A.1. Algorithm for Computing PBSV

We created the Python package anatomy to implement the algorithms for computing the oShapley-VI\textsubscript{p} and PBSV\textsubscript{p} metrics developed in Section 2 of the paper. The algorithms divide the estimation procedure into two steps: (1) evaluating the fitted models using coalitions of predictors from the sampled permuted orders and storing forecasts; (2) computing the Shapley-based metrics from the stored forecasts. After the models are evaluated in the computationally expensive first step, combinations of models and transformations of the forecasts can be evaluated inexpensively in the second step to compute the desired metric. By dividing the algorithm into two parts, the computationally expensive first step is embarrassingly parallelizable. Algorithm 1 provides the structure for the first step. Using results from the first step, the metrics can be computed inexpensively in the second step without the need to rerun the first step.

A.2. Inflation Predictors

The data for the inflation predictors are from three sources. The first is the FRED-MD database. Table A.1 lists the 121 variables from FRED-MD and the their abbreviations. The second source is the Institute for Supply Management, from which we use seven variables: Manufacturing Inventories Index (\texttt{man\_invent}), Manufacturing Production Index
Algorithm 1: Forecast Evaluation for Permuted Orders of Predictors

**Result:** $\hat{Y}$: $D \times K \times P \times 2 \times 2M$ array of forecasts for $D$ out-of-sample periods and $K$ models evaluated over coalitions of $P$ predictors deactivated and activated in $M$ forward and reversed permuted orders; $\bar{Y}$: $D \times K$ matrix of base predictions (i.e., model evaluations with empty coalitions)

**Input:** $\hat{F}$: $D \times K$ matrix of model prediction functions; $X$: $D$ training data matrices of sizes $T_t \times P$ for $t = 1, \ldots, D$; $\mathcal{X}$: $P \times D$ matrix, with columns comprised of vectors of $P$ predictors for producing forecasts; $M$: number of ordered permutations to draw from $\pi(P)$

1. Generate permutation matrix $O$ of size $M \times P$ containing $M$ permutations of $\{1, \ldots, P\}$
2. for $t = 1$ to $D$ do
   // loop over time
   for $k = 1$ to $K$ do
      // loop over models
      Store forecast with all predictors deactivated (base value): $\bar{Y}_{t,k} = \frac{1}{T_t} \sum_{s=1}^{T_t} \hat{F}_{t,k} \left( X_{s}^{(t)} \right)$
      for $m = 1$ to $M$ do
         // loop over permutations
         Copy order to preserve it across runs: $o = O_{m, \cdot}$
         for $i \in \{0, 1\}$ do
            // forward and reverse order
            Copy training data to preserve it across runs: $X^{(m)} = X^{(t)}$
            for $p \in \{o_1, \ldots, o_P\}$ do
               // loop over predictors
               Store forecast with previously activated predictors:
               $\tilde{Y}^{(o)}_{t,k,p,1,M+m} = \frac{1}{T_t} \sum_{s=1}^{T_t} \hat{F}_{t,k} \left( X_{s}^{(m)} \right)$
               Activate predictor $p$ in $X^{(m)}$ by setting all elements of column $p$ to $\mathcal{X}_{p,t}$:
               $X^{(m)}_{p} = \mathcal{X}_{p,t}$
               Store forecast with $p$ and previously activated predictors:
               $\tilde{Y}^{(o)}_{t,k,p,2,M+m} = \frac{1}{T_t} \sum_{s=1}^{T_t} \hat{F}_{t,k} \left( X_{s}^{(m)} \right)$
            end
            Reverse $o$ for antithetic sampling;
         end
      end
   end
end

(man_prod), Manufacturing New Orders Index (man_neworders), Manufacturing Employment Index (man_empl), Manufacturing Prices Index (man_prices), Manufacturing Supplier Deliveries Index (man_deliv), and PMI Composite Index (man_pmi). The final source is the University of Michigan Survey of Consumers, from which we use three variables: Index of Consumer Sentiment (soc_ics), Index of Consumer Expectations (soc_ice), and Index of Current Economic Conditions (soc_icc). The variables from the Institute for Supply Management and University of Michigan Survey of Consumers are specified in levels.
Table A.1: FRED-MD Variables

The table lists the 121 variables from FRED-MD that are used as inflation predictors. The variables are transformed following McCracken and Ng (2016).

