**Spam Classification Task**

Goal: Classify spam vs. non-spam using Spambase dataset. I built two models:

**Model 1 – Maximize Accuracy**

* **Model**: GradientBoostingClassifier (tuned via Optuna)
* **Best Hyperparameters**:
  + n\_estimators=160,
  + max\_depth=7,
  + learning\_rate=0.192397,
  + subsample=0.761491,
  + min\_samples\_split=8,
  + min\_samples\_leaf=2,
  + random\_state=42
* **Performance**:
  + Accuracy: **96%**
  + Precision: 0.97
  + Recall: 0.93
  + F1 Score: 0.95
  + ROC AUC: 0.985
  + Avg. Cost (FN=10, FP=1): 0.000

**Confusion Matrix**:

* TN = 519, FP = 12, FN = 28, TP = 362

**Model 2 – Minimize Cost**

* **Model**: LogisticRegression (tuned with Optuna)
* **Cost Function**: 10×FN + 1×FP
* **Best Hyperparameters**:
  + C = 13.011741761419996,
  + penalty = 'l2',
  + solver = 'lbfgs',
  + random\_state = 42,
  + max\_iter = 478,
* **Performance**:
  + Accuracy: **92%**
  + Precision: 0.92
  + Recall: 0.88
  + F1 Score: 0.90
  + ROC AUC: 0.970
  + Avg. Cost (FN=10, FP=1): **0.5299**

**Confusion Matrix**:

* TN = 503, FP = 28, FN = 46, TP = 344

**Comparison Table**

| **Metric** | **Accuracy Model**  **[Gradient Boost Model]** | **Cost-Sensitive Model**  **[Logistic Regression]** |
| --- | --- | --- |
| Accuracy | 96% | 92% |
| ROC AUC | 0.985 | 0.970 |
| Avg. Cost (10:1) | 0.000 | **0.5299** |

**Interpretation**:

* Accuracy model does better on traditional metrics.
* Cost-sensitive model sacrifices a bit of accuracy but reduces costly false negatives.
* This is especially useful in spam filtering where missing spam (FN) is more expensive.

**Final Reflections**

* I explored 6+ regression models and 4+ classification models.
* Best practices like normalization, Optuna tuning, and nested CV were consistently applied.
* Trade-offs between accuracy and cost are domain-specific. Business context matters.

**Best Models Summary**:

* **Spending Prediction**: RandomForestRegressor (restricted dataset)
* **Spam Classification**:
  + Task (i): GradientBoostingClassifier
  + Task (ii): LogisticRegression (Cost-sensitive)

All code and visualizations are provided in the submitted notebooks.