Big Data for Finance

Group Project Brief

Spring 2026

1. Overview

The group project accounts for 25% of your final module grade. You will work in pre-assigned groups of 5 students to apply machine learning methods to a financial problem. The project tests your ability to implement ML techniques covered in class, interpret results critically, and communicate findings effectively.

1.1 Project Assignment

Each group has been assigned a specific sub-option. You will find your assignment in the Group_Project_Assignments.xlsx file posted on Insendi. Sub-options are assigned to ensure that each group works on a unique dataset/market combination, enabling fair comparison across groups and avoiding duplication of effort.

Your assigned sub-option determines your dataset and research question. You may not switch sub-options with another group.

1.2 Deliverables

1. **Presentation (Week 8):** 10 minutes per group. Please consider the time limit when preparing the slides and presentation materials.
2. **Slides used during the presentation**: The slides to be used for the presentation. Consider max 10/12 slides for the entire presentation due to time constraints. Note that you do not need to submit code. However, I expect that in the presentation you will clearly discuss all methodological choices, including data cleaning and preprocessing, cross-validation strategy, and hyperparameter tuning.

1.3 Grading Criteria

Criterion	Weight	Description
Implementation Quality	40%	Correct methodology, appropriate validation, sound hyperparameter tuning
Interpretation & Analysis	40%	Understanding results, economic intuition, critical evaluation
Deployment Considerations	20%	Production readiness, monitoring, documentation

2. Project Options

There are three main options (A, B, D), each with multiple sub-options. Your group has been assigned one specific sub-option. The descriptions below explain the requirements for each.

Option A: Asset Return Prediction

Apply regression methods to predict asset returns using macroeconomic, technical, and fundamental predictors. This option applies the market-timing framework from Weeks 2-3 across asset classes, including bonds, equities, commodities, and cryptocurrencies.

Core Requirements

- Implement at least 3 regression methods (e.g., Ridge, Lasso, Random Forest, etc)
- Use proper time-series cross-validation (expanding or rolling window)
- Evaluate out-of-sample R^2 and economic performance. You are free to consider other alternative statistical accuracy measures or economic evaluation metrics.
- For the statistical and economic comparison, use a rolling or expanding window mean as the benchmark (see the exercises shown in class and the tutorials).

ID	Sub-Option	Indicative objective of the project
A1	U.S. Treasury Bonds	Predict monthly excess returns on 10-year Treasury bonds using yield curve slopes, term spreads, and macro variables. Assess whether bond returns are more predictable than equity returns.
A2	U.S. Investment-Grade Corporate	Predict excess returns on investment-grade corporate bonds. Evaluate whether credit spread dynamics improve predictability beyond Treasury factors.
A3	U.S. High-Yield Bonds	Predict high-yield bond returns using default spreads, equity volatility, and liquidity measures. Analyse performance across different economic regimes (crisis vs. normal).
A4	U.S. TIPS	Predict TIPS returns using inflation expectations and real yield dynamics. Compare to nominal Treasury timing.
A5	Cryptocurrencies	Predict the return of a cryptocurrency of your choice (e.g., Bitcoin, ETH, Litecoin, Ripple, etc) using momentum, volume, and volatility indicators. Discuss statistical predictability and economic evaluation in comparison to more mainstream assets, such as the market portfolio return (see results in the tutorial).
A6	Global Regional Markets	Predict the return on a regional market of your choice between Europe, Japan, Asia Pacific ex Japan, North America. Discuss statistical predictability and economic evaluation in comparison to more mainstream assets, such as the market portfolio return (see results in the tutorial).
A7	U.S. Industry Portfolios	Predict the return on an industry portfolio based on the Fama-French classification using momentum, volatility, and macroeconomic indicators (you could use the same we used in class from the GWZ data). Evaluate sector-timing strategy.
A8	Global Commodity Prices	Predict the return on a commodity of your choice using momentum, volatility, dollar correlation, and inflation beta.

Option B: Credit Risk Modeling

Build classification models to predict loan defaults. This applies the classification methods from Week 4 to credit datasets different from the LendingClub data used in class.

