BANA 630: Project Term Paper

Jessica Becerra, Ashley Cortez, Robert Pimentel, Angelica Verduzco

December 9, 2024

Table of Contents

Phase 1

Topic Selection & Objective

Our topic is optimizing order fulfillment and shipment efficiency for superstores in the United States. We aim to assist the superstores in maximizing profit by analyzing geographical trends, customer segment trends, and product category profitability.

Practical Implications

This problem is important because there may be certain regions, customer segments or products that are costlier than the revenue they generate. It is also important for superstores to be aware of these metrics in order to better understand the trends that can optimize operational efficiency as well as profitability overall. Understanding these trends can lead to improved efficiency, cost optimization, customer satisfaction, and Market Expansion.

Role of Predictive and Prescriptive Analysis

We will use predictive analytics to estimate and forecast which states/regions are projected to be most profitable, based on the historical trends. We can also explore the price trends of each product category. As each of these variables are continuous, we plan to use linear regression to assess profitability.

Prescriptive analysis will then be used to provide recommendations for the optimal product/category combination for each recommended region that should be prioritized. We will consider resource allocation based on the forecasted demand for specific regions.

Known and Unknown Variables

Known Variables:

- Shipping metrics: Actual and scheduled days to ship
- Store attributes: Store categories and product types
- Financial metrics: Forecasted sales and profit ratios for the measurement period and discounts provided
- Customer Demographics: Customer segment data
- Operational costs: Shipping prices and associated fees

Unknown Variables:

- Future demand: Regional or product-specific demand fluctuations.
- Market dynamics: Price changes due to inflation, production costs, or other economic factors.
- Unforeseen disruptions: Potential impact of external events (e.g., supply chain disruptions, pandemics).

To address the unknown variables we will leverage forecasting models and sensitivity analysis to account for variability.

Objective Function and Limitations/Constraints

The objective function aims to maximize profitability for superstores by determining the most efficient combination of product categories, regions and customer segments. This involves:

- Minimizing shipping and operating costs
- Optimizing inventory levels, in alignment with predicted demand
- Prioritizing high-margin products for the targeted regions and customer segments.

Key constraints include historical demand for each product, the given sales targets, and shipping times, resource availability, and market fluctuations.

- Demand Trends: Historical sales data may not fully reflect demand for new or seasonal products, creating a challenge for accurate forecasting
- Shipping Times: carrier performance and customers locations impose logistical constraints that impact efficient delivery times
- Resource availability: Limitations in warehousing capacity and transportation can affect scalability
- Market Variability: Unexpected economic changes can disrupt established trends and make predictions less reliable

Finding a Dataset

We plan to use a publicly available dataset from Kaggle on Superstore Orders from 2017 - 2020. The dataset contains valuable information on fulfillment, sales, and profitability metrics, including:

- Regional breakdowns of sales and shipping data
- Product category details and corresponding profitability
- Customer segment trends and purchasing patterns

Superstore Orders Analysis File

Phase 2

Dataset Preparation

Dataset Description:
The dataset used contained superstore data from 2017 - 2020 inclusive of shipping details, product category descriptions, product prices, discounts, consumer segments and sales target goals/status. Overall, the dataset shape consisted of 30 columns and nearly 10,000 rows. This dataset was sourced from another Kaggle project titled "Superstore Orders - EDA, Forecast Model Insights."

Preprocessing Steps:
As we initialized the data to begin exploratory analysis, we followed the steps below to address any data cleaning that was needed:

- Duplicates and Missing Values: Upon initial investigation, this dataset did not have any duplicate or missing values to address or clean as it was previously prepared for exploratory analysis and insights from the original source Kaggle project.
- Feature Engineering: As mentioned in our project proposal, we wanted to explore different variables for predictive analysis such as shipping times, profit ratio, state, region and product categories.
    a. First we created a metric called "Days to Ship Difference" which was the calculated difference between "Days to Ship Actual" and "Days to Ship Scheduled". This was created to determine whether shipping times affect overall profitability
    b. Next we categorized and labeled the Profit Ratio values to "Loss", "Low", "Medium" and "High" for predictive analysis and model implementation.
    c. Lastly we made sure to convert the Order and Ship Date columns to datetime objects for consistency.
- Data Transformation: We then proceeded with LabelEncoder in Google Colab to encode categorical variables for the prediction models. The variables that were encoded included: Segment, State, Region, Category, Sub-Category, Sales Target Status, and Sales Forecast Status
- Standardization: In our last preprocessing steps, we standardized the numerical columns in the dataset using the StandardScaler approach.

