Machine Learning Projects in Finance

Course : Introduction to Machine Learning in Finance

Project 1 — Global Commodity Trade Data for GDP or CPI Estimation

Dataset: Global Commodity Trade Statistics (UN).

Aim: Use international commodity trade flows as an **alternative indicator** to nowcast or forecast macroeconomic variables such as GDP growth, CPI inflation, or industrial production.

How to Realise the Project:

- Select a set of countries (OECD, BRICS, or EU).
- Aggregate import/export volumes by commodity groups (energy, metals, agriculture).
- Build features: month-on-month growth, trade balance, volume vs value, etc.
- Train ML models (regression, tree-based, or LSTM if time series) to predict macro variables.
- Evaluate predictive power and economic interpretability.

Additional Data Needed: Yes — macro indicators such as GDP or CPI from the World Bank, OECD, or FRED for validation.

Project 2 — Yield Curve Prediction Using Machine Learning

Dataset: US Treasury Yield Curve Data (FRED) — for example: FRED: Treasury Constant Maturity Series.

Aim: Use machine learning methods to **predict the future shape of the yield curve** (level, slope, curvature) using historical term-structure data and macroeconomic signals.

How to Realise the Project:

- Collect daily Treasury yields across maturities (1M to 30Y).
- Construct yield curve factors:
  - level (long-term rates),
  - slope (long vs short rates),
  - curvature (medium-term bending).
- Create features: past yield values, lagged factors, macro indicators (optional).
- Train ML models such as:
  - Random Forest,
  - Gradient Boosting,

  - LSTM or GRU networks (sequence modelling).

- Predict next-day or next-month yield curve factors or the full curve for each maturity.
- Reconstruct the curve from predicted factors and evaluate forecasting accuracy.

Additional Data Needed: Optional — macro variables (inflation, Fed funds rate, employment indicators) may improve predictions.

Project 3 — Credit-Risk Modelling with Default of Credit Card Clients

Dataset: Default of Credit Card Clients (UCI/Kaggle).

Aim: Build and compare machine learning models to estimate the **probability of default (PD)** of borrowers using demographic information and past payment behaviour.

How to Realise the Project:

- Clean and preprocess categorical and numerical features.
- Handle class imbalance (oversampling or class weights).
- Train Logistic Regression, Random Forest, XGBoost, and/or MLP.
- Evaluate with AUC, F1-score, confusion matrix, and calibration.
- Use SHAP or LIME to interpret important risk drivers.

Additional Data Needed: None required — the dataset is self-contained.

Project 4 — News-Based Stock Market Prediction (NLP + Finance)

Dataset: Daily News for Stock Market Prediction (Kaggle).

Aim: Combine NLP and financial time-series to predict daily stock market direction (up/down) or return magnitude for the Dow Jones or S&P 500.

How to Realise the Project:

- Preprocess news headlines (cleaning, tokenisation, embeddings).
- Convert text into features using TF-IDF, Word2Vec, or BERT embeddings.
- Merge text features with daily index prices.
- Build models: Logistic Regression, LSTM, or Transformer.
- Backtest a small trading strategy using predicted signals.

Additional Data Needed: Optional — more recent news sources or additional financial indicators.

Project 5 — Stock Market Dataset for Portfolio Allocation

Dataset: Stock Market Dataset (NASDAQ Universe, Kaggle).

Aim: Build a machine-learning-driven **portfolio allocation model** (minimum variance, maximum Sharpe, or risk-parity) using predictive signals extracted from returns and volatility forecasts.

How to Realise the Project:

- Select a subset of stocks (e.g., 50–200 tickers).
- Compute features: returns, volatility, sector categories, correlations.
- Predict next-day or next-week returns/volatility using ML models.
- Use predicted risk/return to construct a portfolio (Markowitz, ML-based).
- Evaluate performance: Sharpe ratio, drawdown, turnover.

Additional Data Needed: Optional — sector classifications or macro variables to improve predictions.