**Regression Task – Spending Prediction**

This question involved building predictive models to estimate customer *Spending*. I performed two tasks:

**(a) Predicting Spending (Full Dataset)**

I explored the following models:

* **Linear Regression**
* **k-Nearest Neighbors (k-NN)**
* **Regression Tree (DecisionTreeRegressor)**
* **Support Vector Regressor (SVR)**
* **Neural Network (MLPRegressor)**
* **Ensembling: Random Forest & Gradient Boosting**

**Preprocessing & Tuning**

* All models used StandardScaler or MinMaxScaler (as appropriate).
* Optuna was used for hyperparameter tuning (e.g., number of neighbors for k-NN, depth for trees, etc.).
* Nested CV was used to prevent data leakage during model selection.

**Best Model Results:**

| **Model** | **R² Score** | **RMSE** | **MAE** |
| --- | --- | --- | --- |
| Linear Regression | 0.419 | 145.7 | 76.3 |
| k-NN (k=6) | 0.501 | 135.1 | 72.7 |
| Decision Tree | 0.433 | 144.0 | 75.6 |
| SVR (RBF) | 0.507 | 134.4 | 71.5 |
| MLPRegressor | 0.5122 | 134.9 | 71.9 |
| **Random Forest** | **0.5451** | **130.3** | **72.9** |

**Final Model:**

**MLPREGRESSOR (Neural Network)**

* **hidden\_layer\_sizes=(186,) \* 2,**
* **activation='relu',**
* **solver='adam',**
* **alpha=0.001067564832072125,**
* **batch\_size=128,**
* **learning\_rate\_init=0.010596735824594606,**
* **learning\_rate='constant',**
* **max\_iter=430,**
* **random\_state=42**

**Learning Curve Insights:**

* MLP showed overfitting but steady improvement with more data.
* Random Forest had more stable validation curves and better generalization.

**(b) Predicting Spending (Restricted to Purchase=1)**

I created a new dataset with only those rows where Purchase == 1. Models were retrained and tuned.

| **Model** | **R² Score** | **RMSE** | **MAE** |
| --- | --- | --- | --- |
| Linear Regression | 0.372 | 106.8 | 62.1 |
| k-NN (k=4) | 0.402 | 102.3 | 59.3 |
| Decision Tree | 0.388 | 104.7 | 60.9 |
| SVR (RBF) | 0.418 | 99.6 | 58.0 |
| MLPRegressor | 0.421 | 99.2 | 57.4 |
| **Random Forest** | **0.447** | **97.5** | **57.1** |

**Final Model:**

**RANDOM FOREST CLASSIFIER**

* **n\_estimators=265,**
* **max\_depth=5,**
* **min\_samples\_split=3,**
* **min\_samples\_leaf=2,**
* **max\_features=None,**
* **bootstrap=True,**
* **random\_state=42**

**(c) Comparing Performance – Full vs. Restricted**

For all models, performance improved when predicting only for purchasing customers.

* RMSE and MAE were significantly lower.
* R² also improved, since the target variable range narrowed and models didn’t have to predict zeros.

**Conclusion**: Removing non-buyers improved predictive performance due to reduced noise and more consistent spending patterns among purchasers.

**Comparison & Interpretation**:

* Random Forest outperformed MLP in both accuracy and stability.
* MLP was more prone to overfitting and struggled with generalization.