<table>
<thead>
<tr>
<th>(1) Variable</th>
<th>(2) Abbreviation</th>
<th>(3) Variable</th>
<th>(4) Abbreviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Real personal income</td>
<td>rpi</td>
<td>All Employees, Total Nonfarm</td>
<td>payems</td>
</tr>
<tr>
<td>Real personal income excluding current transfer receipts</td>
<td>w875rxt1</td>
<td>All Employees, Goods-Producing</td>
<td>usgood</td>
</tr>
<tr>
<td>Real consumption</td>
<td>dpcera36006abea</td>
<td>All Employees, Mining</td>
<td>ces1021000001</td>
</tr>
<tr>
<td>Real Manu. and Trade Industries Sales</td>
<td>cmrstapl</td>
<td>All Employees, Construction</td>
<td>uscon</td>
</tr>
<tr>
<td>Advance Retail Sales; Retail Trade</td>
<td>retailix</td>
<td>All Employees, Manufacturing</td>
<td>nansamp</td>
</tr>
<tr>
<td>Industrial Production; Total Index</td>
<td>imdpro</td>
<td>All Employees, Durable Goods</td>
<td>ndnansamp</td>
</tr>
<tr>
<td>Industrial Production; Final Products and Nonindustrial Supplies</td>
<td>ipfpms</td>
<td>All Employees, Nondurable Goods</td>
<td>ndnansamp</td>
</tr>
<tr>
<td>Industrial Production; Final Products</td>
<td>ipfinal</td>
<td>All Employees, Service-Providing</td>
<td>avgserv</td>
</tr>
<tr>
<td>Industrial Production; Consumer Goods</td>
<td>ipcmsgd</td>
<td>All Employees, Trade, Transportation, and Utilities</td>
<td>usutpl</td>
</tr>
<tr>
<td>Industrial Production; Durable Consumer Goods</td>
<td>ipdcmgd</td>
<td>All Employees, Wholesale Trade</td>
<td>usvstrade</td>
</tr>
<tr>
<td>Industrial Production; Non-Durable Consumer Goods</td>
<td>ipncmgd</td>
<td>All Employees, Retail Trade</td>
<td>usstrade</td>
</tr>
<tr>
<td>Industrial Production; Equipment: Business Equipment</td>
<td>ipbusseq</td>
<td>All Employees, Financial Activities</td>
<td>usfire</td>
</tr>
<tr>
<td>Industrial Production; Materials</td>
<td>ipmat</td>
<td>All Employees, Government</td>
<td>usugov</td>
</tr>
<tr>
<td>Industrial Production; Durable Goods Materials</td>
<td>ipdmat</td>
<td>Average Weekly Hours of Production and</td>
<td>ces0600800007</td>
</tr>
<tr>
<td>Industrial Production; Non-Durable Goods Materials</td>
<td>ipmmat</td>
<td>Nonsupervisory Employees, Goods-Producing</td>
<td>avgserv</td>
</tr>
<tr>
<td>Industrial Production; Manufacturing (SIC)</td>
<td>ipmsncs</td>
<td>Average Weekly Hours of Production and</td>
<td>avgserv</td>
</tr>
<tr>
<td>Industrial Production; Non-Durable Consumer Energy Products Residential Utilities</td>
<td>ipb51222a</td>
<td>New Privately Owned Housing Units Started: Total Units</td>
<td>houst</td>
</tr>
<tr>
<td>Industrial Production; Non-Durable Consumer Energy Products Fuels</td>
<td>ipfuels</td>
<td>New Privately Owned Housing Units Started: Total Units in the Northeast Census Region</td>
<td>houstne</td>
</tr>
<tr>
<td>Capacity Utilization; Manufacturing (SIC)</td>
<td>cumfns</td>
<td>New Privately Owned Housing Units Started: Total Units in the Midwest Census Region</td>
<td>houstm</td>
</tr>
<tr>
<td>Non-farm vacancies</td>
<td>hvi</td>
<td>New Privately Owned Housing Units Started: Total Units in the South Census Region</td>
<td>housts</td>
</tr>
<tr>
<td>HWI/(# unemployed)</td>
<td>hviratio</td>
<td>New Privately Owned Housing Units Started: Total Units in the West Census Region</td>
<td>houstw</td>
</tr>
<tr>
<td>Civilian Labor Force Level</td>
<td>clf16ov</td>
<td>New Privately Owned Housing Units Authorized in Permit-Issuing Places: Total Units</td>
<td>permit</td>
</tr>
<tr>
<td>Employment Level</td>
<td>ce16ov</td>
<td>New Privately Owned Housing Units Authorized in Permit-Issuing Places: Total Units in the Northeast Census Region</td>
<td>permitns</td>
</tr>
<tr>
<td>Unemployment Rate</td>
<td>unrate</td>
<td>New Privately Owned Housing Units Authorized in Permit-Issuing Places: Total Units in the Midwest Census Region</td>
<td>permitmw</td>
</tr>
<tr>
<td>Average Weeks Unemployed</td>
<td>uempmean</td>
<td>New Privately Owned Housing Units Authorized in Permit-Issuing Places: Total Units in the West Census Region</td>
<td>permitm</td>
</tr>
<tr>
<td>Number Unemployed for Less Than 5 Weeks</td>
<td>uemp15t</td>
<td>New Privately Owned Housing Units Authorized in Permit-Issuing Places: Total Units in the West Census Region</td>
<td>permitw</td>
</tr>
<tr>
<td>Number Unemployed for 5-14 Weeks</td>
<td>uemp5to14</td>
<td>New Orders for Durable Goods</td>
<td>andnmx</td>
</tr>
<tr>
<td>Number Unemployed for 15 Weeks &amp; Over</td>
<td>uemp15ov</td>
<td>Unfilled Orders for Durable Goods</td>
<td>andnmx</td>
</tr>
<tr>
<td>Number Unemployed for 15–26 Weeks</td>
<td>uemp1526</td>
<td>Total Business Inventories</td>
<td>businvx</td>
</tr>
<tr>
<td>Number Unemployed for 27 Weeks &amp; Over</td>
<td>uemp27ov</td>
<td>Total Business: Inventories to Sales Ratio</td>
<td>isratios</td>
</tr>
<tr>
<td>Initial Claims</td>
<td>claims</td>
<td>MI Money Supply</td>
<td>stest</td>
</tr>
</tbody>
</table>

A.3. Forecasting the Equity Premium

This section uses the time-series metrics developed in Section 2 of the paper to analyze out-of-sample forecasts of the US equity premium. Out-of-sample equity premium prediction is obviously quite challenging, as equity markets are reasonably efficient, so that the predictable component in the market excess return is inherently small; thus, we are forecasting in a
provided that we adequately guard against overfitting.

