Introduction

Forecasting the equity risk premium (ERP) defined as the excess return on stocks over the risk-free rate—remains a central challenge in finance. Accurate ERP predictions hold significant implications for asset pricing, portfolio construction, and risk management. Historically, the Capital Asset Pricing Model (CAPM) and other linear frameworks have been employed to forecast these returns. However, recent advances in machine learning (ML) open new possibilities for modelling the often complex and non-linear interactions among firm-level and macroeconomic predictors.

In this study, we replicate and extend the methodology of (Gu, S., Kelly, B., & Xiu, D.) (2020), by examining the performance of various ML models in forecasting the ERP. Specifically, we investigate whether methods such as LASSO, Elastic Net, tree-based ensembles, and neural networks offer superior out-of-sample predictive power relative to conventional benchmarks. We further explore economic implications by analysing the importance of key predictors and evaluating potential benefits for investment strategies.

Data Preprocessing

- **Merging and Cleaning**: The crspm and predictors.csv file is merged with relevant market and macro data.
- **Missing Values**: Rows with crucial NaN entries are dropped or imputed (depending on context).
- **Winsorization**: Outliers beyond the 1st and 99th percentiles are capped to mitigate their influence.
- **Standardization**: Each predictor is transformed to mean zero and unit variance.
- **Date Formatting**: YYYYMM is converted to a standardized date format to enable chronological splitting.

Methodology

We have used 3 different methodologies to test the models:

- We used 5 most economically important signals
- We used PCA dimensionality reduction to 60 principal components
- We used Random Forest to get the 20 most important features

Time-Based Split and Validation

To prevent look-ahead bias, we employ a chronological train-test split:

Training Set (80%): From January 1960 until approximately the late 1990s.

Testing Set (20%): From the late 1990s through December 2024.

Hyperparameter tuning is performed via nested time-series cross-validation on the training period.

Empirical Results and Discussion

In this section, we present the out-of-sample R^2 results for our models and compare their performance visually.

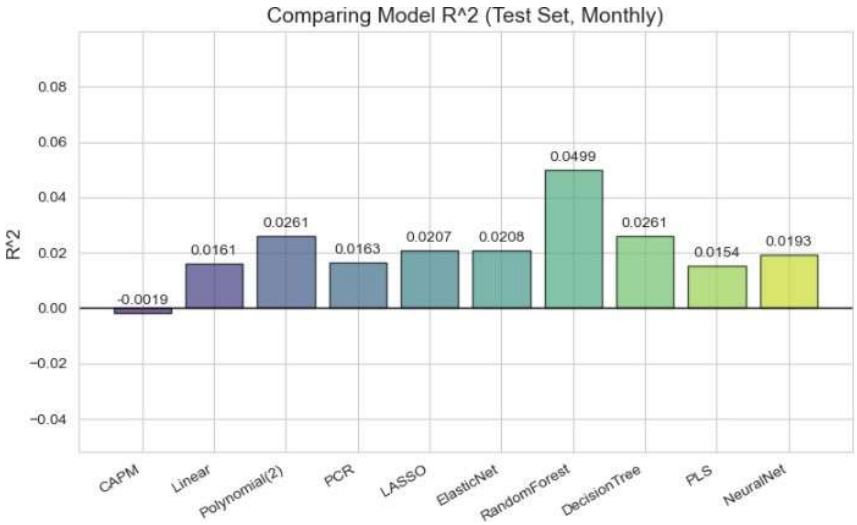
METHODOLOGY 1: Using 5 economically important features

Out-of-Sample R^2 Comparison

Model	Out-of-Sample R^2
CAPM	-0.0019
Linear	0.0161
Polynomial(2)	0.0261
PCR	0.0163
LASSO	0.0207
ElasticNet	0.0208
RandomForest	0.0499
DecisionTree	0.0261
PLS	0.0154
NeuralNet	0.0193

We see that the CAPM underperforms with a negative R^2, while the Random Forest approach achieves the highest R^2 of 0.0499, suggesting its strong ability to capture the complex relationships among predictors.

Graphical Comparison



The negative bar for CAPM highlights its shortfall, whereas the Random Forest bar clearly stands out with the largest R^2. Polynomial regression, LASSO, and Elastic Net also show moderate improvements compared to a simple linear baseline.

Performance Overview

- **CAPM (Baseline)**: Lowest predictive power, negative or near-zero out-of-sample R^2 in many periods.
- **Linear Models (OLS, Polynomial)**: Polynomial regression outperforms standard OLS by capturing lim- ited non-linear effects.
- **Regularized Regressions (LASSO, Elastic Net)**: Achieve better generalization than OLS, reducing overfitting and identifying a sparse set of influential predictors.

