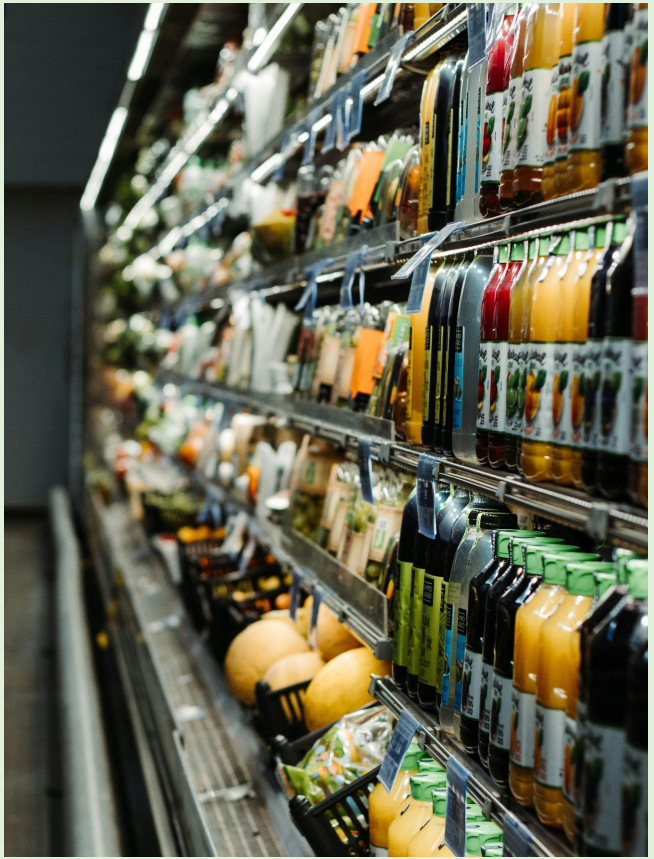
Group 25: Superstore Project

Zaira Ramji, Maryam Alsayed, MD Tanjim Hossain

Problem

A Superstore Giant has requested help in understanding what practices will be best for their profit margins. For this, they have provided sales data across several regions for several categories of products.

The objective of our project is to predict which factors lead to profitability for the superstore using 3 machine learning models.



Dataset

The dataset was an excel file containing 9994 entries with the the following meta data:

- Row ID => Unique ID for each row.
- Order ID => Unique Order ID for each Customer.
- Order Date => Order Date of the product.
- Ship Date => Shipping Date of the Product.
- Ship Mode=> Shipping Mode specified by the Customer.
- Customer ID => Unique ID to identify each Customer.
- Customer Name => Name of the Customer.
- Segment => The segment where the Customer belongs.
- Country => Country of residence of the Customer.
- City => City of residence of of the Customer.
- State => State of residence of the Customer.
- Postal Code => Postal Code of every Customer.
- Region => Region where the Customer belong.
- Product ID => Unique ID of the Product.
- Category => Category of the product ordered.
- Sub-Category => Sub-Category of the product ordered.
- Product Name => Name of the Product
- Sales => Sales of the Product.
- Quantity => Quantity of the Product.
- Discount => Discount provided.
- Profit => Profit/Loss incurred.