Predictive Modeling

Choice of Predictive Model

For this project, we used a Random Forest Regression model for predictive analytics. This choice was based on its ability to handle non-linear relationships and capture feature importance effectively. As the problem at hand involves forecasting predictability (a continuous variable), a regression-based model was the most appropriate instead of a classification model. In addition to this, hyperparameters were tuned using GridSearchCV with a 5-fold cross-validation strategy to ensure robustness.

Linear regression was also considered, however it was not ultimately selected as it assumes a linear relationship between features and the target, which is not the case for our selected dataset.

Model Training and Validation

As mentioned in the earlier section, the dataset was preprocessed by scaling numerical variables (Sales, Discount, and Price) using StandardScaler to ensure uniformity in feature scaling. Categorical variables (State, Region, Category) were label-encoded to be used in the regression model. A train-test split of 80% training and 20% testing was performed to evaluate the model's performance on unseen data.

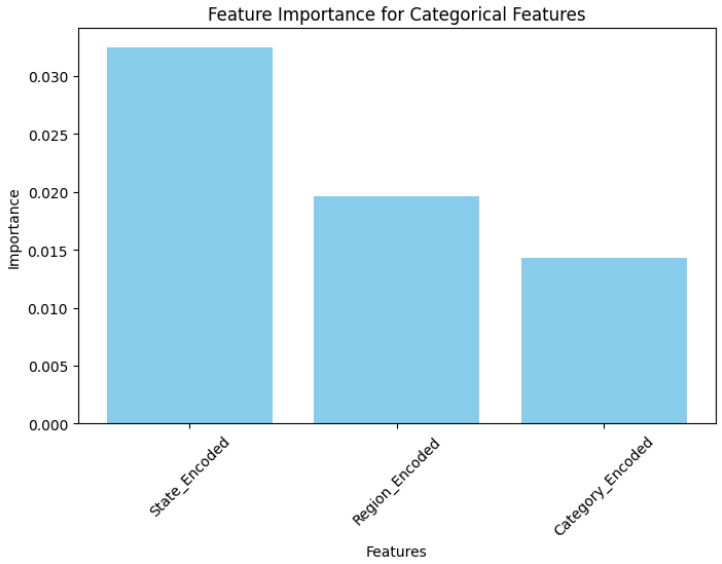
The model's performance was ultimately validated using Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and R^2 (Coefficient of Determination). Results for each metric are outlined below:

- MSE: 1.057
- RMSE: 1.028
- R^2 (After GridSearch): 0.703

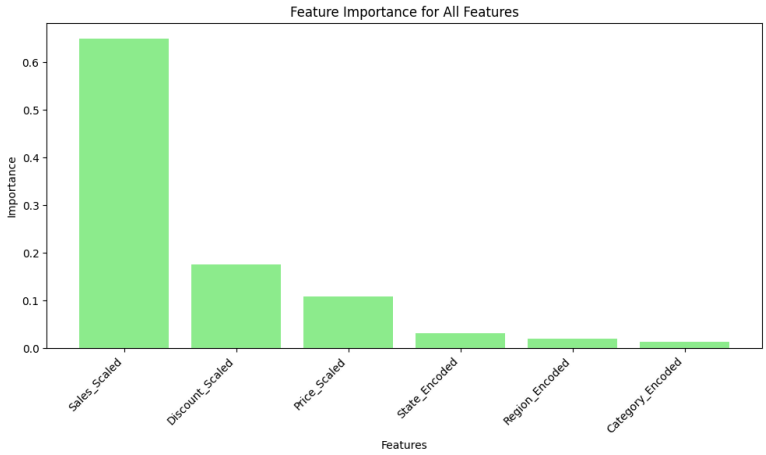
The MSE/RMSE indicates that the random forest model has reasonable predictive power, and hyperparameter tuning via GridSearch techniques indicates that the model explains 70% of the variance in profitability across the training data.