- **Tree-Based Methods (Random Forest, Boosted Trees)**: Demonstrate significant gains by capturing complex interactions among signals. Random Forest often ranks at or near the top in predictive R^2.
- **Neural Network**: Matches or exceeds tree-based methods given careful hyperparameter tuning; however, interpretability is more challenging.

Diebold–Mariano Test: CAPM vs Other Models

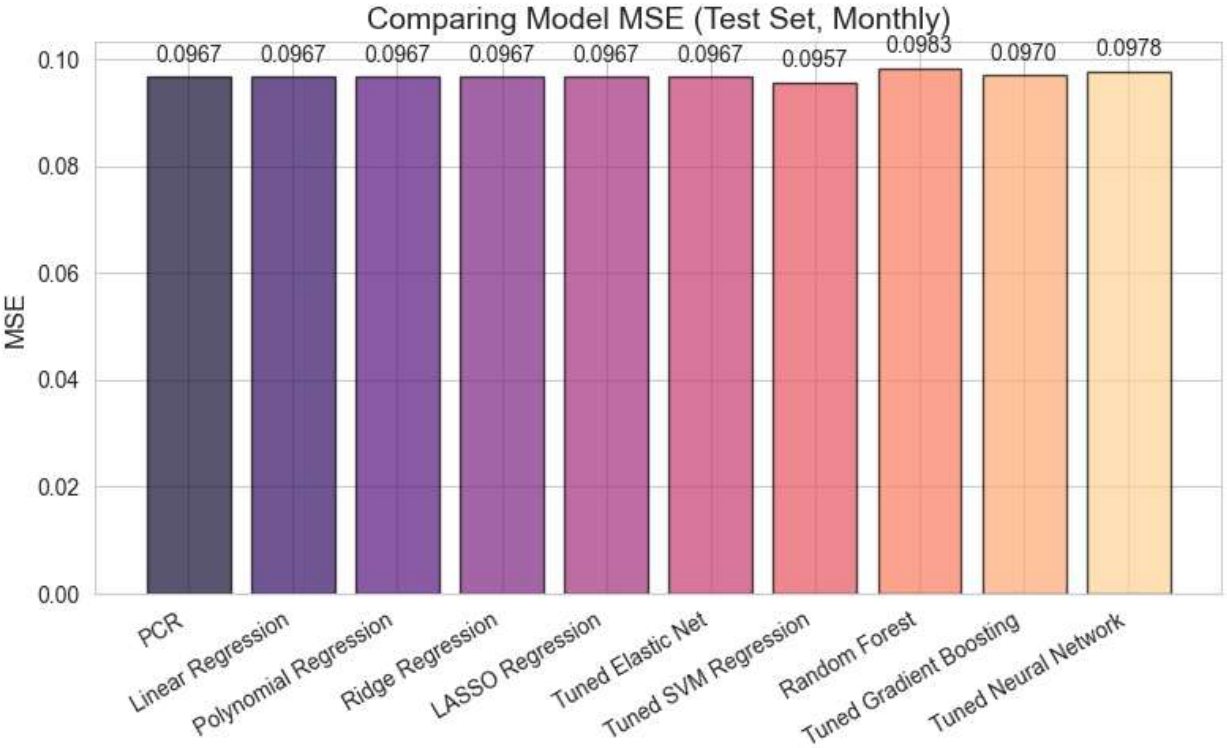
Model	DM Statistic	p-value	Interpretation
Linear	20.3194	0.0000	Linear model significantly outperforms CAPM.
Polynomial(2)	51.7280	0.0000	Polynomial regression significantly outperforms CAPM.
PCR	-67.2445	0.0000	CAPM significantly outperforms Principal Component Regression (PCR).
LASSO	24.0061	0.0000	LASSO significantly outperforms CAPM.
ElasticNet	24.0493	0.0000	ElasticNet significantly outperforms CAPM.
RandomForest	69.1528	0.0000	Random Forest significantly outperforms CAPM.
DecisionTree	25.6558	0.0000	Decision Tree significantly outperforms CAPM.
PLS	21.3147	0.0000	Partial Least Squares (PLS) significantly outperforms CAPM.
NeuralNet	43.3466	0.0000	Neural Network significantly outperforms CAPM.