Dataset

Row ID	Order ID	Order Date	Ship Date	Ship Mode	Customer ID	Customer Na	Segment	Country	City	State	Postal Code	Region	Product ID	Category	Sub-Categor	Product Nam	Sales	Quantity	Discount	Profit
1	CA-2016-152	11/8/2016	11/11/2016	Second Class	CG-12520	Claire Gute	Consumer	United State	Henderson	Kentucky	42420	South	FUR-BO-100	Furniture	Bookcases	Bush Somers	261.96	2	0	41.9136
2	CA-2016-152	11/8/2016	11/11/2016	Second Class	CG-12520	Claire Gute	Consumer	United State	Henderson	Kentucky	42420	South	FUR-CH-100	Furniture	Chairs	Hon Deluxe F	731.94	3	0	219.582
3	CA-2016-138	6/12/2016	6/16/2016	Second Class	DV-13045	Darrin Van H	Corporate	United State	Los Angeles	California	90036	West	OFF-LA-1000	Office Suppli	Labels	Self-Adhesive	14.62	2	0	6.8714
4	US-2015-108	10/11/2015	10/18/2015	Standard Cla	SO-20335	Sean O'Donn	Consumer	United State	Fort Lauderd	Florida	33311	South	FUR-TA-1000	Furniture	Tables	Bretford CR4	957.5775	5	0.45	-383.031
5	US-2015-108	10/11/2015	10/18/2015	Standard Cla	SO-20335	Sean O'Donn	Consumer	United State	Fort Lauderd	Florida	33311	South	OFF-ST-1000	Office Suppli	Storage	Eldon Fold 'N	22.368	2	0.2	2.5164
6	CA-2014-115	6/9/2014	6/14/2014	Standard Cla	BH-11710	Brosina Hoffr	Consumer	United State	Los Angeles	California	90032	West	FUR-FU-100(	Furniture	Furnishings	Eldon Expres	48.86	7	0	14.1694
7	CA-2014-115	6/9/2014	6/14/2014	Standard Cla	BH-11710	Brosina Hoffr	Consumer	United State	Los Angeles	California	90032	West	OFF-AR-100(	Office Suppli	Art	Newell 322	7.28	4	0	1.9656
8	CA-2014-115	6/9/2014	6/14/2014	Standard Cla	BH-11710	Brosina Hoffr	Consumer	United State	Los Angeles	California	90032	West	TEC-PH-100(	Technology	Phones	Mitel 5320 IP	907.152	6	0.2	90.7152
9	CA-2014-115	6/9/2014	6/14/2014	Standard Cla	BH-11710	Brosina Hoffr	Consumer	United State	Los Angeles	California	90032	West	OFF-BI-1000	Office Suppli	Binders	DXL Angle-Vi	18.504	3	0.2	5.7825
10	CA-2014-115	6/9/2014	6/14/2014	Standard Cla	BH-11710	Brosina Hoffr	Consumer	United State	Los Angeles	California	90032	West	OFF-AP-100(	Office Suppli	Appliances	Belkin F5C20	114.9	5	0	34.47
11	CA-2014-115	6/9/2014	6/14/2014	Standard Cla	BH-11710	Brosina Hoffr	Consumer	United State	Los Angeles	California	90032	West	FUR-TA-1000	Furniture	Tables	Chromcraft F	1706.184	9	0.2	85.3092
12	CA-2014-115	6/9/2014	6/14/2014	Standard Cla	BH-11710	Brosina Hoffr	Consumer	United State	Los Angeles	California	90032	West	TEC-PH-100(	Technology	Phones	Konftel 250 C	911.424	4	0.2	68.3568

Kaggle link for the dataset: https://www.kaggle.com/datasets/vivek468/superstore-dataset-final
Tableau link for the dataset:
https://community.tableau.com/s/question/0D54T00000CWeX8SAL/sample-superstore-sales-excelxls

Deviation from proposal

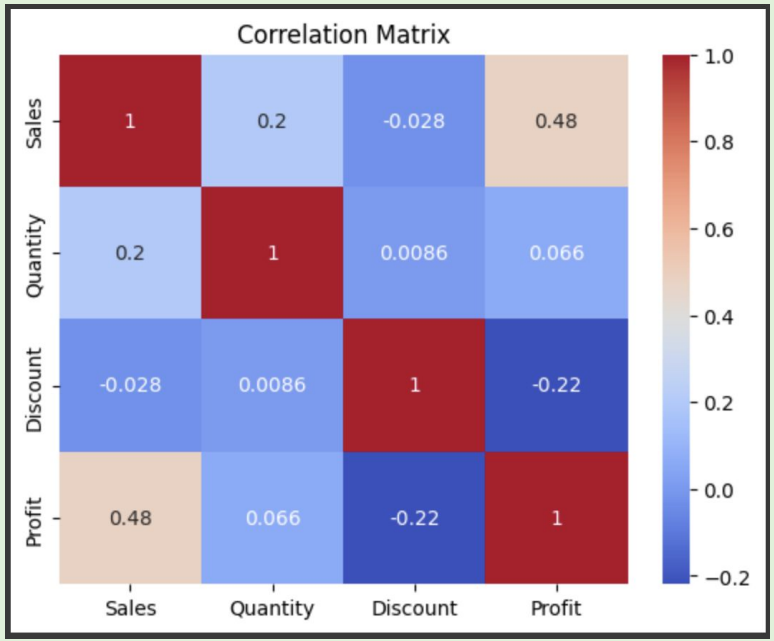
We originally proposed that we would predict the profitability of various products, and specifically which products would result in the most financial gain or loss. Through further analysis we found that our use of the models was best suited to identifying the features that are most important for profit. We have changed our goal to identifying which factors have the most impact on the stores profitability overall.

