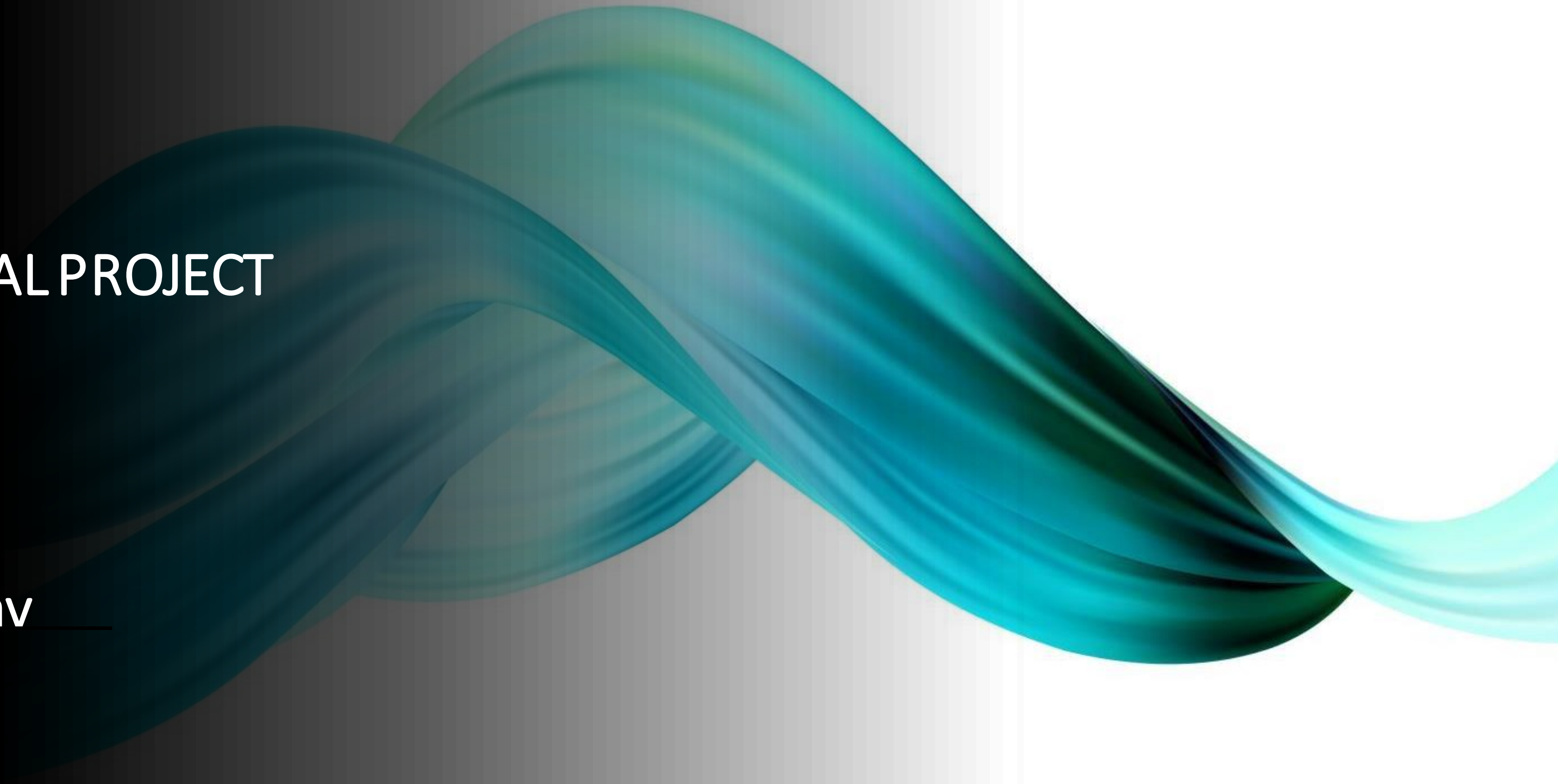
MACHINE LEARNING FINAL PROJECT

- Syed Rizwan
- Karthik Mantri
- Kashappa Omkar Jadhav



Contents

- Business Scenario
- Data Description
- Data Preprocessing
- Model Selection
- Data Analysis
- Conclusion

