highest price of the item a customer can buy, which directly contributes to the item outlet sales (Thakur 2015). The outlet types describe where items are sold, which includes supermarkets and grocery stores. The sales of an item vary based on supermarket tiers and grocery stores. While supermarkets cover a wide range of food choices, grocery stores tend to categorize their products for a local customer base (Campbell 2020). The outlet location types use tiers to determine the size of a city, meaning that the more concentrated the city is, the more attracted people will be, leading to more competition around the area, which impacts the sale of multiple stores. For instance, in the Greater Cincinnati Area, while Kmart is more accessible in terms of shorter driving time from all tracts than Walmart, Kmart must compete with different stores, which heavily impacts their outlet sales (Li & Liu 2012, p. 596-597). The variable concerning outlet size illustrates the ground area covered by the supermarket or the grocery store. The bigger the outlet is, the more sales the supermarket will make, or the more products will be displayed (Campbell 2020). The item type is also a factor in contributing to sales to guide sellers towards product trends and customer demands. A comparison of customers' shopping lists in Tesco revealed that bananas and strawberries are among the best-selling products (Bellos 2015). This suggests a concentration on household fruits and helps sellers manage their inventory and decide whether to promote other products for better sales. Item visibility determines how much display area the item occupies in the store, and how the shop organizes these items to attract customers' attention, playing a role in the outlet sales. Creating a focal point to identify products easier, especially when they are almost sold out, allows customers to optimize their route and keep them always interested (Musa et al 2014).

A few hypotheses can be formed from the connections between product and location features and product sales to build a prediction model. Items with a higher maximum retail price will generate more sales since customers are not troubled by prices beyond the limit and can negotiate lower prices with sellers. Items located in supermarkets are likely to sell more

than grocery stores because supermarkets have a wider product range and higher product availability than grocery stores. If the size of the store is categorized as 'high', then the store will produce more sales because customers have options to choose from numerous items. If the stores were established earlier, then the store might generate more sales than newly established stores as newly established stores have not had the time to produce as many sales. If the item is frequently displayed more than others, then it is likely to be purchased more than other items because customers are aware of the item when they enter the shop.

