Project Proposal — *Retail Customer Purchase Prediction*

Team Members: Benjamin Tran, Matthew Nguyen, Victor Bui, Gustavo Buenrostro

Introduction

The problem we will be investigating focuses on retail customer purchases. This topic intrigues us because it allows us to explore patterns in consumer behavior and develop models to predict future purchasing trends.

To develop a clean, supervised ML task to apply course methods end-to-end (preprocessing → modeling → evaluation → ablation) to build personal understanding of e-commerce purchase intent for practical insights (targeting, promotions, remarketing).

Background Readings

- Tom Mitchell, *Machine Learning* — supervised learning setup and evaluation.
- Russell & Norvig, *Artificial Intelligence: A Modern Approach* — classification, metrics, and model comparison.
- Goodfellow, Bengio, Courville, *Deep Learning* — feedforward neural networks (MLP) fundamentals.

Dataset

- UCI *Online Shoppers Purchasing Intention* (~12k sessions; label Revenue). We will extract actionable session features (e.g., page duration, bounce-like signals, product-page ratios).
- Kaggle *Online Retail II* aggregated to session/customer RFM (Recency, Frequency, Monetary) for a small robustness check.

Methodology

To reprocess data (one-hot categoricals, scale numerics, train/val/test split, leakage audit). To the models: **Logistic Regression** (baseline), **Decision Tree → Random Forest → Gradient Boosting** (if permitted), and a shallow **MLP** (ReLU, dropout/L2, sigmoid output).

Approach

(1) Audit data and engineer features → train LR baseline → add tree ensembles → add MLP. (2) Choose best model by validation, then finalize on held-out test set. (3) Ablations: compare feature groups (behavioral, temporal, RFM) and model classes (linear, tree, MLP). (4) Conclude with business takeaways on which signals drive purchase intent.

Evaluations

- **Quantitative:** prioritize **ROC-AUC** (primary) and **F1/Precision/Recall** to address potential class imbalance; report accuracy for context.
- **Qualitative:** ROC curves, confusion matrices; tree-based feature importance; brief learning curves (MLP).
- **Expectation:** the advanced tree ensembles and the MLP will outperform logistic regression, illustrating the trade-off between interpretability and predictive power on structured retail data and yielding practical insight into which behaviors most influence conversions.