Nevertheless, there is growing evidence that the equity premium is predictable on an out-of-sample basis using a large number of predictors, provided that we adequately guard against overfitting.

---

1The predictable component can reflect a rational time-varying risk premium and/or inefficiencies related to behavioral factors and market frictions.
Consider a linear predictive regression model:

\[ r_{m,t+1:t+h} = \alpha + x'_t \beta + \varepsilon_{t+1:t+h}, \]  

(A.1)

where \( r_{m,t+1:t+h} = (1/h) \sum_{k=1}^{h} r_{m,t+k}, \) is the market log excess return, \( x_t = [x_{1,t} \ldots x_{P,t}]' \) is a \( P \)-dimensional vector of predictors, \( \alpha \) is the intercept, and \( \beta = [\beta_1 \ldots \beta_P]' \) is a \( P \)-dimensional vector of slope coefficients. Estimating Equation (A.1) via ordinary least squares (OLS) produces the following forecast:

\[ \hat{r}_{m,t+1:t+h}^{\text{OLS}} = \hat{\alpha}^{\text{OLS}} + x'_t \hat{\beta}^{\text{OLS}}, \]  

(A.2)

where \( \hat{\alpha}^{\text{OLS}} \) and \( \hat{\beta}^{\text{OLS}} \) are the OLS estimates of \( \alpha \) and \( \beta \), respectively, in Equation (A.1) based on data through \( t \). While straightforward to compute, the forecast in Equation (A.2) is likely to perform poorly in practice, especially in a low signal-to-noise environment.

**A.3.1. Forecasting Strategies**

We consider three different strategies for forecasting the equity premium based on a large number of predictors. The strategies all essentially rely on shrinkage to alleviate overfitting and have proven useful in the literature. All of the approaches employ an expanding estimation window.\(^2\)

**A.3.1.1. Principal Component Regression**

The first approach is principal component regression (PCR), which Ludvigson and Ng (2007), Neely et al. (2014), Çakmakli and van Dijk (2016), and Dong et al. (2022) find improves out-of-sample equity premium prediction in the presence of a large number of predictors.

\(^2\)Although a rolling window better accommodates potential structural breaks, an expanding window provides more observations to improve estimation precision, which can be beneficial in light of the bias-variance trade-off. This is especially the case when forecasting in an environment with a low signal-to-noise ratio.
Let \( z_t = [z_{1,t}, \ldots, z_{C,t}]' \) denote the vector containing the first \( C \) principal components corresponding to \( x_t \), where \( C \ll P \). The PCR specification is given by

\[
r_{m,t+1:t+h} = \alpha_z + z_t' \beta_z + \varepsilon_{t+1:t+h},
\]

where \( \beta_z = [\beta_{z,1}, \ldots, \beta_{z,C}]' \) is a \( C \)-dimensional vector of slope coefficients. The forecast corresponding to Equation (A.3) is as follows:

\[
\hat{r}_{m,t+1:t+h}^{\text{PCR}} = \hat{\alpha}_z^{\text{OLS}} + z_t' \hat{\beta}_z^{\text{OLS}},
\]

where \( \hat{\alpha}_z^{\text{OLS}} \) and \( \hat{\beta}_z^{\text{OLS}} \) are the OLS estimates of \( \alpha_z \) and \( \beta_z \), respectively, in Equation (A.3), and \( z_t \) is the \( C \)-dimensional vector of the first \( C \) principal components extracted from \( x_t \), all of which are based on data through \( t \). As in Neely et al. (2014), we select \( C \) by maximizing the adjusted \( R^2 \) for Equation (A.3) based on data through \( t \), allowing for a maximum value of four for \( C \).

A.3.1.2. Elastic Net

The second approach relies on the elastic net (ENet, Zou and Hastie 2005) to estimate Equation (A.1). The objective function for ENet estimation of Equation (A.1) can be expressed as

\[
\arg\min_{\alpha, \beta} \frac{1}{2[t - (h - 1) - 1]} \left\{ \sum_{s=1}^{t-(h-1)-1} \left[ r_{m,s+1:s+h}^{(i)} - (\alpha + x_s' \beta) \right]^2 + \lambda P_\delta(\beta) \right\},
\]

where

\[
P_\delta(\beta) = 0.5(1 - \delta)||\beta||^2_2 + \delta||\beta||_1;
\]