Data Preprocessing

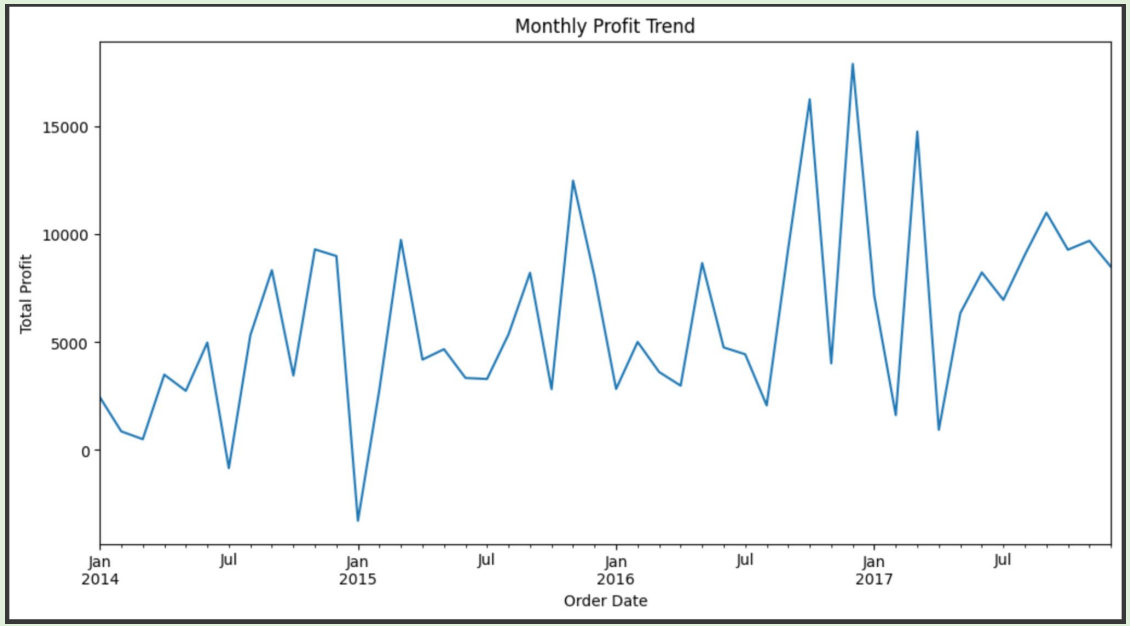
- Target Variable: “Profit”
- Extract date-based features
- Check for missing values
- Encode categorical columns with get_dummies
- Scaling