Visualizations

Feature Importance Plots

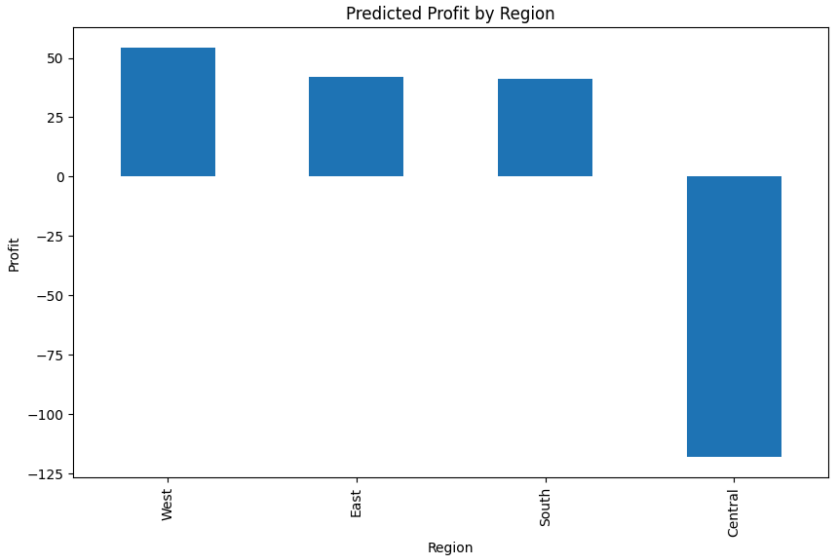


Insights: For the bar chart above, State_Encoded is the most important categorical variable. This indicates that the model uses state-level information more effectively to predict profitability. Region_Encoded is moderately important, but it is less impactful that state since it is a more aggregated variable. (Product) Category_Encoded is less important than both geographical variables.

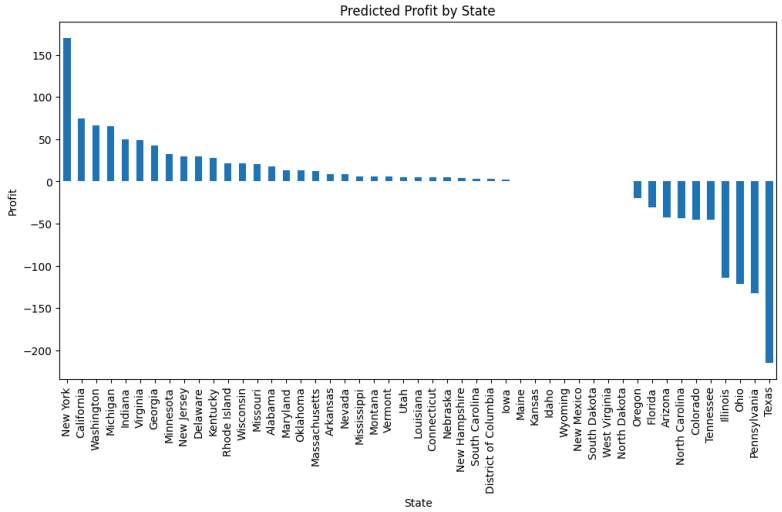


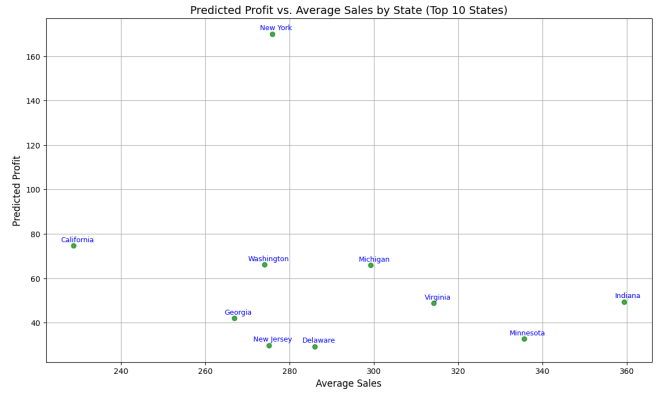
Insights: When observing feature importance across all variables (categorical and numerical) Sales data is the most important feature by far, suggesting that profitability is heavily dependent on sales volume (however this is an expected result). Discounts are the next important in predicting profitability as they can reduce profit margins (despite driving sales by motivating purchases).