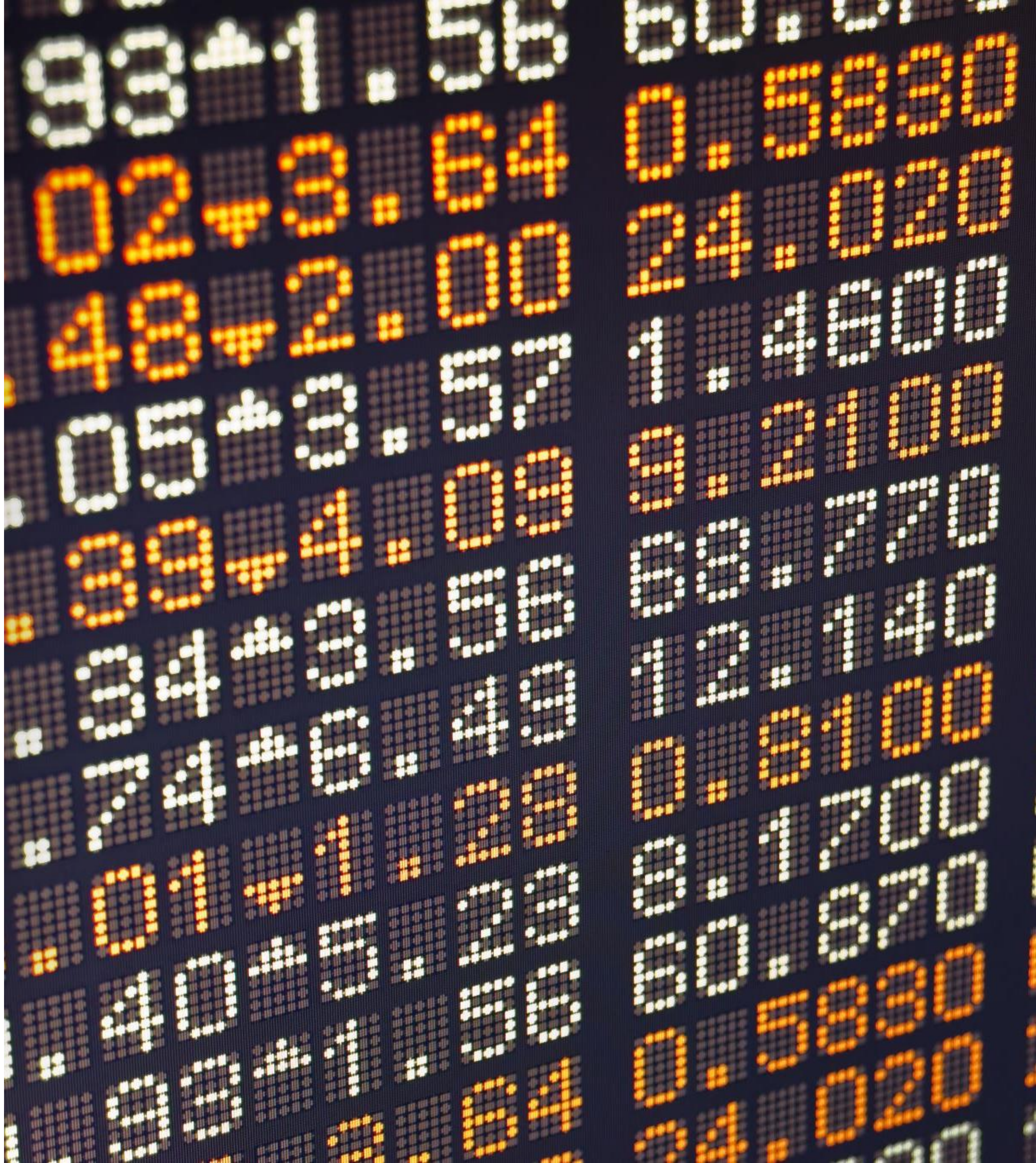
Business Scenario

- Predicting the ideal sell price for supermarket goods while taking into account variables like brand, product size, maximum retail price (MRP), and discounts is the main goal. For precise price forecasts, two machine learning models.