Exploratory Data Analysis

Correlation matrix of numerical data



Monthly Profit Trend

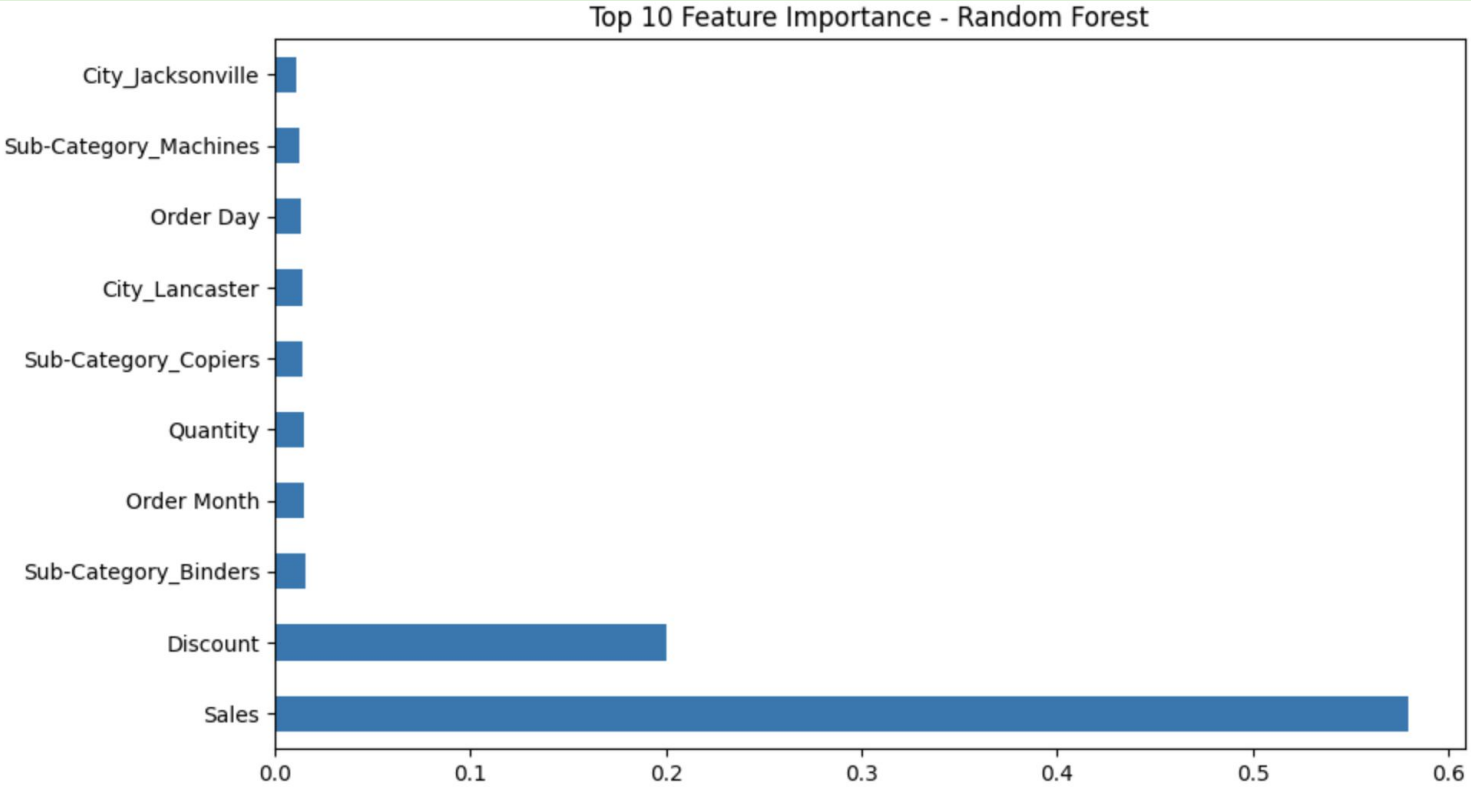


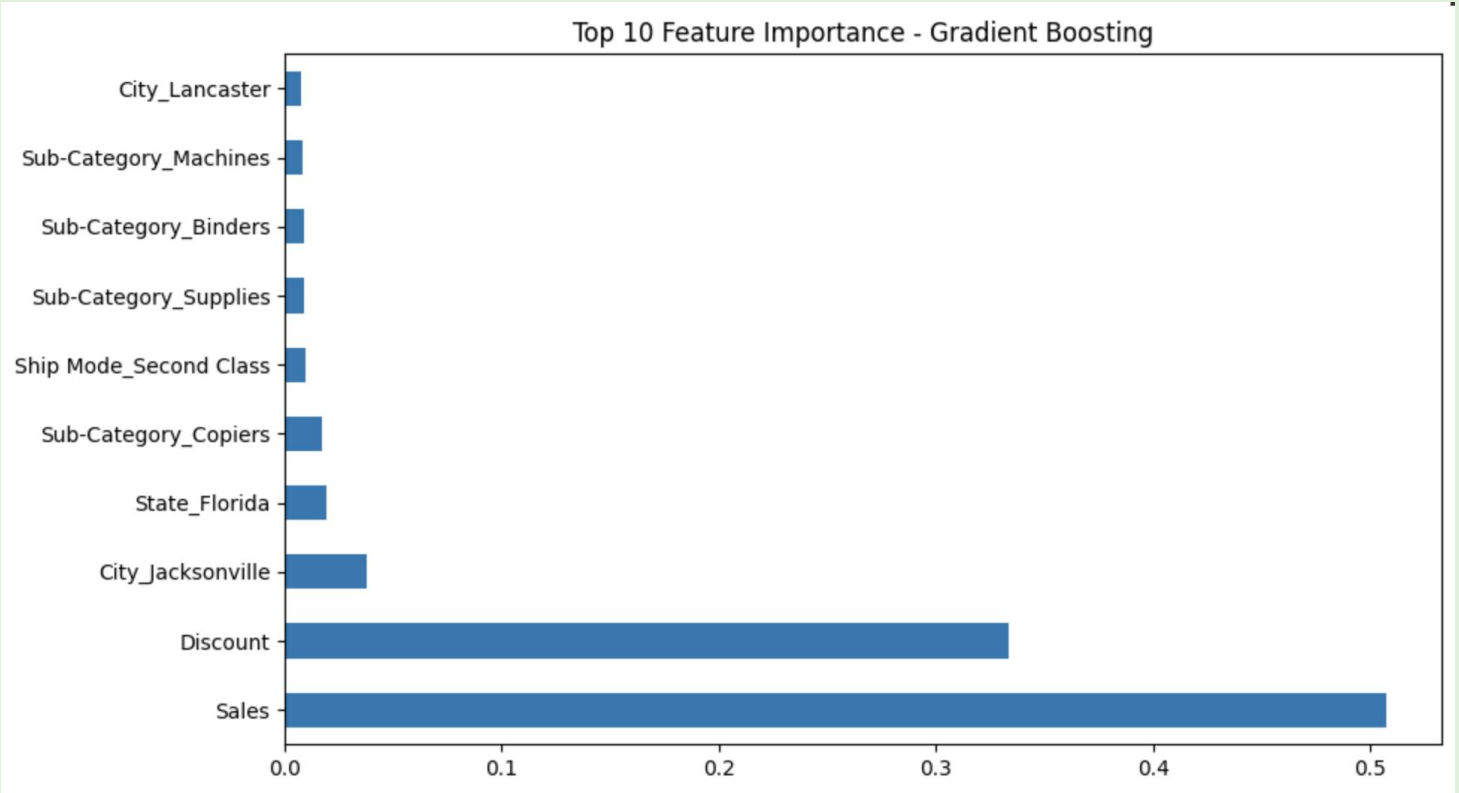
Model Implementation

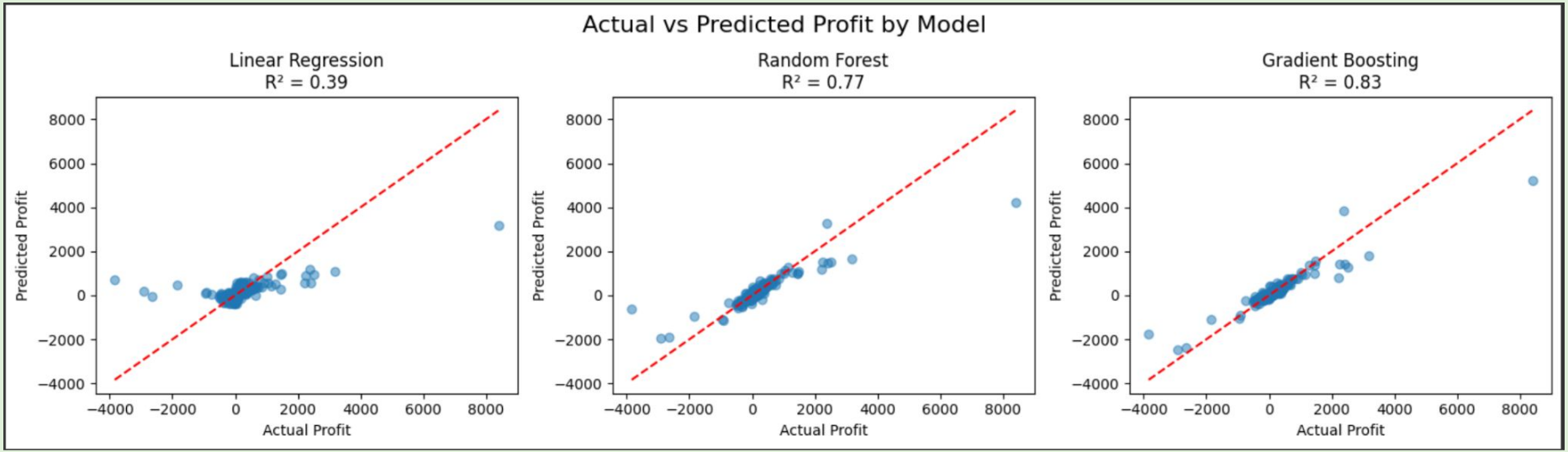
Model Implementation

- Linear Regression
- Random Forest Regressor
- Gradient Boosted Regressor
- Standard Scaler
- Test size of 20%
- n_estimators = 100, random_state = 0

	RMSE	R2	CV Mean R2
Linear Regression	230.910683	0.387676	0.211752
Random Forest	142.478003	0.766875	0.639864
Gradient Boosting	120.409796	0.833499	0.678270


Top 10 Feature Importance - Random Forest
City_Jacksonville