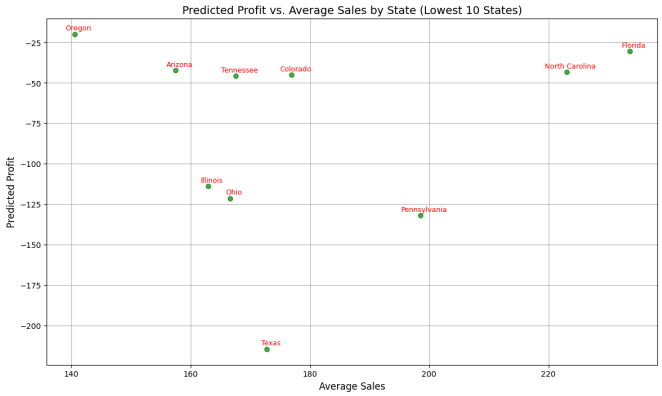
Profitability by Region and State



Insights: Central states were the least profitable with an overall negative predicted profit. Upon further analysis, this is primarily due to this region having a combination of lower prices and higher discounts. Ultimately making it a less profitable region to do business with. An initial potential solution would be to adjust the discrepancy between prices and discounts (i.e., increase prices for products in these areas or lower discounts).

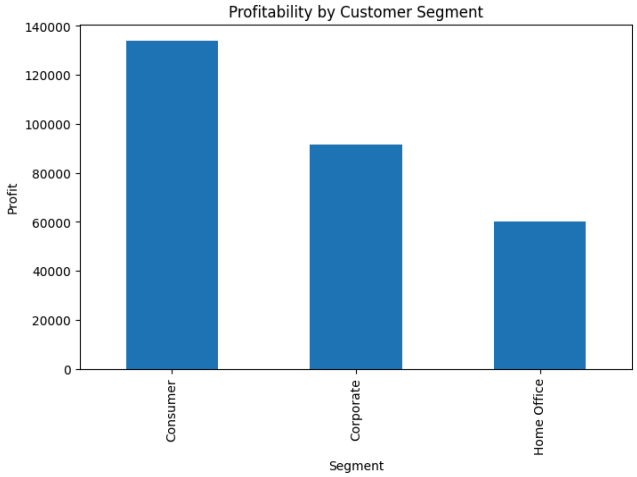






Insights: Higher performing states include California and New York. Our analysis shows that these states typically have higher prices and less discounts for consumers, resulting in overall higher profitability than states in other regions. The lower performing states on the right-hand size of the graph can be considered targets for improvement strategies mentioned in the region analysis.

Profitability by Customer Segment



Insights: In terms of customer segments, the consumer segment is the most profitable, followed by corporate, with the Home Office lagging behind. Emphasis on consumer-focused strategies can likely yield higher returns, while Home Office should be investigated further.

Product Profitability

```
# Group by product and calculate metrics
product_metrics = df_superstore.groupby('Product Name').agg({
    'Sales': 'sum',
    'Discount': 'mean',
    'Profit': 'sum'
}).sort_values(by='Profit', ascending=False)

# Display top 10 profitable products
print(product_metrics.head(10))
```

```
                                                   Sales   Discount  \
Product Name
Canon imageCLASS 2200 Advanced Copier            61600.0  12.000000
Fellowes PB500 Electric Punch Plastic Comb Bind...  27454.0  24.000000
Hewlett Packard LaserJet 3310 Copier             18840.0  20.000000
Canon PC1060 Personal Laser Copier               11620.0  15.000000
HP Designjet T520 Inkjet Large Format Printer -...  18375.0  16.666667
Ativa V4110MDD Micro-Cut Shredder                 7700.0   0.000000
3D Systems Cube Printer, 2nd Generation, Magenta    14300.0   0.000000
Plantronics Savi W720 Multi-Device Wireless Hea...   9368.0   5.714286
Ibico EPK-21 Electric Binding System             15876.0  33.333333
Zebra ZM400 Thermal Label Printer                 6966.0   0.000000

                                                    Profit
Product Name
Canon imageCLASS 2200 Advanced Copier            25200.0
Fellowes PB500 Electric Punch Plastic Comb Bind...   7751.0
Hewlett Packard LaserJet 3310 Copier             6984.0
Canon PC1060 Personal Laser Copier               4571.0
HP Designjet T520 Inkjet Large Format Printer -...   4095.0
Ativa V4110MDD Micro-Cut Shredder                3773.0
3D Systems Cube Printer, 2nd Generation, Magenta     3718.0
Plantronics Savi W720 Multi-Device Wireless Hea...   3697.0
Ibico EPK-21 Electric Binding System             3345.0
Zebra ZM400 Thermal Label Printer                3344.0
```

Insights: We also decided to analyze the profitability of individual products, as shown in the snapshot above. Given that there are many products recorded in this dataset, we decided not to move forward with granular analysis to include this variable.

