**Predicting Stock Return Jumps**

**Introduction**

This report investigates the prediction of extreme stock return jumps using firm-level characteristics and macroeconomic indicators. A “jump” is defined as a one-month return exceeding 10% in absolute value. I examine the time-series behavior of jump frequencies, explore relationships with macro variables (Unemployment, CPI, Industrial Production, and VIX), and evaluate predictive models ranging from logistic regressions with regularization to tree-based methods. The analysis uses CRSP monthly stock data supplemented with FRED macroeconomic series.

**Jump Frequency and Macro Indicators**

I first compute the fraction of stocks experiencing jumps in each month. This jump share is plotted against macroeconomic indicators, especially the VIX.

A graph showing a number of blue lines

AI-generated content may be incorrect.

Observations:

* Jump frequency spikes during crisis periods such as the dot-com crash, the 2008 financial crisis, and the COVID-19 shock.
* The VIX rises in near-lockstep with the jump share, making it a strong contemporaneous indicator of jump risk.
* Unemployment and Industrial Production lag more, aligning with economic downturns but not offering clear short-term predictive signals.
* CPI changes appear weakly related to jump risk.

Conclusion: among the indicators, the VIX shows the most potential as a near-term precursor of high jump likelihood .

**Model Evaluation**

I estimated several classification models to predict whether a stock will jump in the following month. Models are trained on 1996–2017 data and tested on 2018–2023. Out-of-sample performance is assessed by the Area Under the ROC Curve (AUC) and the Kolmogorov–Smirnov (KS) statistic.

**Logistic Regression**

A baseline logistic regression provides a reference point.

A graph of a curve

AI-generated content may be incorrect.

* AUC = 0.693, KS = 0.290
* Performance is modest but significantly better than random guessing

**LASSO Logistic**

Apply an L1 penalty to select predictive features.

A graph with a line

AI-generated content may be incorrect.

* AUC = 0.690, KS = 0.281
* Sparsity yields interpretability but sacrifices little predictive power

**Post-LASSO Logistic**

Refit a standard logistic on variables selected by LASSO.

A graph of a logistic

AI-generated content may be incorrect.

* AUC = 0.683, KS = 0.267
* Slight decline in predictive accuracy relative to baseline

**Ridge Logistic**

An L2-penalized logistic regression controls for multicollinearity.

A graph with a line

AI-generated content may be incorrect.

* AUC = 0.690, KS = 0.282
* Results nearly identical to LASSO, highlighting robustness

**XGBoost**

Finally, evaluated a boosted tree model.

A graph of a positive rate

AI-generated content may be incorrect.

* AUC = 0.711, KS = 0.326
* Outperforms logistic models, indicating nonlinear interactions and threshold effects matter for jump prediction

**Comparative Summary**

* Logistic-based models (plain, LASSO, Ridge, Post-LASSO) achieve AUCs around 0.68–0.69
* XGBoost improves performance to 0.71 AUC, the best among tested models
* The KS statistic similarly favors XGBoost (0.326 vs ~0.28 for logistic)
* This suggests machine learning approaches capture nonlinearities and higher-order interactions that logistic regressions miss

**Conclusion**

The analysis confirms that stock return jumps cluster around macroeconomic crises and correlate strongly with volatility indices like the VIX. While logistic regression provides a transparent baseline, its predictive power is limited. Regularization does not significantly improve accuracy. In contrast, tree-based ensemble methods like XGBoost deliver stronger out-of-sample performance, making them a promising tool for forecasting extreme stock returns.

Future work could extend the feature set to include option-implied measures, sentiment indices, or more granular firm fundamentals, and test more advanced models such as neural networks or random forests.