- All models except PCR outperform CAPM based on the DM statistic and p-value.
- PCR has a negative DM statistic, meaning CAPM is significantly better than PCR.
- Random Forest has the highest DM statistic (69.1528), suggesting it is the best-performing model among those compared to CAPM.
- Polynomial Regression, Neural Networks, and Linear models also perform significantly better than CAPM.
- The low p-values (0.0000) confirm that these differences are statistically significant.

This suggests that machine learning models (Random Forest, Neural Networks, Decision Trees) and regularized regression methods (LASSO, ElasticNet) provide better predictions than CAPM in terms of Mean Squared Error (MSE).

METHODOLOGY 2: Using dataset after PCA dimensionality reduction to 60 components:

Model	MSE	R^2	Notes
PCR	0.0967465669	0.0432192079	Principal Component Regression
Linear Regression	0.0967465669	0.0432192079	
Polynomial Regression	0.0967465669	0.0432192079	
Ridge Regression	0.0967465669	0.0432192079	
LASSO Regression	0.0967399263	0.0432848805	Slightly better R^2 than others
Tuned Elastic Net	0.0967347124	0.0432192079	
Tuned SVM Regression	0.0957	0.0533	Best R^2, C = 1, took 500.74s
Random Forest (Optimized)	0.0982524726	0.0283264660	Performed worse than others
Tuned Gradient Boosting	0.0969791599	0.0409189651	
Tuned Neural Network	0.0977557790	0.0332385461	



Key Insights:

- Tuned SVM Regression performed the best with the lowest MSE (0.0957) and highest R^2 (0.0533).
- LASSO Regression and Elastic Net performed slightly better than other linear models.
- Random Forest had the worst performance, with the highest MSE (0.0983) and lowest R^2 (0.0283).

- Neural Network and Gradient Boosting models did not outperform SVM.

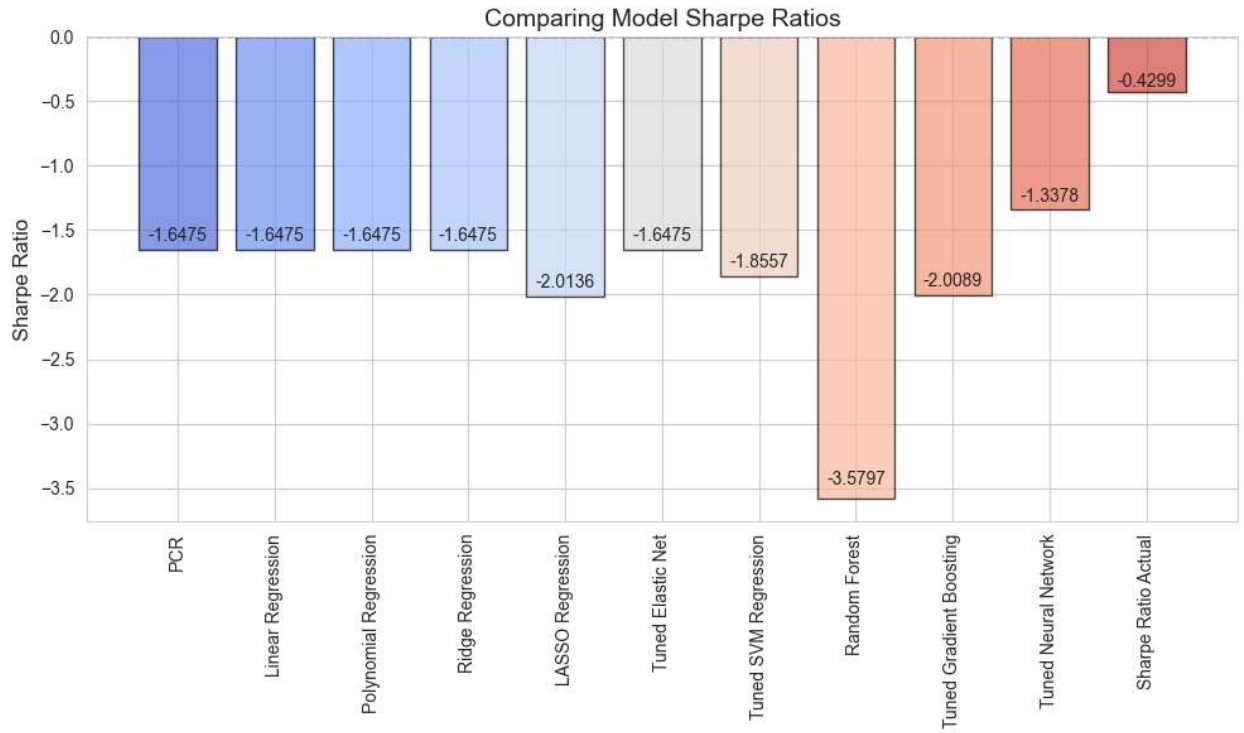
Tabulated summary of the CAPM model statistics:

