Predictive Model & Final Report

Section 1: Setup

A. Research Questions

1. Can a predictive model accurately predict the profit based on the given categorical features (Category, Sub-category, Region) and numeric features (Sales, Quantity, Discount)?
    - Null hypothesis (H0): There is no significant relationship between the features and the profit, and the predictive model's predictions are not better than random chance.
    - Alternative hypothesis (H1): There is a significant relationship between the features and the profit, and the predictive model's predictions are better than random chance.
2. Which features are the most important in predicting profit using a predictive model?
    - Null hypothesis (H0): All features have equal importance in predicting the profit, and there is no significant difference in the feature importance scores.
    - Alternative hypothesis (H1): There are significant differences in the feature importance scores, indicating that certain features are more important than others in predicting the profit.

B. Quantify Reliability

1. Mean Squared Error (MSE)
   MSE is a measure of the average squared difference between the predicted values and the actual values. It is calculated by taking the average of the squared differences between each predicted value and the corresponding actual value. MSE penalizes larger errors more heavily than smaller errors because of the squaring operation. A lower MSE indicates a better model fit, where the predicted values are closer to the actual values.

2. Mean Absolute Error (MAE):
   MAE is a measure of the average absolute difference between the predicted values and the actual values. It is calculated by taking the average of the absolute differences between each predicted value and the corresponding actual value. Unlike MSE, MAE does not square the differences, which means it treats all errors equally regardless of their direction. MAE provides a more intuitive understanding of the average magnitude of errors in the model predictions.

3. R-squared (Coefficient of Determination):
   R-squared is a statistical measure that represents the proportion of variance in the dependent variable (target) that can be explained by the independent variables (features) in a regression model. R-squared is often interpreted as the percentage of the target variable's variance that is explained by the model. A higher R-squared value indicates a better fit of the model, implying that the model's predictions capture a larger proportion of the variability in the target variable.

C. Identify Dataset

1. The dataset utilized in this study, acquired from Kaggle, comprises a total of 21 columns and 9994 rows.
2. Three columns with categorical values (Category, Sub-Category, Region) and four columns with numeric values (Sales, Quantity, Discount, Profit) are utilized in the experiment.

Section 2: Approach

A. Random Forest

The machine learning model utilized in this study is Random Forest. It is an ensemble learning algorithm used for both classification and regression tasks. It consists of a collection of decision trees, where each tree is constructed independently using a random subset of the training data. It utilizes majority voting or aggregation to make predictions.

B. Features

Three categorical features (Category, Sub-Category, Region) and three numerical features (Sales, Quantity, Discount) are utilized in this experiment. Three categorical features are converted into numerical features (0, 1, 2) by using function LabelEncoder().

C. Model Tuning

In this study, two distinct methods were employed for model tuning, focusing on optimizing the hyperparameters n_estimator and max_depth.

- The first method involved employing cross-validation in conjunction with a for loop to iteratively evaluate different values for n_estimator. This iterative process allowed for the identification of the optimal n_estimator value that yielded the best model performance. Two performance metrics, Mean Squared Error (MSE) and Mean Absolute Error (MAE), are utilized to determine the optimal hyperparameter configuration.
- The second method utilized the gridsearchcv technique, which is a systematic and exhaustive search over specified hyperparameter combinations. The gridsearchcv process facilitated the identification of the optimal combination of n_estimator and max_depth that resulted in the highest model performance.

D. Analysis Approach

1. Research Question 1: Can a predictive model accurately predict the profit based on the given categorical features and numeric features?
   Model tuning are utilized to leverage the best model configuration to generate predictions, denoted as y_pred, for the target variable. Subsequently, a comprehensive evaluation was conducted to compare the predicted values, y_pred, with the corresponding ground truth values, y_test. By comparing y_test and y_pred using MSE, MAE, and R-squared, a comprehensive evaluation of the model's performance, enables the determination of its efficacy in capturing the variability and trends within the dataset.

2. Research Question 2: Which features are the most important in predicting profit using a predictive model?
   An assessment was conducted to rank the importance of three categorical features (Category, Sub-Category, Region) and three numeric features (Sales, Quantity, Discount) within the dataset. The feature importance ranking aimed to discern the relative significance of these features in relation to the target variable (Profit).