Core Requirements

- Implement at least 3 classification methods (e.g., Logistic Regression, Classification Tree, Random Forest)
- Address class imbalance appropriately (e.g., class weights or threshold adjustment)
- Statistical evaluation (e.g., AUC-ROC, precision-recall, etc.)
- Develop business recommendations (e.g., approval thresholds, economic costs)

ID	Sub-Option	Indicative objective of the project
B1	Taiwan Credit Card Default	Predict credit card default using payment history and demographic features. Develop a credit scorecard and recommend approval thresholds that balance default risk with customer acquisition.
B2	German Credit (Statlog)	Predict creditworthiness using mixed numerical and categorical features. Address the small sample challenge and discuss model interpretability for regulatory compliance.
B3	Polish Company Bankruptcy	Predict corporate bankruptcy using financial ratios. Compare early warning indicators across different prediction horizons (1-year vs. 5-year).
B4	South German Credit	Predict loan default with emphasis on model interpretability. Provide feature importance analysis and actionable lending criteria.

Option D: Unsupervised Learning Applications

Apply dimensionality reduction (PCA) and clustering methods to discover structure in financial data. This extends Week 5 concepts to new applications.

Core Requirements

- Apply PCA and/or clustering methods (K-means, hierarchical)
- Determine the optimal number of components/clusters using appropriate criteria
- Provide an economic interpretation of the discovered structure
- Demonstrate downstream application (forecasting, risk management, etc)

ID	Sub-Option	Indicative objective of the project
D1	Cryptocurrency Market Structure	Extract principal components from a cross-section of cryptocurrency returns. Identify "groups" and assess whether PCA provide any benefit in terms of out-of-sample forecasting as a downstream application.
D2	Multi Regime Detection	Cluster time periods using cross-asset returns (equities, bonds, commodities). Identify market regimes and provide an economic narrative to the results (e.g., separate periods of high equity/low bond returns vs high bond/low equity returns).

D3	Industry Clustering	Extract principal components from U.S. industry portfolio returns. Cluster industries by return correlations. Compare discovered clusters to traditional sector classifications (e.g., GICS sectors)
D4	Credit Applicant Segmentation	Cluster borrowers in Taiwan Credit Card or German Credit data based on some preferred household characteristics (or maybe more than one). Discuss the implications of clustering for downstream credit risk assessment.

Some Hints on How to Structure the Project Flow

To ensure consistent quality and a manageable workload, a sensible starting point is to simply follow the tutorial structure for a given application.

Option A: Follow Tutorials 2–3 (Return Prediction)

Your project may mirror the Week 2 and Week 3 tutorials:

Step 1 – Data Preparation: Load data, create features (lags, transformations), chronological train/test split.
Step 2 – Baseline (OLS): Fit OLS regression. Evaluate out-of-sample R^2.
Step 3 – Regularised Regression: Fit Ridge, Lasso, Elastic Net with cross-validated λ.
Step 4 – Tree-Based Methods: Fit Random Forest and Gradient Boosting/XGBoost.
Step 5 – Statistical Evaluation: Report out-of-sample R^2 vs. historical mean benchmark.
Step 6 – Economic Evaluation: Market timing strategy: Sharpe ratio, cumulative returns.

Key insight: Statistical accuracy ≠ Economic value. Report both statistical and economic performance!

Option B: Follow Tutorial 4 (Credit Risk)

Your project may mirror the Week 4 tutorial:

Step 1 – Data Preparation: Handle missing values, encode categoricals, train/test split.
Step 2 – Train Models: Logistic Regression, Random Forest, Gradient Boosting.
Step 3 – Confusion Matrix: Understand TP, FP, TN, FN in credit context.
Step 4 – Evaluation Metrics: Precision, Recall, F1, AUC-ROC curves.
Step 5 – Threshold Optimisation: Do NOT use 0.5. Optimise based on asymmetric costs.
Step 6 – Business Decisions: Approve/deny rules, risk-based pricing tiers.

Key insight: Costs are asymmetric—missing a defaulter costs more than denying a good borrower.

Option D: Follow Tutorial 5 (Unsupervised Learning)

Your project may mirror the Week 5 tutorial:

Step 1 – Data Preparation: Standardise all features (critical for PCA/clustering).
Step 2 – PCA: Scree plot, choose components, interpret loadings economically.
Step 3 – Clustering: K-means or hierarchical. Elbow method.

Step 4 – Interpret Clusters: Name clusters economically (e.g., 'Crisis', 'Normal', 'Bull').
Step 5 – Downstream Application: Forecasting using the principal components, regime-switching based on clusters, or estimating segment-specific models based on clusters.

Key warning: Avoid look-ahead bias! Fit PCA/clustering only on past data when backtesting.

3. Presentation Guidelines

Presentations will take place in Week 8. Each group has exactly 10 minutes.