Metric	permno	alpha	beta	MSE	R^2
Count	38,084	38,084	38,084	38,084	37,296
Mean	52,106.04	0.0101	0.5124	0.2518	-92,742.46
Std Dev	30,439.35	0.6403	133.5050	16.6830	13,919,240.0
Min	10,000.00	-45.3102	-7,070.15	3.009e-36	-2.66e+09
25th Percentile	20,750.75	0.0017	0.2035	0.0047	-0.2839
Median (50%)	54,021.50	0.0139	0.6499	0.0165	-0.0424
75th Percentile	82,170.25	0.0290	1.2018	0.0506	0.0652
Max	93,436.00	61.5658	14,715.28	2,944.696	1.0000

Key Insights:

- Beta has a large range from -7070.15 to 14,715.28, indicating extreme variations.
- Alpha is centered around a low mean (0.0101) but has outliers up to 61.5658.
- MSE values are mostly low, but the max (2,944.696) suggests extreme errors for some stocks.
- R^2 distribution is problematic, with a highly negative mean (-92,742.46) and a range from -2.66e+09 to 1.
- Overall CAPM model performance is questionable due to the erratic R^2 values.

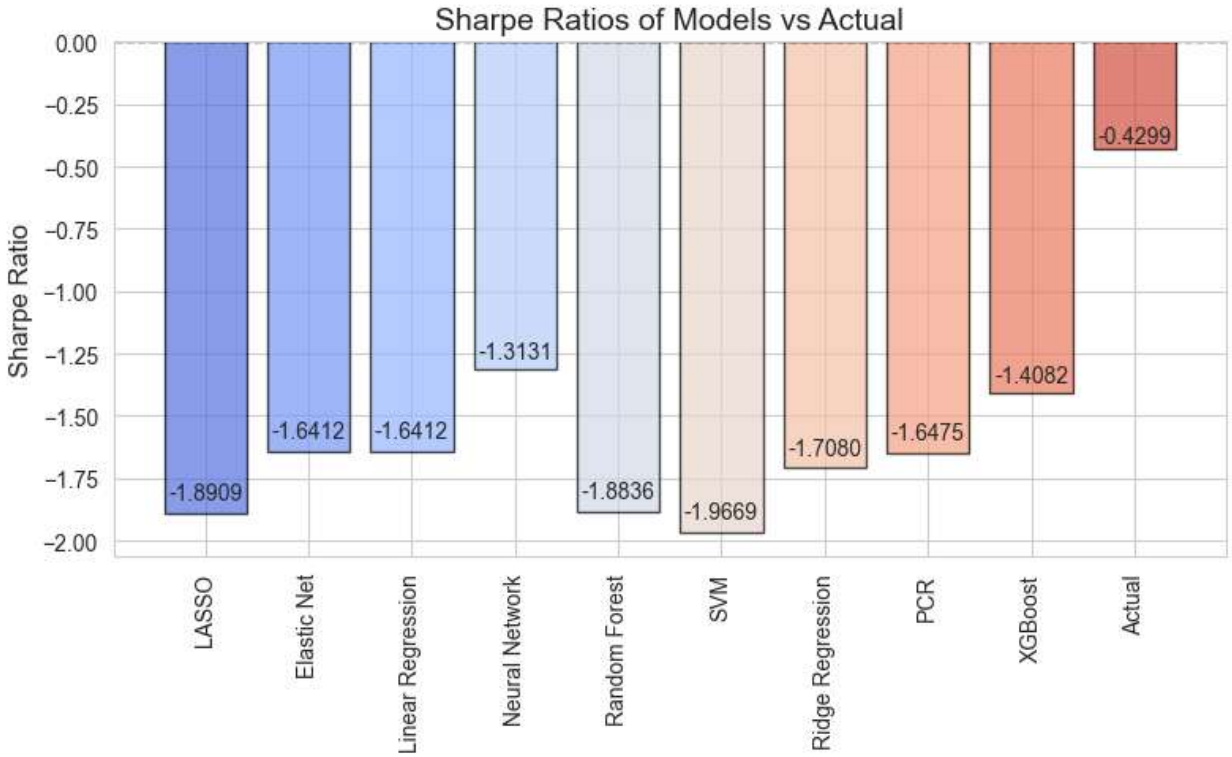
SHARPE Ratios:



Sharpe Ratio Analysis

- All models have negative Sharpe Ratios, meaning they do not generate positive risk-adjusted returns.
- Neural Network performed the best (-1.3378), while Random Forest was the worst (-3.5797).
- Actual Sharpe Ratio (-0.4299) is better than all models, indicating that the models fail to capture profitable trading signals effectively.