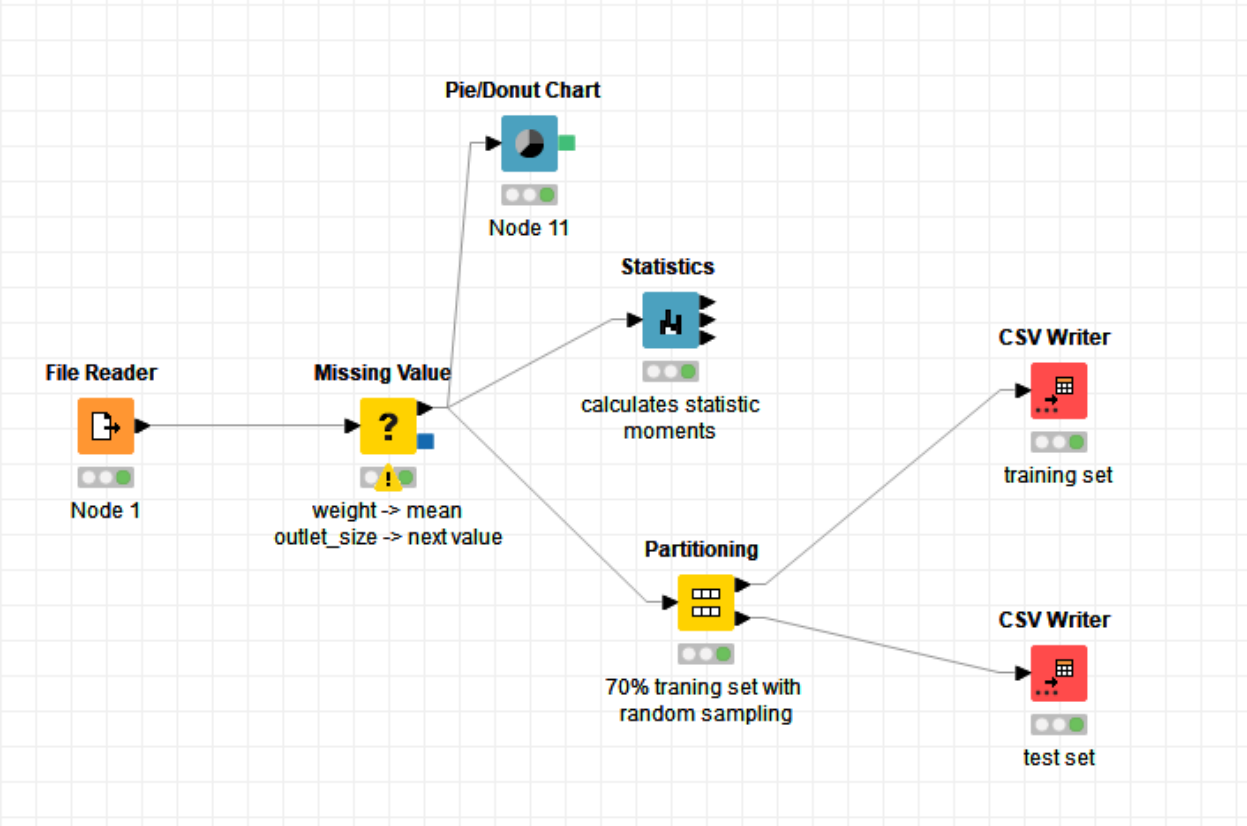

Figure 1: Filtering and partitioning the dataset in KNIME

For the pre-processing workflow, the BigMart dataset is imported as a CSV file into the KNIME Analytics Platform through the 'File Reader' node (Scott 2020). The 'Missing Value' node allows null or incorrect variables to be assigned with a value (Knime 2020). In this case, some products from the dataset do not have their weight included, and some outlet location sizes are not labeled. Any missing values of the product weight are replaced with the mean of labeled values in the same column. Similarly, any missing values of the size of outlet locations are filled with the next labeled value in its column. Other variables that do not have missing values are excluded from this node (Knime 2020). Next, to create meaningful categorical data from existing numerical data, the 'Statistics' and 'Donut Chart' nodes are some of the options utilized in this assignment. Specifically, the 'Statistics' node computes data points such as mean, median, and standard deviation and illustrates histograms for each

possible variable (Knime 2020). The simple histogram below shows the number of outlets established from 1985 to 2009, which seemingly presents more outlet establishments between 1985 and 2009 than the two extremes. On the other hand, the 'Donut Chart' node creates a donut chart based on a column input (Knime 2020). In this case, the donut chart divides product types into slices, with fruits and vegetables and snack foods as the more abundant products. The 'Partitioning' node splits the dataset formed from the 'Missing Value' node into two datasets, a training dataset, and a test dataset (Knime 2020). For this assignment, the split method is constructed by randomly selecting rows from the original dataset, in which 70% of that dataset goes into the training data, leaving the other 30% for the test data. The training dataset is then extracted to a CSV file by connecting the upper outlet port of the 'Partitioning' node with the 'CSV Writer' node (Knime 2020). After that, to retrieve the test dataset without the prediction target attribute, it is required to configure the 'Missing Value' node and remove every column customization done before. Then, using a fixed seed when randomly generating the training dataset in the 'Partitioning' node, the test dataset will be created. To extract the test dataset into a CSV file, the lower outlet port is linked with a different "CSV Writer" node.