- Our job is to Develop accurate models for predicting product prices and providing valuable insights into supermarket pricing and inventory management.

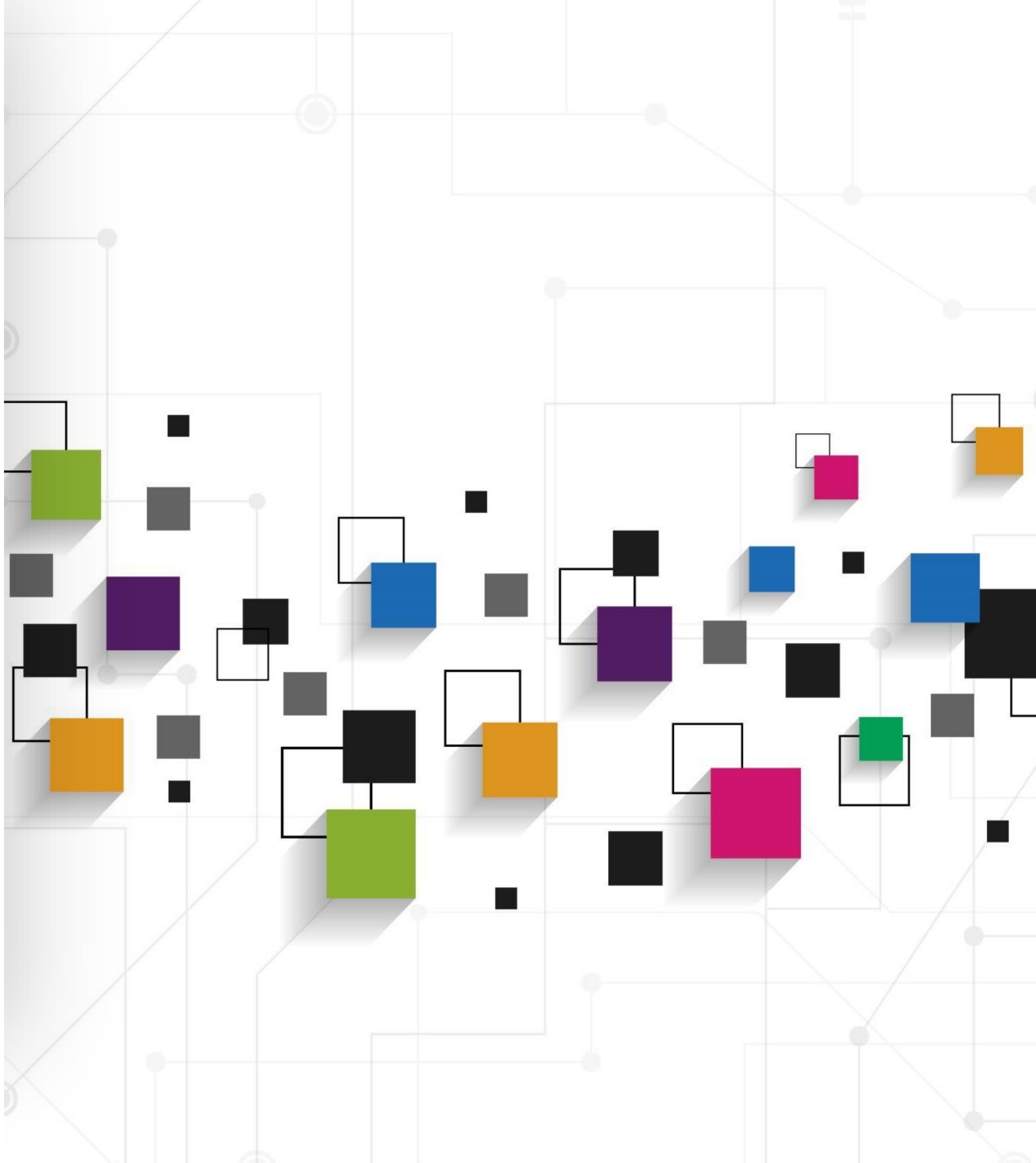
Data Description

- Brand Name: Brand of the product.
- Product ID: Unique identifier for each product.
- Product Name: Name of the product.
- Brand Desc: Description of the brand.
- Product Size: Size of the product.
- Currency: Currency used for pricing.
- MRP: Maximum Retail Price of the product.
- Sell Price: Actual selling price of the product.
- Discount: Discount applied to the product.
- Category: Category to which the product belongs.