Shipping Efficiency

```
Correlation between shipping delays and profitability: 0.004150725224732755
Average shipping delay by region:
Region
Central   -0.388291
East      -0.346221
South     -0.399752
West      -0.336040
Name: Days to Ship Difference, dtype: float64
```

Insights: In the bar graph, we can see that the raw data shows that when orders are shipped early, they have higher profitability than those that ship late or on time. Interestingly, orders that shipped late had a slightly higher profitability than those that shipped on time. This prompted us to look at the correlation between shipping delays and profitability, which was 0.004. Shipping delays were also not specific by geographic location and seemed relatively similar across regions. With that being said, shipping times were not considered significant in predicting profitability.

Initial Recommendations based on Predictive Analysis:

1. **For High-Performing States**:
    - **Double Down**:
        - Prioritize investment in states like those in the top 10 list.
        - Expand successful pricing and discount strategies to similar states.
2. **For Low-Performing States**:
    - **Targeted Campaigns**:
        - Increase visibility for high-profit categories in low-performing states.
        - Launch localized marketing efforts.
    - **Discount Optimization**:
        - Reduce overuse of discounts where they do not drive sufficient sales.

Phase 3

Clear Problem Formulation

Optimization Problem Variables:

- Decision Variables: Quantity of items ordered in each product category within each region
- Objective Function: **Maximize** Profit Across All States and Product-Categories
- Constraints:
    - Product Category Allocation must not Exceed the Estimated Annual Inventory*
        i. *Original dataset did not specify inventory levels, so annual inventory for the next year was calculated based on a 12-month moving average.
    - Product Category Allocation by region must not exceed the average % allocation seen in the historical data
        i. Each region should have a quantity sold goal that reflects the historical distribution for each product category.
        ii. This is important because it does not make logical sense to try to sell more in regions that have historically only counted for less than 20% of profit. (i.e., the Central Region should not account for more than 23% of Furniture sales because the Western region historically has had the majority of Furniture orders across regions)

Mathematical Formulation of Problem:

x = quantity of items ordered

Maximization of
P= [Average Category Price]x - ([Average Category Price]*[Discount%])
x <= Estimated Annual Inventory
x <= Average % Allocation by Category

Optimal Solution

Region	Product Category	Average Monthly Quantity	Average Monthly Orders	Average Quantity/Order	Average Category Price	Quantity to Sell (Decision Variable)	%Allocation by Category	Discount	Calculated Profit
Central	Furniture	152	40	4	$87.04	1,826	23%	10%	$143,053.82
	Office Supplies	451	119	4	$31.47	2,211	14%	10%	$62,613.17
	Technology	129	35	4	$107.91	0	0%	10%	$0.00
East	Furniture	184	50	4	$91.05	467	6%	15%	$36,116.10
	Office Supplies	538	143	4	$32.26	5,412	34%	14%	$150,124.89
	Technology	162	45	4	$135.38	6,454	93%	14%	$751,437.09
South	Furniture	107	28	4	$88.31	3,779	47%	12%	$293,667.31
	Office Supplies	315	83	4	$34.55	7,239	45%	17%	$207,611.47
	Technology	93	24	4	$124.05	482	7%	11%	$53,171.25
West	Furniture	225	59	4	$94.42	1,944	24%	13%	$159,709.74
	Office Supplies	603	158	4	$31.28	1,114	24%	9%	$31,722.01
	Technology	194	50	4	$114.28	0	24%	13%	$0.00
									$1,889,226.84

	Estimated Annual Quantity	Average Annual Quantity (Next Year Inventory)
Furniture	8,016	8,016
Office Supplies	15,976	22,884
Technology	6,936	6,936

Category	Region	Volume Allocation of Category by Region	%Allocation of Category by Region
Furniture	Central	1,826	23%
	East	2,211	28%
	South	1,284	16%
	West	2,695	34%
Office Supplies	Central	5,412	24%
	East	6,454	28%
	South	3,779	17%
	West	7,239	32%
Technology	Central	1,546	22%
	East	1,944	28%
	South	1,114	16%
	West	2,332	34%

Based on the snapshot above, Solver allocates inventory for each category within each region to maximize overall profit. In this scenario, it does not allocate inventory to the Technology category in both Central and Western regions.

