**Sales Classification Using Random Forest**

**Report**

**1. Project Overview**

This project aims to analyze and identify key attributes that drive high sales performance. By applying machine learning techniques, specifically the **Random Forest Classifier**, we classify sales data into **“High”** or **“Low”** categories and extract valuable business insights.

**2. Objective**

To develop a classification model that predicts whether product sales will be "High" or "Low" based on various independent attributes like **Price, Income, Advertising, Shelve Location**, and others.

3. **Solution Architecture**

+--------------------+

| Raw Sales Data |

+---------+----------+

|

+--------v--------+

| Data Preprocessing |

| - Label encoding |

| - Sales categorization (High/Low) |

+--------+--------+

|

+--------v--------+

| Exploratory Data Analysis |

| - Visualizations with Seaborn |

| - Feature insights |

+--------+--------+

|

+--------v--------+

| Model Building: Random Forest |

| - Train/Test split |

| - Hyperparameter tuning |

+--------+--------+

|

+--------v--------+

| Evaluation & Inference |

| - Accuracy, Report, Confusion Matrix |

| - Feature Importance |

+--------------------+

**4. Methodology**

| **Phase** | **Description** |
| --- | --- |
| Data Understanding | Studied feature distributions, correlations, and potential business impact |
| Data Preparation | Converted continuous sales to binary categories; encoded categorical vars |
| Exploratory Analysis | Conducted visual analysis using Seaborn for trends and insights |
| Model Training | Built Random Forest Classifier on training data |
| Evaluation | Measured accuracy, precision, recall, and feature importance |
| Reporting & Inference | Explained findings and model outcomes in business terms |

**5. Technologies Used**

* **Language**: Python
* **Libraries**: Pandas, Seaborn, Matplotlib, Scikit-learn
* **Tools**: Jupyter Notebook
* **Output Format**: HTML (Notebook)

**6. Time Taken**

| **Activity** | **Estimated Time** |
| --- | --- |
| Data Loading & Understanding | 1 hour |
| EDA & Graphs | 2 hours |
| Data Preprocessing | 1 hour |
| Model Development | 2 hours |
| Evaluation & Tuning | 1 hour |
| Documentation | 1 hour |
| **Total Time** | **8 hours** |

**7. Challenges Faced**

* Determining the right threshold to convert Sales into categorical class
* Handling categorical variables without introducing bias
* Avoiding overfitting on small feature sets
* Interpreting model outputs in a business context

**8. Complexity Level**

* **Level**: Moderate
* **Reason**: Requires classification of continuous variables, categorical encoding, and multi-feature interaction modeling (handled well by Random Forest).

**9. Business Impact**

**Targeted Marketing Strategy**

* Helps identify what factors boost product sales—enabling focused promotional campaigns

**Pricing & Promotion Optimization**

* Understand how price and advertising impact performance

**Product Placement Insight**

* Shelve location’s high importance indicates how layout impacts consumer choices

**Scalable Model**

* The solution is adaptable to different product categories and markets

**10. Conclusion**

The model successfully identifies the drivers behind high product sales. With clear EDA visualizations and a robust classifier, the approach enhances decision-making across pricing, advertising, and merchandising.