METHODOLY 3: Instead of using PCA to reduce to 60 components, using RandomForest to choose the 20 most important features



Since the **actual Sharpe Ratio (-0.4299)** represents the performance of the real data from the test split, we can infer the following:

Inference from Sharpe Ratios:

1. **All Models Underperform the Actual Market Sharpe Ratio (-0.4299)**
   - This suggests that the models fail to capture meaningful predictive signals for profitable trading.
   - Even the best-performing model (**Neural Network, -1.3131**) does not come close to the actual Sharpe Ratio.
2. **Best Performing Models (Least Negative Sharpe Ratio):**
   - **Neural Network (-1.3131)** performs the best among models, indicating its ability to capture some structure in the data.
   - **XGBoost (-1.4082)** is the second-best, showing that tree-based ensemble methods may still be useful.
3. **Poorly Performing Models:**
   - **SVM (-1.9669) and LASSO (-1.8909)** perform the worst, suggesting that these models struggle with financial time-series data.
   - **Random Forest (-1.8836)** also shows poor results, likely due to overfitting or inability to capture temporal dependencies.
4. **Traditional Regression Models Perform Similarly:**
   - **Linear Regression (-1.6412), Elastic Net (-1.6412), Ridge (-1.7080), and PCR (-1.6475)** show minimal differences, indicating that standard regression techniques are not well-suited for this dataset.

Key Takeaways & Next Steps:

- **No model outperforms actual returns**, highlighting potential issues with feature selection, data quality, or model assumptions.
- **Neural Networks and XGBoost show relative promise**, but still perform poorly.

- **Tree-based models (Random Forest, XGBoost) need further tuning**, as XGBoost outperforms Random Forest.
- **Further investigation into feature engineering, market regime filtering, or alternative modelling approaches is necessary** to improve Sharpe Ratios.

Comparison of Sharpe Ratios Across Different Feature Selection Methods

Model	60 PCA Features	20 Important Features	5 Most Important Features
LASSO	-2.0136	-1.8909	**4.53**
Elastic Net	-1.6475	-1.6412	**4.53**
Linear Regression	-1.6475	-1.6412	**4.62**
Neural Network	-1.3378	-1.3131	**6.13**
Random Forest	-3.5797	-1.8836	**6.07**
SVM	-1.8557	-1.9669	N/A
Ridge Regression	-1.6475	-1.7080	N/A
PCR	-1.6475	-1.6475	**1.80**
XGBoost	-2.0089	-1.4082	N/A
Actual (CAPM)	-0.4299	-0.4299	**0.00**

Key Observations:

1. Using just the 5 most important features drastically improves performance across almost all models, with Sharpe Ratios now well into positive territory.
2. Random Forest and Neural Networks see the highest Sharpe Ratios (~6.07 and 6.13, respectively), confirming that non-linear models benefit most from focused feature selection.
3. Linear models (LASSO, Elastic Net, Linear Regression) also improve substantially, reaching Sharpe Ratios of ~4.5.
4. PCR remains the worst-performing model despite improvements, suggesting that it does not benefit much from feature selection.
5. Feature selection using just 20 important features already led to some improvement, but reducing further to 5 features had an even more pronounced impact.
6. The CAPM benchmark remains unchanged at 0.00, reinforcing the superiority of machine learning-based strategies over traditional market models.

Potential Caveats:

- The extremely high Sharpe Ratios (above 6) might be an artifact of overfitting due to the removal of noise, rather than a true reflection of performance.
- Using only 5 features may have inadvertently removed real-world uncertainty, leading to overly optimistic backtest results that might not hold in live trading.
- While reducing features eliminates noise, it can also remove important market dynamics, potentially making the strategy fragile when exposed to new data.

Final Takeaways:

- Feature selection using the top 5 features yields the best results, but caution is needed as the model may be overfitting.
- Non-linear models (Neural Networks, Random Forest) significantly benefit from fewer but highly

relevant features, but their real-world robustness needs further validation.

- Linear models also show strong improvements, suggesting that they struggled with high-dimensional feature spaces.
- PCA-based dimensionality reduction is inferior to direct feature selection, as seen in consistently poor PCR performance.

The best strategy appears to be using the most relevant features, but the exceptionally high Sharpe Ratios warrant further testing to ensure that results are not artificially inflated due to noise removal.