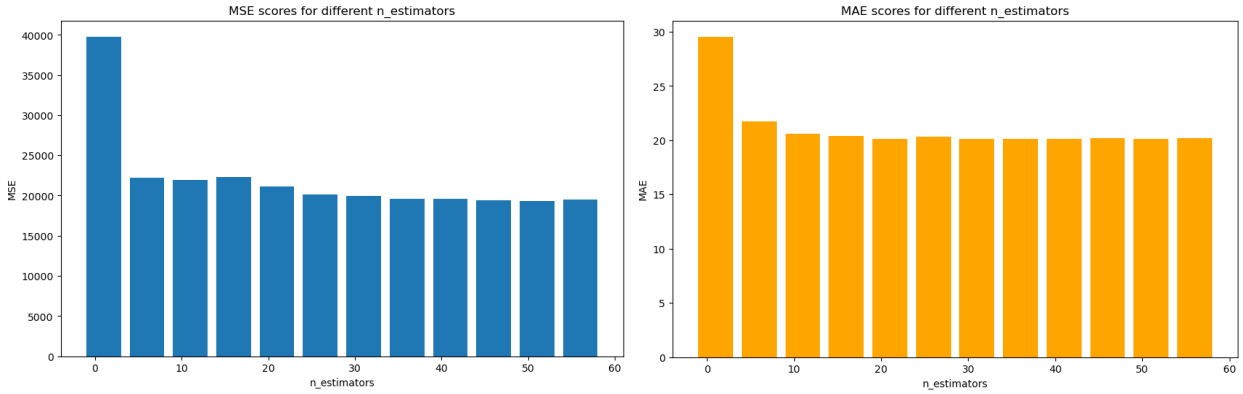
Section 3: Results

A. Summary

1. Optimal hyperparameter (n_estimator and max_depth)

Graph 1 and 2 shows Mean Squared Error (MSE) and Mean Absolute Error (MAE) based on different n_estimators. When the n_estimator is close to 51, both MSE and MAE achieve the lowest value. Hence, 51 is the optimal n_estimator for the random forest model.

Graph 3 displays the result that comes from gridsearchcv technique. It facilitated the identification of the optimal combination of n_estimator and max_depth. Hence, 'max_depth is None' and 'n_estimator is 51' is the optimal combination of hyperparameter for random forest model.



Graph 1 & 2

```
The best parameters are: {'max_depth': None, 'n_estimators': 51}
```

Graph 3

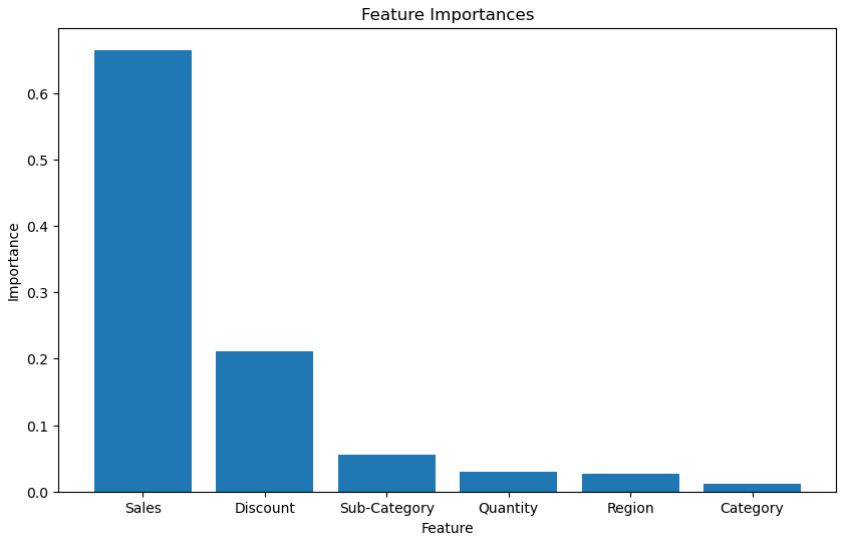
2. Comparison of y_pred and y_test

   Graph 4 shows the results of the comparison of y_pred and y_test. According to graph 4, MSE value of 6856 implies that the model's predictions have a larger overall deviation from the true values. The MAE value of 17.79 indicates the model's predictions are closer to the true values with smaller deviations or errors on average. The R-squared value of 0.85 indicates a stronger correlation between the predictors and the target variable.

```
Mean Squared Error (MSE): 6855.9908499049525
Mean Absolute Error (MAE): 17.79375575233118
R-squared:  0.8495707150908479
```

Graph 4

3. Feature importance
   - According to Graph 5, feature Sales is more important than others in predicting the profit, while feature category is less important than others.



Graph 5

B. Critically Analyze Results

1. Setup or Approach Limitations:
   The chosen hyperparameter configuration for the Random Forest model, though optimal within the given evaluation, may not necessarily capture the best possible model performance. Exploring a wider range of hyperparameter values or employing advanced techniques like Bayesian optimization could lead to further improvements.
2. Possible Improvements:
   - Exploring alternative algorithms or ensemble techniques beyond Random Forest may provide improvements in prediction accuracy.
   - Adding additional relevant features might enhance the model's predictive capabilities.

Section 4: Conclusion

From this study, several key insights can be gained that would enhance my abilities as a data analyst, like importance of model evaluation, hyperparameter tuning, feature importance analysis, critical analysis and limitations. Considering the provided context, I would recommend the approach outlined in this study as a potential solution. The reported performance metrics indicate promising results, such as relatively low MAE and a high R-squared value, which suggest the model's ability to capture patterns and explain the target variable's variance.

Overall, the study provides valuable insights and methodologies that enhance my skills as a data analyst, emphasizing the importance of rigorous evaluation, thoughtful analysis, and continuous improvement in modeling techniques.