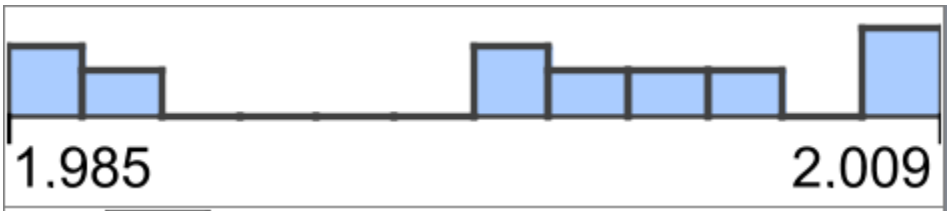

Figure 2: Outlet establishment year between 1985 and 2009

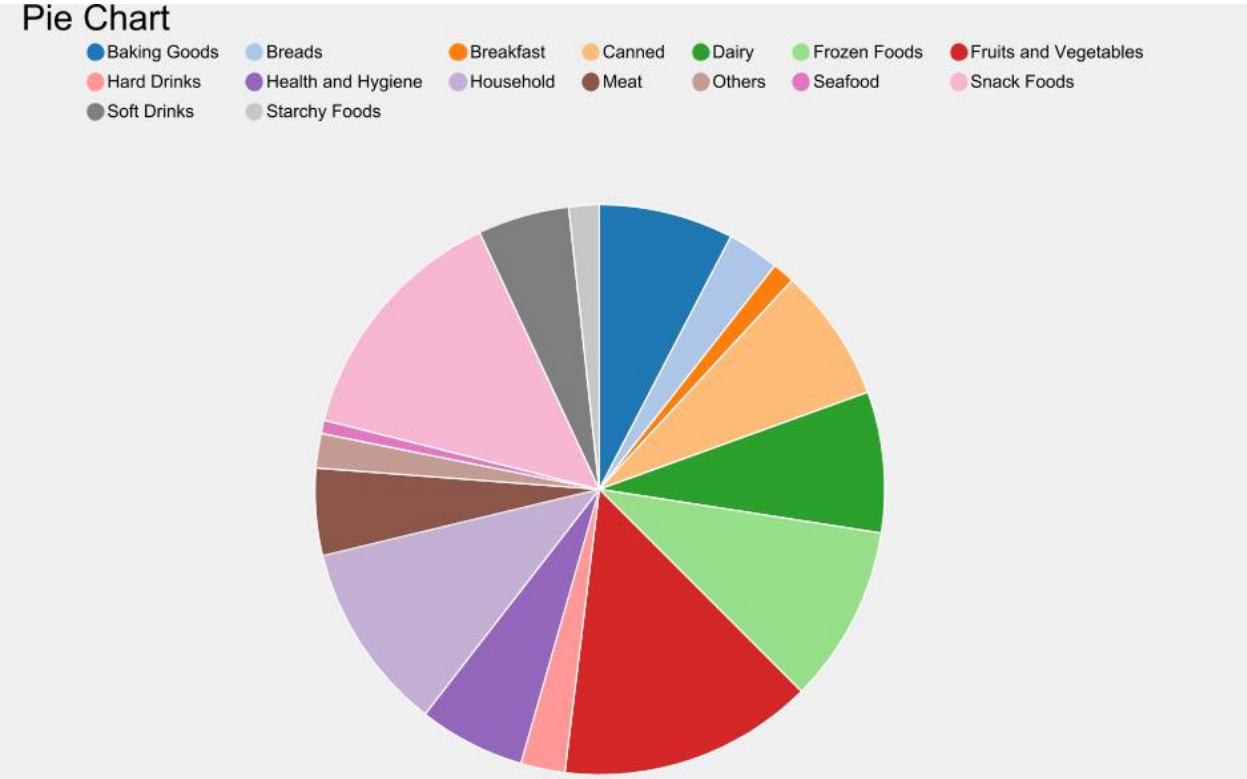

Figure 3: A donut chart of the types of products in the BigMart dataset

4. OUTPUT VARIABLE: CATEGORICAL OR CONTINUOUS?

Based on this specific assignment, the output variable is the outlet sales of each product in each store. Outlet sales are generally considered a continuous variable because they can have an infinite number of values between an interval and can be expressed as any value. For example, the graph of annual sales of Christmas decorations will likely rise by early December and fall by mid-January because the sales of the product change continuously in response to customer demand. Therefore, if the output variable is continuous, it is easier to produce accurate predictions as the data points are plentiful and obtained objectively. Furthermore, the collection of data is relatively easy to analyze using statistical models or relational databases. Understanding the analyzing techniques will simplify the problem and conduct prediction and actual trends efficiently, saving time and resources in the meantime. However, the research of a continuous output variable may provide data that cannot explain complex problems and can lead to false conclusions. For instance, during the data collection stage, the method does not consider any social or competitive phenomena between the outlets or stores, which leads to unanswered questions or why or how a store makes more profit than its competitors. Moreover, the data might be hard to replicate in future predictions as the data gathered is only reliable then. When looking into a long-term viewpoint, it would be preferable not to draw any conclusions from past research.