\[\text{A.6}\]
\( \lambda \geq 0 \) is a hyperparameter that governs the degree of shrinkage; \( \| \cdot \|_1 \) and \( \| \cdot \|_2 \) are the \( \ell_1 \) and \( \ell_2 \) norms, respectively; and \( 0 \leq \delta \leq 1 \) is a hyperparameter for blending the \( \ell_1 \) and \( \ell_2 \) components in the penalty term. We set \( \delta = 0.5 \) (Hastie and Qian 2016) and tune \( \lambda \) using a walk-forward cross-validation procedure. The ENet forecast based on Equation (A.1) is given by

\[
\hat{r}_{m,t+1:t+h}^{\text{ENet}} = \hat{\alpha}^{\text{ENet}} + x'_t \hat{\beta}^{\text{ENet}}, \tag{A.7}
\]

where \( \hat{\alpha}^{\text{ENet}} \) and \( \hat{\beta}^{\text{ENet}} \) are the ENet estimates of \( \alpha \) and \( \beta \), respectively, in Equation (A.1) based on data through \( t \). Dong et al. (2022) find that the ENet is useful for improving out-of-sample equity premium prediction with a large number of predictors.

### A.3.1.3. Forecast Combination

The third approach uses forecast combination, which Rapach, Strauss, and Zhou (2010) show improves out-of-sample equity premium prediction based on a large number of predictors. We begin with bivariate predictive regressions based on each individual predictor (considered in turn):

\[
r_{m,t+1:t+h}^{(p)} = \alpha^{(p)} + \beta^{(p)} x_{p,t} + \varepsilon_{t+1:t+h} \tag{A.8}
\]

for \( p = 1, \ldots, P \). A forecast based on Equation (A.8) is given by

\[
\hat{r}_{m,t+1:t+h}^{(p)} = \hat{\alpha}^{(p)} + \hat{\beta}^{(p)} x_{p,t} \tag{A.9}
\]

for \( p = 1, \ldots, P \), where \( \hat{\alpha}^{(p)} \) and \( \hat{\beta}^{(p)} \) are the OLS estimates of \( \alpha^{(p)} \) and \( \beta^{(p)} \), respectively, in Equation (A.8) based on data through \( t \). A simple combination forecast takes the form of an equal-weighted average of the individual bivariate predictive regression forecasts in
Equation (A.9):
\[ \hat{r}_{m,t+1:t+h}^{\text{Combine}} = \frac{1}{P} \sum_{p=1}^{P} \hat{r}_{m,t+1:t+h}^{(p)}. \]  
(A.10)

As shown by Rapach, Strauss, and Zhou (2010), relative to the forecast in Equation (A.2) based on OLS estimation of the multiple predictive regression in Equation (A.1), the combination forecast in Equation (A.10) makes two adjustments. First, it replaces the multiple regression slope coefficient estimates with their univariate counterparts, which reduces the role of multicollinearity in producing imprecise parameter estimates. Second, it decreases the magnitudes of the slope coefficient estimates by the factor \(1/P\). These adjustments induce a strong shrinkage effect, which helps guard against overfitting in a noisy data environment.

A.3.2. Data

We measure the equity premium as the log excess return on the S&P 500 portfolio, which we forecast using the set of 28 predictors from Neely et al. (2014). The predictors include 14 macroeconomic and financial variables from Goyal and Welch (2008), as well as 14 technical indicators. The sample period is 1950:12 to 2020:12. We use 1950:12 to 1965:12 as the initial in-sample period and generate out-of-sample forecasts for 1966:01 to 2020:12.

The equity premium predictors are comprised of 14 predictors from Goyal and Welch (2008) and 14 indicator variables based on technical signals, as in Neely et al. (2014). The first 14 predictors are constructed from data available on Amit Goyal’s webpage:

- Log dividend yield (dy): log of a twelve-month moving sum of dividends for the S&P 500 index minus the log of the lagged S&P 500 price index;

\(^{3}\)Data for the S&P 500 return and risk-free return are from Amit Goyal’s webpage.
• Log earnings-price ratio (\(ep\)): log of a twelve-month moving sum of earnings for the S&P 500 index minus the log of the S&P 500 price index;

• Log payout ratio (\(de\)): log of a twelve-month moving sum of dividends for the S&P 500 index minus the log of a twelve-month moving sum of earnings for the S&P 500 index;

• Return volatility (\(rvol\)): volatility of the S&P 500 excess return based on a twelve-month moving standard deviation estimator (Mele 2007);\(^4\)

• Book-to-market ratio (\(bm\)): book-to-market ratio for the Dow Jones Industrial Average;

• Net equity issuance (\(ntis\)): twelve-month moving sum of net equity issues by NYSE-listed stocks dividend by the total end-of-year market capitalization of NYSE stocks;

• Three-month Treasury bill yield (\(tbl\));

• Long-term government bond yield (\(lty\)): yield on a ten-year Treasury bond;

• Long-term government bond return (\(ltr\)): return on a ten-year Treasury bond;

• Term spread (\(tms\)): ten-year Treasury bond yield minus the three-month Treasury bill yield;

• Default yield spread (\(dfy\)): BAA-rated corporate bond yield minus the AAA-rated corporate bond yield;

• Default return spread (\(dfr\)): long-term corporate bond return minus the long-term government bond return;

• Inflation (\(infl\)): CPI inflation rate.\(^5\)