Sub-Category_Machines
Order Day
City_Lancaster
Sub-Category_Copiers
Quantity
Order Month
Sub-Category_Binders
Discount
Sales
0.0
0.1
0.2
0.3
0.4
0.5
0.6


Top 10 Feature Importance - Gradient Boosting
City_Lancaster
Sub-Category_Machines
Sub-Category_Binders
Sub-Category_Supplies
Ship Mode_Second Class
Sub-Category_Copiers
State_Florida
City_Jacksonville
Discount
Sales
0.0
0.1
0.2
0.3
0.4
0.5


Actual vs Predicted Profit by Model
Linear Regression
R² = 0.39
Random Forest
R² = 0.77
Gradient Boosting
R² = 0.83
Predicted Profit
Actual Profit
8000
6000
4000
2000
0
−2000
−4000
−4000
−2000
0
2000
4000
6000
8000

Hyperparameter Tuning

Hyperparameter Tuning

- We used a parameter grid and grid search to find the best model
- Pre-pruning to prevent overfitting
- Due to the scale of the data, the grid search took an extensive amount of time, so we had to reduce the amount of parameters we tried

```
param_grids = {
        'Linear Regression': {},
        'Random Forest': {
        'n_estimators': [100],
        'max_depth': [None, 10, 20],
        'min_samples_split': [2, 5],
        'min_samples_leaf': [1, 2]
        },
        'Gradient Boosting': {
        'n_estimators': [50, 100],
        'max_depth': [3, 5, 7]
        }
}
```

