Enid Roman

Data 622 Machine Learning and Big Data

Exploratory Analysis and Essay – 100 and 10,000 Sales Records

In the realm of data analytics, uncovering correlations within datasets is an essential step towards extracting meaningful insights and making informed decisions. In this exploration, we delve into two distinct datasets—a small one with 100 observations and a larger counterpart with 10,000 observations. Both datasets encompass 14 variables, including categorical features like Region, Country, Item Type, Sales Channel, Order Priority, Order Date, and Ship Date, along with numerical features such as Order ID, Units Sold, Unit Price, Unit Cost, Total Revenue, Total Cost, and Total Profit.

In the smaller dataset, the 'Units Sold' variable stands out with its considerable variability, suggesting diverse sales patterns. 'Unit Price' reflects stability, indicative of a consistent pricing strategy, while 'Unit Cost' displays moderate variability, potentially influenced by production factors. The financial indicators - 'Total Revenue,' 'Total Cost,' and 'Total Profit' - exhibit substantial fluctuations, emphasizing the complexity of overall sales and operational expenses.

Correlation analysis reveals intriguing relationships. The positive correlation between 'Units Sold' and 'Total Profit' signifies that higher unit sales tend to be associated with increased total profits. Strong positive correlations between 'Unit Price' and 'Unit Cost,' as well as 'Total Revenue' and 'Total Cost,' underscore interconnections within pricing and operational dynamics. Histograms and correlation coefficients provide a comprehensive view of the dataset's characteristics.

Transitioning to the larger dataset, which boasts 10,000 observations, we encounter a diverse range of regions, countries, and product types. The numerical features in this dataset - 'Units Sold,' 'Unit Price,' 'Unit Cost,' 'Total Revenue,' 'Total Cost,' and 'Total Profit' - exhibit variability, indicative of intricate sales and financial landscapes. Notably, correlations between these variables echo patterns observed in the smaller dataset, with strong positive correlations highlighting interconnected aspects of production costs, pricing, and overall revenue.

Boxplot analysis in the larger dataset reveals a positive distribution with right-skewed patterns, particularly in 'Total Revenue,' 'Total Cost,' and 'Total Profit.' The presence of outliers in these variables adds an extra layer of complexity, hinting at potential challenges that may influence algorithm performance or necessitate careful preprocessing.

In conclusion, both datasets offer valuable insights through correlation exploration. Understanding these relationships aids in making informed decisions, whether in refining pricing strategies, optimizing production costs, or predicting financial outcomes. The presence of outliers and the complexity observed in the larger dataset underscore the importance of robust algorithm selection and thorough data preprocessing to ensure accurate modeling and reliable predictions.

I first chose the smaller dataset of 100, but because of the overfitting of my choice of algorithm, I went with the large dataset. I also had the same problem with overfitting with the large dataset using Linear Regression. Also, given the insights from your exploratory analysis, such as the presence of outliers and right-skewed distributions of the large dataset, linear regression was not the most robust choice.

I then selected Decision Tree and KNN. Decision Trees for their versatility, in handling various types of data in mixed datasets. They are easy to interpret, providing clear decision structures. Capable of grasping non-linear patterns, they work well with complex data. While resilient to outliers, prudent preprocessing is recommended. Exploring ensemble methods like Random Forests can boost predictive performance. To counter overfitting, consider pruning techniques or ensembles. Decision Trees reveal feature importance, aiding in identifying impactful variables.

One downside of Decision Trees is their susceptibility to overfitting, especially when the tree becomes too complex. This complexity can result in poor generalization to new, unseen data. Pruning techniques and controlling the tree's depth can help mitigate overfitting. Additionally, Decision Trees may not perform well with small datasets, and their sensitivity to minor variations in the data can lead to different tree structures, the reason I chose the large dataset over the small dataset.

I also selected KNN because it scales well to large datasets and identifies natural groupings in data. A downside is it requires specifying the number of clusters and is sensitive to initialization.

The analysis of the two models, Decision Tree and KNN, on the given datasets reveals differences in their performance metrics. The Decision Tree model, evaluated using Mean Squared Error (MSE), yielded a value of 10,369,206,128, indicating the average squared difference between predicted and actual values. Meanwhile, the KNN model was assessed with a metric of [1] 0.5125, likely representing classification accuracy or error for discrete outcomes. It's crucial to note that these metrics are not directly comparable, as MSE is tailored for regression tasks, while the KNN metric is more relevant to classification tasks. Further exploration of the models' strengths and weaknesses is necessary to determine their suitability for the specific dataset and analysis goals.

The analysis of a dataset can be affected by the amount of data used. Using too much data may lead to overfitting, where the model becomes too specific to the training set and struggles to generalize to new data. Conversely, using too little data can result in underfitting, where the model oversimplifies relationships. Striking a balance with an appropriate amount of representative data is crucial. This ensures the model generalizes well without being overly complex or too simplistic. Techniques like cross-validation and thorough exploration of performance metrics aid in finding this balance for a more reliable analysis.

Reference: https://excelbianalytics.com/wp/downloads-18-sample-csv-files-data-sets-for-testing-sales/