\(^5\)We lag the inflation rate by one month to account for the delay in the release of CPI data.
The technical indicators are based on three popular trend-following strategies. The first is a MA rule that compares a short MA to a long MA and generates a buy (sell) signal of one (zero) if the short MA is greater than or equal to (less than) the long MA:

\[
S_t = \begin{cases} 
1 & \text{if } MA_{s,t} \geq MA_{l,t}, \\
0 & \text{if } MA_{s,t} < MA_{l,t},
\end{cases}
\]  

(A.11)

where

\[
MA_{j,t} = \frac{1}{j} \sum_{i=0}^{j-1} P_{t-i} \quad \text{for } j = s, l;
\]

(A.12)

\(P_t\) is the level of the S&P 500 price index; and \(s (l)\) is the length of the short (long) MA \((s < l)\). The MA indicator with MA lengths \(s\) and \(l\) is denoted by \(MA(s, l)\). We consider six MA indicators: \(MA(1, 9) (ma0109)\), \(MA(1, 12) (ma0112)\), \(MA(2, 9) (ma0209)\), \(MA(2, 12) (ma0212)\), \(MA(3, 9) (ma0309)\), and \(MA(3, 12) (ma0312)\).

The next strategy relies on momentum by comparing the current stock price with its value \(l\) periods ago:

\[
S_t = \begin{cases} 
1 & \text{if } P_t \geq P_{t-l}, \\
0 & \text{if } P_t < P_{t-l}.
\end{cases}
\]

(A.13)

We compute momentum indicators for \(l = 9 (mom09)\) and \(l = 12 (mom12)\).

The final strategy employs on-balance volume (Granville 1963), which is defined as

\[
OBV_t = \sum_{k=1}^{t} VOL_k D_k,
\]

(A.14)

where \(VOL_k\) is a period-\(k\) measure of trading volume and \(D_k\) is a binary variable that takes
a value of 1 if $P_k - P_{k-1} \geq 0$ and $-1$ otherwise.\footnote{Data for trading volume are from Yahoo Finance.} A trading signal is generated by

$$S_t = \begin{cases} 1 & \text{if } MA_{s,t}^{OBV} \geq MA_{t,t}^{OBV}, \\ 0 & \text{if } MA_{s,t}^{OBV} < MA_{t,t}^{OBV}, \end{cases} \quad (A.15)$$

where

$$MA_{j,t}^{OBV} = \frac{1}{j} \sum_{i=0}^{j-1} OBV_{t-i} \quad \text{for } j = s, l. \quad (A.16)$$

Denoting an on-balance volume indicator by $VOL(s, l)$, we consider six indicators: $VOL(1, 9)$ ($vo10109$), $VOL(1, 12)$ ($vo10112$), $VOL(2, 9)$ ($vo10209$), $VOL(2, 12)$ ($vo10212$), $VOL(3, 9)$ ($vo10309$), and $VOL(3, 12)$ ($vo10312$).

### A.3.3. Results

We analyze the PCR, ENet, and combination forecasts of the equity premium via the out-of-sample $R^2 (R_{OS}^2)$ statistic (Fama and French 1989; Campbell and Thompson 2008), a widely used measure of out-of-sample forecasting performance in the literature on equity premium prediction. The $R_{OS}^2$ statistic is defined as

$$R_{OS}^2 = 1 - \frac{\text{MSE}_{\text{Compete}}}{\text{MSE}_{\text{PM}}} = 1 - \frac{\sum_{s=1}^{D-(h-1)} \left( \hat{r}_{m,T_i+s:T_{i+h}+(s-1)} - \hat{r}_{m,T_i+s:T_{i+h}+(s-1)}^{\text{Compete}} \right)^2}{\sum_{s=1}^{D-(h-1)} \left( \hat{r}_{m,T_i+s:T_{i+h}+(s-1)} - \hat{r}_{m,T_i+s:T_{i+h}+(s-1)}^{\text{PM}} \right)^2}, \quad (A.17)$$

where $\text{MSE}_{\text{Compete}}$ is the MSE for a competing forecast generically denoted by $\hat{r}_{m,T_i+s:T_{i+h}}^{\text{Compete}}$, $\text{MSE}_{\text{PM}}$ is the MSE for the prevailing mean benchmark forecast denoted by $\hat{r}_{m,T_i+s:T_{i+h}}^{\text{PM}}$, $D$ is the length of the out-of-sample period, and $h$ is the forecast horizon. The prevailing mean
The forecast is given by
\[
\hat{r}_{m,t+1:t+h} = \frac{1}{t - (h - 1)} \sum_{s=1}^{t-(h-1)} r_{m,s:s+h-1}. \tag{A.18}
\]

Equation (A.18) is a conventional benchmark forecast for assessing out-of-sample market return predictability. It corresponds to the constant expected excess return model, which assumes that the market excess return is not predictable (apart from its mean value). The competing forecast is based on Equation (A.1), so that it incorporates the information in the vector of predictors \(x_t\). The \(R^2_{OS}\) statistic in Equation (A.17), which can be viewed as an “inverted” loss function, measures the proportional reduction in MSE for the competing forecast that utilizes the information in the predictors vis-à-vis the benchmark forecast that ignores the information. Due to the intrinsically small predictable component in the equity premium, the prevailing mean forecast is a stringent benchmark (e.g., Goyal and Welch 2008).