5. WHAT BUSINESS VALUES CAN WE EXTRACT FROM AUTOMATICALLY PREDICTING THE SALES OF A PRODUCT?

Knowing the trajectory of a business's revenues and profits can give a firm understanding of expectations and budget optimization. This assists the business in approaching its business values and paves a path to discovering more business strategies for greater success. One of the business values gained from the automatic prediction of product sales is strategic planning. Being able to predict the number of sales of a product allows businesses to plan accurately based on reliable statistics in the past. The more accurate the prediction is, the lesser the risks and the more concrete and efficient choices in planning can be made. If the opposite occurs, the profitability of the business will be affected negatively. For example, American company Blockbuster faced a challenge regarding how they must avoid under-forecast or over-forecast in 2012 (Chung et al 2012, p. 852). Both extremes of the issue would result in sales loss and put the company in a bad position while coming up against other competitors (Chung et al 2012, p. 852). If they were to predict the number of products to stock per weekly or monthly demand, their revenue would not decrease considerably in contrast to previous years (Chung 2012 et al, p. 872). Similarly, if Blockbuster had achieved success in sales forecasting, it would also have helped to reduce the cost of purchasing items to stock up, since it would be possible to estimate the number of sales in the following days and weeks. At the same time, the business could choose to buy fewer items that are considered to not likely be bought, thus leading to saving more money for the company. Because there were roughly 55 new releases at Blockbuster weekly, monitoring sales, and strategic planning could have significantly reduced their operating costs (Chung et al 2012, p. 852). Lastly, the automatic prediction of product sales would bring better customer satisfaction as it allows businesses to satisfy their customers more efficiently and more often, by purchasing the right products and delivering them at the right time, to the right demographics. Better forecasting would further enhance the accuracy, and frequency of satisfying their wants and needs and increase their loyalty to the business. For instance, Lucent Technologies, a telecommunications equipment company, succeeded in communicating customer requirements and building accurate forecast models by conducting a sale forecasting audit and adapting to challenges they encountered directly (Moon et al, p. 19, p. 21-22). They managed their inventory, logistics, and production costs accordingly based on their customer demands for their products as well as strategically planned to minimize their supply chain costs (Moon et al, p. 19, pg. 24).

6. CONCLUSION

The brief analysis of the BigMart Sales Dataset above has been assessed through variables and possible hypotheses regarding the prediction of the sales of a product. The KNIME Analytics Platform has cleaned and organized the dataset into training and test

datasets, with one serving as a prediction model and the other as a baseline to predict the model accuracy. The histogram and the donut chart are two examples of possibilities that are taken into consideration in future research on predicting the actual model. As businesses understand their customers' demands and strategically plan to avoid revenue loss based on prediction models and machine algorithms, they can manage their supply effectively and respond accordingly to every customer desire in their supermarket.

REFERNCES

Bellos, A, 2015, 'Top bananas: shopping list survey reveals bananas are number 1 supermarket impulse buy', *The Guardian,* 5 June 2015, viewed 27 February 2021, https://www.theguardian.com/science/alexs-adventures-in-numberland2015/jun/05/top-bananas-shopping-list-survey-reveals-bananas-are-number-1-supermarket-impulse-buy

Campbell, J, 2020, 'What's the Difference Between a Grocery Store & a Supermarket?', *The Grocery Store Guy,* 10 June 2020, viewed 27 February 2021, https://thegrocerystoreguy.com/whats-the-difference-between-a-grocery-store-a-supermarket/#:~:text=The%20terms%20supermarket%20%26%20grocery%20store,of%20food%20or%20targeted%20demographic.

Chung, C, Niu, S-C., Sriskandarajah, C, 2012, 'A Sales Forecast Model for Short-Life-Cycle Products: New Releases at Blockbuster', *Production and Operations Management,* vol. 21, no. 5, pp. 851-873.

Knime, 2020, 'CSV Writer', *Knime hub,* viewed 27 February 2021, https://kni.me/n/mPMx3pzMHtI9Qk5G

Knime, 2020, 'Missing Value', *Knime hub,* viewed 27 February 2021, https://kni.me/n/uVmaGQkzUOFCTqUe

Knime, 2020, 'Partitioning', *Knime hub,* viewed 27 February 2021, https://kni.me/n/T0AhxSe0Yi42rQK8

Knime, 2020, 'Statistics', *Knime hub,* viewed 27 February 2021, https://kni.me/n/7N4LxTHoPJrHN5lT

Li, Y, Liu, L, 2012, 'Assessing the impact of retail location on store performance: A comparison of Wal-Mart and Kmart stores in Cincinnati', *Applied Geography,* vol. 32, no. 2, pp. 591-600

Moon, M, Mentzer, J, Thomas, D.E, 2000, 'Customer Demand Planning at Lucent Technologies', *Industrial Marketing Management,* vol. 29, no. 1, pp. 19-26.

Musa, A, Gunasekaran, A, Yusuf, Y, 2014, 'Supply chain product visibility: Methods, systems, and impacts', *Expert Systems and Applications,* vol. 41, no. 1, pp. 176-194.

Scott, F, 2020, 'Read a CSV File', *Knime hub,* 10 January 2020, viewed 27 February 2021, https://hub.knime.com/knime/spaces/Examples/latest/01_Data_Access/01_Common_Type_Files/02_Read_a_CSV_file

Thakur, M, 2015, 'Full Form of MRP', *Wall Street Mojo,* 28 July 2015, viewed 27 February 2021, https://www.wallstreetmojo.com/full-form-of-mrp/