Sensitivity Analysis

Variable Cells

	Cell	Name	Final Value	Reduced Cost	Objective Coefficient	Allowable Increase	Allowable Decrease
Central	H3	Furniture Quantity to Sell (Decision Variable)	1826	0.95	78.34	1E+30	0.95
	H4	Office Supplies Quantity to Sell (Decision Variable)	2211	28.32	28.32	1E+30	28.32
	H5	Technology Quantity to Sell (Decision Variable)	0	-13.28	97.12	13.28	1E+30
East	H6	Furniture Quantity to Sell (Decision Variable)	467	0.00	77.39	0.32	77.39
	H7	Office Supplies Quantity to Sell (Decision Variable)	5412	27.74	27.74	1E+30	27.74
	H8	Technology Quantity to Sell (Decision Variable)	6454	6.02	116.42	1E+30	6.02
South	H9	Furniture Quantity to Sell (Decision Variable)	3779	0.32	77.71	1E+30	0.32
	H10	Office Supplies Quantity to Sell (Decision Variable)	7239	28.68	28.68	1E+30	28.68
	H11	Technology Quantity to Sell (Decision Variable)	482	0.00	110.40	6.02	10.98
West	H12	Furniture Quantity to Sell (Decision Variable)	1944	4.76	82.14	1E+30	4.76
	H13	Office Supplies Quantity to Sell (Decision Variable)	1114	28.47	28.47	1E+30	28.47
	H14	Technology Quantity to Sell (Decision Variable)	0	-10.98	99.42	10.98	1E+30

Constraints

Cell	Name	Final Value	Shadow Price	Constraint R.H. Side	Allowable Increase	Allowable Decrease
D18	Furniture Estimated Annual Quantity	8016	77.39	8016	2227.97	466.69
D19	Office Supplies Estimated Annual Quantity	15976	0.00	22884	1E+30	6908.18
D20	Technology Estimated Annual Quantity	6936	110.40	6936	1064.17	481.62

In our Sensitivity Analysis, we observed why certain decision variables were given 0 allocation and the effect of volume on other regions with more allocation to certain product categories.

For reduced cost, the negative values indicate that the objective function coefficients for these decision variables would need to increase by that amount to be included in the optimal solution. As mentioned above, the current solution does not allocate product volume to Technology in both Central and Western regions. This is likely due to their sales price to discount ratios (i.e., certain areas have a higher discount rate and lower sales prices for different product categories).

For shadow prices, office supplies had a shadow price of zero, indicating that the inventory amount would not affect the optimal solution. The amount of Office Supplies sold in this scenario is below the available inventory amount, indicating that there is room to allocate more toward other portions of the problem. For Furniture and Technology categories, increasing the amount of inventory constraint would increase the optimal solution by their respective shadow prices.

Results Interpretation & Future Directions

Central and Western Technology was not included in the optimization because it wasn't the best way to fulfill orders in order to maximize overall profit considering the given discounts. Shadow prices show that more inventory can be allocated to Furniture and Technology in higher performing regions, which can ultimately help with driving better results.

We suggest that the Superstore should focus on tailoring efforts to each region and product category to make the biggest impact in overall profitability. Additionally, the company should be mindful when offering discounts in regions where prices are already significantly lower than others to avoid unnecessary profit loss.

Prescriptive Analytics is useful in scenarios like this as it can help identify where there are weaknesses in a sales area. This kind of analysis could also be applied to customer segmentation/target audience analyses in marketing. For example, if you have a product and you are trying to determine which customer segments would be best to target in future marketing efforts. In initial data collection and analysis, you could find that there are certain types of customers that do not have an affinity for your product. Prescriptive analysis would allow you to take that insight and decide to not invest further marketing efforts against that type of customer segment. Overall, prescriptive analytics is useful in many scenarios and can help business owners and analysts with decision making to optimize their overall business performance.