Results & Evaluation

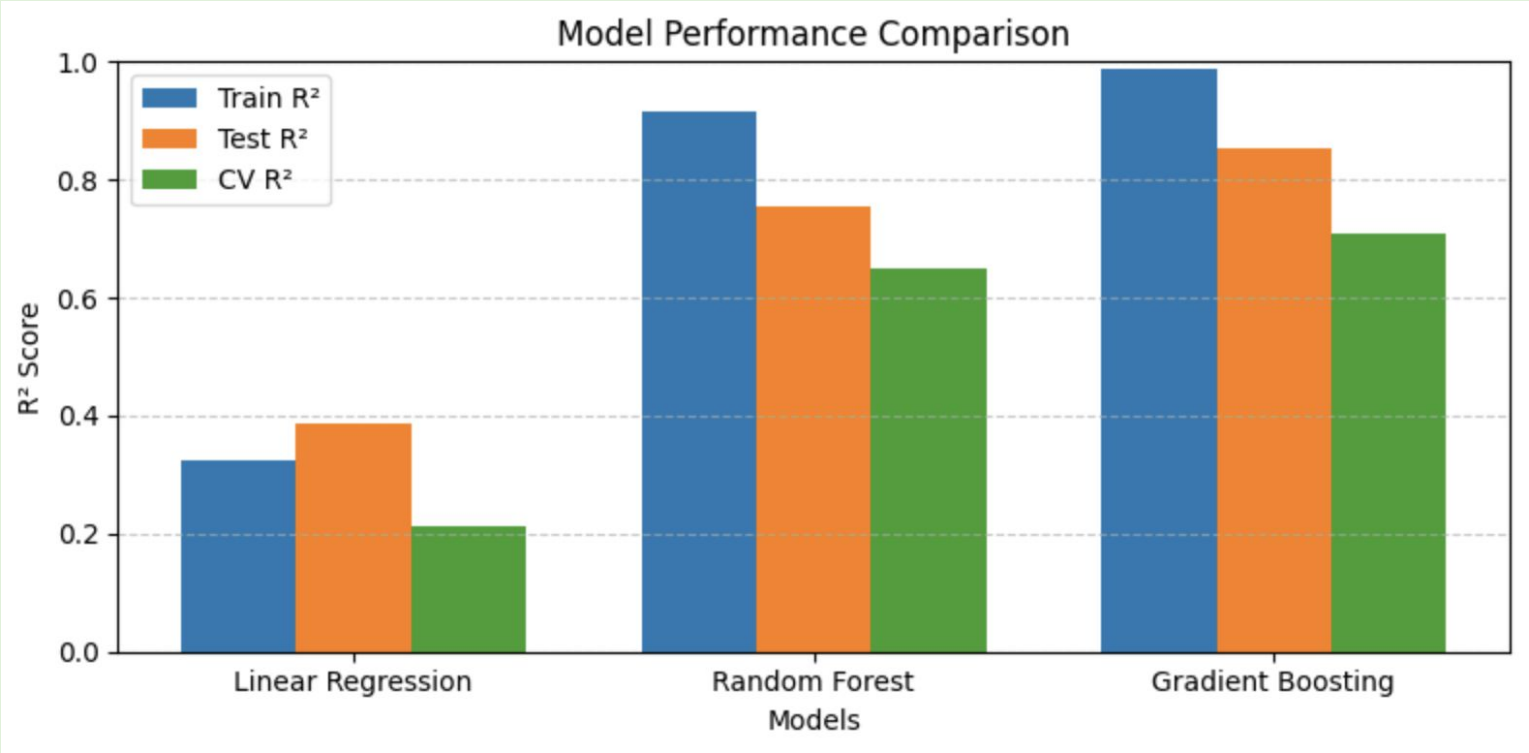
```
Performing grid search for Linear Regression...
Fitting 3 folds for each of 1 candidates, totalling 3 fits

Performing grid search for Random Forest...
Fitting 3 folds for each of 12 candidates, totalling 36 fits

Performing grid search for Gradient Boosting...
Fitting 3 folds for each of 6 candidates, totalling 18 fits

Model Comparison:
                                                Best Parameters  \
Linear Regression                                            {}
Random Forest      {'max_depth': 20, 'min_samples_leaf': 1, 'min_...
Gradient Boosting              {'max_depth': 5, 'n_estimators': 100}

                  Train R2   Test R2 Cross Validation Train Score
Linear Regression  0.325829  0.387676                    0.211752
Random Forest      0.915826  0.755738                    0.648254
Gradient Boosting  0.986953   0.85335                    0.709766
```


Model Performance Comparison
Train R²
Test R²
CV R²
R² Score
1.0
0.8
0.6
0.4
0.2
0.0
Linear Regression
Random Forest
Gradient Boosting
Models

Conclusion

Summary of findings:

- Best model before hyperparameter tuning: Gradient Boosted Regressor
  - Best R2 score (0.833)
  - Lowest RMSE (120.4)
- Best model after hyperparameter tuning: Gradient Boosted Regressor
  - Best test accuracy (0.853)
  - Best CV score (0.709)
- Most important features for profitability: Discount, Sales

Thank You!