Table A.2 reports \(R^2_{OS}\) statistics for the PCR, ENet, and combination forecasts for horizons of one, three, six, and twelve months. We use the Clark and West (2007) statistic to test the null hypothesis that the MSE (in population) for the prevailing mean forecast is less than or equal to that for the competing forecast \((R^2_{OS} \leq 0)\) against the (one-sided, upper-tail) alternative that the prevailing mean forecast MSE is greater than the competing forecast MSE \((R^2_{OS} > 0)\).

All of the \(R^2_{OS}\) statistics in Table A.2 are positive, so that the competing forecasts always outperform the prevailing mean benchmark in terms of MSE. Because the predictable com-

\[^7\]Clark and McCracken (2001) and McCracken (2007) show that the popular Diebold and Mariano (1995) and West (1996) statistic has a nonstandard asymptotic distribution when comparing forecasts from nested models; in particular, it can be substantially undersized, resulting in low power to detect predictability when it exists. Clark and West (2007) adjust the Diebold and Mariano (1995) and West (1996) statistic so that it is approximately asymptotically distributed as a standard normal random variable when comparing forecasts from nested models. We use a robust standard error (Newey and West 1987) to compute the Clark and West (2007) statistic, which accounts for the autocorrelation induced by overlapping observations when \(h > 1\). In Section 3 of the paper, we use the Diebold and Mariano (1995) and West (1996) statistic to analyze inflation forecasts even though the competing and benchmark forecasts are nested, as they are estimated via rolling windows (Giacomini and White 2006).
Table A.2: Out-of-Sample Forecasting Results for the Equity Premium

The table reports out-of-sample $R^2$ ($R^2_{\text{OS}}$) statistics for equity premium forecasts for 1966:01 to 2020:12 and the forecast horizon ($h$) in the column heading. The forecasts are generated using an expanding estimation window; the initial window spans 1950:12 to 1965:12. The $R^2_{\text{OS}}$ statistic is the percent reduction in mean squared error (MSE) for the competing forecast in the first column vis-à-vis the prevailing mean benchmark forecast. Brackets report the $p$-value for the Clark and West (2007) statistic for testing the null hypothesis that the benchmark forecast MSE is less than or equal to the competing forecast MSE against the (one-sided, upper tail) alternative hypothesis that the benchmark forecast MSE is greater than the competing forecast MSE.

<table>
<thead>
<tr>
<th>Forecast</th>
<th>(1)</th>
<th>(2)</th>
<th>(3)</th>
<th>(4)</th>
<th>(5)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Principal component regression</td>
<td>1.61%</td>
<td>3.29%</td>
<td>4.78%</td>
<td>7.71%</td>
<td></td>
</tr>
<tr>
<td></td>
<td>[0.007]</td>
<td>[0.010]</td>
<td>[0.014]</td>
<td>[0.007]</td>
<td></td>
</tr>
<tr>
<td>Elastic net</td>
<td>1.11%</td>
<td>1.20%</td>
<td>1.56%</td>
<td>5.85%</td>
<td></td>
</tr>
<tr>
<td></td>
<td>[0.000]</td>
<td>[0.012]</td>
<td>[0.014]</td>
<td>[0.003]</td>
<td></td>
</tr>
<tr>
<td>Combination</td>
<td>0.64%</td>
<td>1.14%</td>
<td>1.52%</td>
<td>2.40%</td>
<td></td>
</tr>
<tr>
<td></td>
<td>[0.014]</td>
<td>[0.033]</td>
<td>[0.045]</td>
<td>[0.031]</td>
<td></td>
</tr>
</tbody>
</table>

ponent in the market excess return is intrinsically small, the $R^2_{\text{OS}}$ statistics in Table A.2 are relatively small. However, even an apparently small degree of return predictability can be economically meaningful; for example, based on the historical Sharpe ratio for the US equity premium, Campbell and Thompson (2008) suggest that a monthly $R^2_{\text{OS}}$ statistic (corresponding to $h = 1$) above 0.5% represents an economically significant degree of predictability. The $R^2_{\text{OS}}$ statistics for the PCR, ENet, and combination forecasts in the second column of Table A.2 are all above the Campbell and Thompson (2008) threshold.

For each forecast in Table A.2, the $R^2_{\text{OS}}$ statistics increase monotonically as the horizon increases, in line with the pattern in the literature. The PCR forecast delivers the best performance at each horizon, with $R^2_{\text{OS}}$ statistics ranging from 1.61% ($h = 1$) to 7.71% ($h = 12$). According to the $p$-values for the Clark and West (2007) statistics reported in brackets, all of the $R^2_{\text{OS}}$ statistics are significant at the 1% or 5% level. Overall, the results in Table A.2 indicate that out-of-sample equity premium predictability based on a large number of predictors is statistically and economically significant, provided that we use forecasting
approaches that are designed to guard against overfitting, which is consistent with the general finding in the literature.8