Suggested Structure

- Introduction and motivation (1-2 minutes)
- Data and methodology overview (2-3 minutes)
- Key results (3-4 minutes)
- Conclusions and deployment considerations (1-2 minutes)

Tips

- Practice timing - 10 minutes is strict
- Focus on insights, not technical details
- Use clear visualisations; avoid cluttered tables
- All group members should contribute to the presentation
- Be prepared to answer questions about methodology choices. Some of the questions could be about:
    - Data Preprocessing**:** Feature scaling/standardisation approach. Feature engineering decisions (lags, transformations).
    - Cross-Validation: Validation scheme (expanding window, rolling window, blocked CV). Training/validation/test split ratios and dates.
    - Hyperparameter Tuning: Which hyperparameters were tuned for each method. Search method (grid search, random search).

4. Data Access

All datasets are publicly available and require no registration. See the accompanying Excel file (Project_Data_Sources.xlsx) for direct links, download instructions, and suggested predictors.

Key data sources:

- **FRED (Federal Reserve Economic Data):** fred.stlouisfed.org - U.S. macro and bond data
- **UCI Machine Learning Repository:** archive.ics.uci.edu - Credit risk datasets
- Ken French Data Library: mba.tuck.dartmouth.edu/pages/faculty/ken.french/data_library.html – Factor returns, industry portfolios, regional market data (pandas_datareader)

- Kaggle: kaggle.com/datasets – Cryptocurrency historical prices (free account required, direct CSV download)

5. Frequently Asked Questions

Assignment Questions

Q: Can I switch sub-options with another group?

A: No. Sub-options are assigned to ensure unique coverage across groups and fair grading. Please work with your assigned sub-option.

Q: Where do I find my group's assignment?

A: Check the Group_Project_Assignments.xlsx file on Insendi. It lists each group's assigned sub-option, dataset, and objective.

Data Questions

Q: I cannot access the data. What should I do?

A: All data sources are public and do not require registration. Check the Project_Data_Sources.xlsx file for direct URLs. If a specific link is broken, use the main website to search for the dataset.

Q: How much historical data should I use?

A: Use the maximum available history for your asset class. For FRED data, 20+ years is typical. For cryptocurrencies, use data from 2020 onwards. You could use even longer data when sufficient coins existed.

Q: Can I use additional data sources beyond those listed?

A: Yes, as long as they are publicly available and properly cited. Document any additional sources in your presentation.

Methodology Questions

Q: How many models should I implement?

A: Minimum 3 methods as specified in each option. You may include more if time permits, but depth is more important than breadth.

Q: What cross-validation method should I use?

A: For time-series data (Options A, D1-D3), use expanding window or rolling window CV. Never use standard k-fold CV as it causes look-ahead bias. For cross-sectional credit data (Option B), a standard k-fold is acceptable, but stratified sampling is recommended.

Q: How should I handle class imbalance in Option B?

A: You may use class weights, threshold adjustment, or a combination. Document your choice and justify it.

Presentation Questions

Q: Should all group members present?

A: Yes, all members should contribute to the presentation. You may divide sections among yourselves.

Q: Will we receive the presentation schedule in advance?

A: Yes, the presentation order will be announced at least one week before Week 8.

Grading Questions

Q: How is the group grade determined?

A: The group receives a single grade based on the slides and presentation. Peer evaluation may be used to adjust individual grades if contributions are significantly unequal.

Q: What if our models perform poorly?

A: Poor predictive performance does not automatically mean a poor grade. Critical interpretation of negative results (e.g., "why market timing is difficult") can demonstrate a strong understanding. Methodology and interpretation matter more than raw performance.

6. Important Dates

Milestone	Date
Group and sub-option assignments released	Week 4
Presentations	Week 8 (in class)
Submission of the slides	23rd February 2026 9 am UK time

Work submitted up to one (1) day after the assessment deadline (date and time) will be marked but capped at the pass mark. Work submitted more than one (1) day late will not be accepted as a valid attempt, and a mark of zero will be recorded. This is the default university penalty for late submissions of assessed work and will be deviated from only in exceptional circumstances at the discretion of the Programme Director.

7. Contact

For project-related questions:

- **General questions:** Email d.bianchi@imperial.ac.uk
- **Administrative questions:** Email a.dean@imperial.ac.uk
- **Tutorial support:** Email yoshita.lakka24@imperial.ac.uk

Please check the FAQ before emailing, as your question may already be answered.