Task Update from SNS Group (Data analyst)

Objective: Evaluate the candidate’s proficiency in data analysis, visualization, and reporting

Part 1: Theoretical Knowledge

Datasource=**https://www.kaggle.com/datasets/arunjangir245/super-market-sales?select=supermarket_sales.csv**

1 Understanding Data Visualization

- Shoppers are drawn to the abundant variety and choices available, enabling them to select from numerous brands and sizes.
- Furthermore, supermarkets leverage their purchasing power to provide competitive prices and frequent discounts, making them an attractive option for budget-conscious consumers. With extended operating hours, including late evenings and weekends, they cater to busy schedules.
- Additionally, their commitment to offering fresh produce, seamless technology integration, exceptional customer experiences, and community engagement initiatives have solidified their appeal. Supermarkets continuously innovate to meet evolving consumer preferences, including the demand for organic and eco-friendly products.
- Their globalization efforts have also made these shopping havens a familiar and trusted presence in international markets.

2 Power BI Fundamentals

- Power BI is a data visualization tool for building and designing reports Power BI Service - the online publishing service for viewing and sharing reports and dashboards.
- Power BI mobile apps - for viewing reports and dashboards on the go.

Part 2: Practical Application:

Data Cleaning and Preparation:

Plot summary:

- EDA is like exploring a new place, looking for clues, and making sense of what you find before making any important decisions. It's a crucial step in the data analysis process.
- With the help of scatter plot The position of each dot on the horizontal and vertical axis indicates values for an individual data point. Scatter plots are used to observe relationships between variables.
- A boxplot is a standardized way of displaying the distribution of data based on a five number summary (“minimum”, first quartile (Q1), median, third quartile (Q3), and

“maximum”). ... It can also tell you if your data is symmetrical, how tightly your data is grouped, and if and how your data is skewed.

- kdeplot is a data visualization technique that employs Kernel Density Estimation (KDE) to estimate and display the probability density function of continuous data. It produces a smoothed, continuous curve that reveals the underlying distribution's shape and characteristics.
- A pairplot plot a pairwise relationships in a dataset. The pairplot function creates a grid of Axes such that each variable in data will by shared in the y-axis across a single row and in the x-axis across a single column.
- A barplot (or barchart) is one of the most common types of graphic. It shows the relationship between a numeric and a categoric variable. Each entity of the categoric variable is represented as a bar. The size of the bar represents its numeric value.
- In this dataset no null value or missing value, One straightforward way to handle missing values is by removing them. Since the data sets we deal with are often large, eliminating a few rows typically has minimal impact on the final outcome.

Part 3: Advanced Analytics

Statistical Analysis

- Correlation analysis - It is a statistical method used to measure the strength of the linear relationship between two variables and compute their association.
- Hypothesis testing -In the practice of statistics, we make our initial assumption when we state our two competing hypotheses -- the null hypothesis (H0) and the alternative hypothesis (HA).

Models used to predict total income:

K Nearest Neighbor

- K-Nearest Neighbour is one of the simplest Machine Learning algorithms based on Supervised Learning technique.
- K-NN algorithm assumes the similarity between the new case/data and available cases and put the new case into the category that is most similar to the available categories.
- Training Score:65.71

Decision Tree

- A decision tree is a flowchart-like structure in which each internal node represents a test on a feature (e.g. whether a coin flip comes up heads or tails) , each leaf node represents a class label (decision taken after computing all features) and branches represent conjunctions of features that lead to those class labels. The paths from root to leaf represent classification rules.

- Training Score:64.75

Random Forest

- Random forest, like its name implies, consists of a large number of individual decision trees that operate as an ensemble. Each individual tree in the random forest spits out a class prediction and the class with the most votes becomes our model's prediction.
- A large number of relatively uncorrelated models (trees) operating as a committee will outperform any of the individual constituent models.

Training Score:100.00

Gradient Boosting Classifier

- The GradientBoostingClassifier is a machine learning model designed for classification tasks. It utilizes gradient boosting, an ensemble technique, to combine the predictions of multiple weak classifiers sequentially.
- With features like weighted voting, adjustable learning rates, and regularization parameters, it provides robust and accurate solutions for a wide range of classification problems. It is particularly useful when dealing with complex datasets and has applications in spam detection, fraud prevention, and image classification, among others.
- Training Score:88.28

Conclusion:

- As we see best Model is given by Random forest classifier(100% Accuracy).