Next, we use the metrics in Section 2 of the paper to analyze the in-sample and out-of-sample relevance of the predictors for the different forecasting approaches. To compute the global PBSV$_p$ for the $R^2_{OS}$ statistic, we change Equation (27) in the paper to

$$\theta_{p,m}^{out}(W, h, R^2_{OS}) = 1 - \frac{1}{\text{MSEP}_M} \frac{1}{|W|} \times \left\{ \sum_{i \in W} r_{T_{in}+i:T_{in}+h+(i-1)} - \frac{1}{|W|} \sum_{s \in W_i} \hat{f}(x_j, T_{in}+(i-1) : j \in \text{Pre}_p(O_m) \cup \{p\}, x_{k,s} : k \in \text{Post}_p(O_m) ; W_i, h) \right\}^2 - \sum_{i \in W} r_{T_{in}+i:T_{in}+h+(i-1)} - \frac{1}{|W|} \sum_{s \in W_i} \hat{f}(x_j, T_{in}+(i-1) : j \in \text{Pre}_p(O_m), x_{k,s} : k \in \text{Post}_p(O_m) \cup \{p\} ; W_i, h) \right\}^2 \right\} \tag{A.19}$$

for $p \in S$, while Equations (28) and (29) in the paper become

$$\hat{\phi}_p^{out}(W, h, R^2_{OS}) = \frac{1}{2M} \sum_{m=1}^{2M} \theta_{p,m}^{out}(W, h, R^2_{OS}) \tag{A.20}$$

and

$$R^2_{OS} = \sum_{p \in S} \hat{\phi}_p^{out}(W, h, R^2_{OS}) + \hat{\phi}_0^{out}(W, h, R^2_{OS}), \tag{A.21}$$

respectively.

For the PCR forecast and each of the 28 predictors, Figure A.1 shows the iShapley-VI$_p$ and oShapley-VI$_p$ metrics, as well as the global PBSV$_p$ in Equation (A.20). The PBSV$_p$ measures in Figure A.1 decompose the $R^2_{OS}$ statistics for the PCR forecasts in Table A.2 into components attributable to the individual predictors. The different panels in Figure A.1 display results for the different horizons. The predictors on the horizontal axis in each panel are ordered according to iShapley-VI$_p$. The red bars and black lines correspond to iShapley-VI$_p$.

---

8As expected, the OLS forecast in Equation (A.2) is substantially less accurate than the prevailing mean benchmark in terms of MSE at all horizons.
Figure A.1. Variable importance and PBSV for PCR equity premium forecasts. The figure shows iShapley-VI (left axis), oShapley-VI (left axis), and PBSV (right axis) for principal component regression (PCR) equity premium forecasts for the 1966:01 to 2020:12 out-of-sample period. The forecasts are generated using an expanding estimation window; the initial window spans 1950:12 to 1965:12. iShapley-VI (oShapley-VI) is the predictor’s importance for all of the in-sample predictions over all of the expanding estimation windows (out-of-sample forecasts); PBSV is the predictor’s contribution to the out-of-sample $R^2$ statistic over the out-of-sample period. The numbers associated with the green bars are rankings of predictors according to PBSV; a positive (negative) ranking indicates predictors that improve (decrease) out-of-sample forecasting accuracy.

and oShapley-VI$_p$, respectively, while the green bars correspond to $\hat{\varphi}^{out}_p(W, h, R^2_{OS})$ in Equation (A.20). The numbers associated with the green bars are rankings for the contributions of the predictors to out-of-sample forecasting accuracy, where predictors with a positive (neg-
ative) ranking contribute positively (negatively) to the $R^2_{OS}$ statistic over the out-of-sample period; for example, a ranking of 1 (−1) signifies the predictor that contributes the most to improving (adversely affecting) out-of-sample forecasting accuracy.

The top panel in Figure A.1 provides results for the one-month horizon. According to iShapley-VI$_p$, the dividend yield ($dy$) and dividend-price ratio ($dp$) are the two most important predictors, followed by the payout ratio ($de$), three-month Treasury bill yield ($tbl$), and term spread ($tms$). Interestingly, the dividend yield, dividend-price ratio, and Treasury bill yield are also arguably the most popular equity premium predictors in the literature. Technical indicators based on on-balance volume ($vo10109$, $vo10209$, $vo10309$, and $vo10112$) also appear among the top ten in terms of in-sample variable importance. Comparing the red bars with the black lines, there is a reasonably close correspondence between in-sample and out-of-sample variable importance according to iShapley-VI$_p$ and oShapley-VI$_p$, respectively. This is not surprising, as the in-sample and out-of-sample predicted target values are based on the same fitted models when determining the importance of individual predictors.

When we compare the out-of-sample PBSV$_p$ to the in-sample iShapley-VI$_p$, sizable discrepancies emerge in the top panel of Figure A.1. Although $dy$ and $dp$ are deemed the two most important predictors by iShapley-VI$_p$, they adversely affect out-of-sample forecasting accuracy in terms of the $R^2_{OS}$ statistic based on PBSV$_p$. The book-to-market ratio ($bm$) is in the top half of predictors in terms of importance according to iShapley-VI$_p$. However, PBSV$_p$ indicates that it worsens out-of-sample accuracy; indeed, according to its ranking, $bm$ contributes most negatively to the $R^2_{OS}$ statistic. There are also cases where a predictor appears to be of limited importance according to iShapley-VI$_p$, while it is quite important according to PBSV$_p$; for example, the long-term government bond return ($ltr$) is ranked in the bottom half of predictors in terms of in-sample variable importance, although it is the predictor that contributes most positively to the $R^2_{OS}$ statistic based on PBSV$_p$.