Data Preprocessing:-

- S no, Currency, Product ID, Brand Desc, column that is not meaningful for the model can be removed from both train and test data.
- Converting categorical data into numerical data in both test and train dataset.
- Category
- Brand Name
- Product Size



Model Selection

- We have selected below regression models for evaluating the performance of the train data.
- Multiple Linear Regression
- Random Forest Regressor(RF)

Data Analysis

- Since only two columns in the datasets have numerical entries (i.e., MRP and SELL PRICE), the majority of the features are categorical data (strings).
- Unwanted features like Sno and currency In addition to removing these features, we also removed the integer from the discount variable.

•In order to avoid having to deal with missing values, we checked for Null/NAN/dirty record values and replaced the data with data that the model could understand.

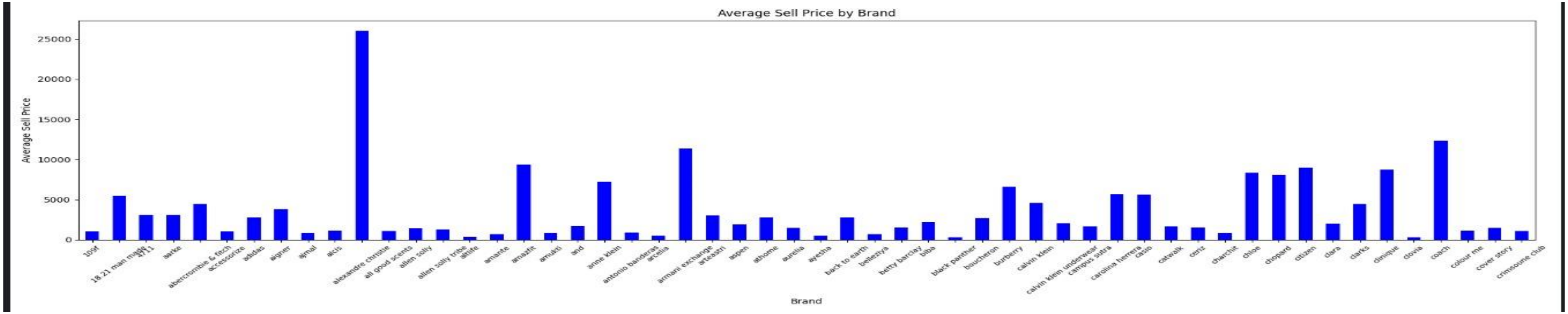
• We also had a check on datatypes and analyzed non numerical values/ features.

•After data cleaning, move on to the following stages of data analysis, such as determining how different features and variables are correlated.

Co-relation matrix between all attributes

	BrandName	MRP	SellPrice	Discount	Category
BrandName	1	0.17263	0.12757	-0.00006	-0.31508
MRP	0.17263	1	0.43494	0.14646	0.12582
SellPrice	0.12757	0.43494	1	-0.38557	-0.03504
Discount	-0.00006	0.14646	-0.38557	1	0.13496
Category	-0.31508	0.12582	-0.03504	0.13496	1

We drew a bar graph to illustrate the relationship between selling price and supermarket brand.



Conclusion

- After doing all the regression model analysis we find Random Forest Regressor is the best model.
- Training Accuracy = 86.20%
- Validation Accuracy = 85.56%
- The Linear Regression model has moderate explanatory power, accounting for approximately 49% of the variation in product selling prices.
- In the future, we can consider different hyperparameters for tuning to improve results.
- Implementing more models to gain better results.