For some of the other predictors, iShapley-VI$_p$ and PBSV$_p$ are much more in line. For example, $tbl$ and $tms$ are among the top five (three) predictors according to iShapley-VI$_p$.
Intuitively, because the PCR forecast generates statistically and economically significant out-of-sample gains, there should be a reasonable degree of coherence between iShapley-VI\textsubscript{p} and PBSV\textsubscript{p}. Nevertheless, the results in the top panel of Figure A.1 caution that in-sample (and out-of-sample) measures of variable importance are not necessarily reliable indicators of individual predictors’ contributions when it comes to out-of-sample forecasting accuracy, even when a forecast provides significant out-of-sample gains. The results in the remaining panels of Figure A.1 are similar to those in the top panel, reinforcing the importance of PBSV\textsubscript{p} for understanding the relevance of predictors for out-of-sample forecasting accuracy.

Figure A.2 presents results for the ENet forecast. Overall, the patterns are similar to those for the PCR forecast in Figure A.1. The iShapley-VI\textsubscript{p} and oShapley-VI\textsubscript{p} measures in Figure A.2 are again reasonably close. The predictor tbl is the most important according to iShapley-VI\textsubscript{p} at horizons of one, three, and six months. PBSV\textsubscript{p} indicates that tbl is also the predictor that contributes the most or second-most to improving out-of-sample performance at these horizons. Other predictors that are among the most important based on iShapley-VI\textsubscript{p} and are also ranked highly by PBSV\textsubscript{p} in terms of improving out-of-sample accuracy are ltr for \( h = 1 \); dp and return volatility (rvol) for \( h = 3 \); net equity issuance (ntis) and tms for \( h = 6 \); and ntis, dp, tms, and tbl for \( h = 12 \).

As in Figure A.1, there are also notable discrepancies between iShapley-VI\textsubscript{p} and PBSV\textsubscript{p} in Figure A.2. For example, for the one-month horizon in the top panel, ma0212 and ma0309 are the fourth and eleventh most important predictors, respectively, according to iShapley-VI\textsubscript{p}, but they make the second and first most negative contributions, respectively, to the \( R^2_{OS} \) statistic based on PBSV\textsubscript{p}. There are also discrepancies in the opposite direction: ma0109 and ma0312 are only the 13th and 19th most important predictors according to iShapley-VI\textsubscript{p}, while PBSV\textsubscript{p} indicates that they are ranked third and fourth, respectively, in terms of their positive contributions to the \( R^2_{OS} \) statistic. Other predictors evincing discrepancies between iShapley-VI\textsubscript{p} and PBSV\textsubscript{p} in Figure A.2 are the default yield spread (dfy), long-
Figure A.2. Variable importance and PBSV for ENet equity premium forecasts. See the notes for Figure A.1 with elastic net (ENet) replacing the principal component regression forecast.

term government bond yield (lty), de, bm, dy, and the earnings-price ratio (ep) for $h = 3$; lty, de, and bm for $h = 6$; and de and lty for $h = 12$.

Figure A.3 displays results for the combination forecast. Overall, there are fewer discrepancies between iShapley-VI$_p$ and PBSV$_p$ in Figure A.3. At the one- and three-month horizons in the top two panels, tbl and tms are the two most important predictors according to iShapley-VI$_p$, and these are also the two predictors that contribute the most to improving out-of-sample performance based on PBSV$_p$. The predictors tbl and tms are also among
the most relevant predictors according to both iShapley-VI<sub>p</sub> and PBSV<sub>p</sub> at the six- and twelve-month horizons in the bottom two panels. There are still some sizable discrepancies between iShapley-VI<sub>p</sub> and PBSV<sub>p</sub> in Figure A.3. For example, ep and bm are both among the ten most important predictors according to iShapley-VI<sub>p</sub> at the six-month horizon; however, they both worsen out-of-sample performance based on PBSV<sub>p</sub>, and bm contributes the most adversely to out-of-sample accuracy.

When comparing the PBSV<sub>p</sub> measures in Figures A.1 to A.3, it is worth noting that
whether a predictor contributes positively or negatively to out-of-sample forecasting accuracy depends on the fitted model. Among other things, this is due to the amount of shrinkage induced in the slope coefficient attached to a predictor and the predictor’s correlations with other predictors. As an example of how the marginal contribution of a predictor to out-of-sample performance can differ across fitted models, $dp$ contributes negatively to the $R^2_{OS}$ statistic at all horizons for the PCR forecasts in Figure A.1, while it contributes positively at all horizons for the ENet and combination forecasts in Figures A.2 and A.3, respectively.

Figure A.4 plots the cumulative difference in squared errors (CDSE, Goyal and Welch 2008) between the prevailing mean benchmark and PCR forecasts. We also compute PBSV$_p$. 

100

0

100

200

300

1967/68

1971/72

1975/76

1979/80

1983/84

1987/88

1991/92

1995/96

1999/00

2003/04

2007/08

2011/12

2015/16

2019/20

Figure A.4. Cumulative difference in squared errors for PCR equity premium forecasts. The figure shows the cumulative difference in squared errors for the prevailing mean benchmark forecast vis-à-vis the principal component regression (PCR) forecast for the 1966:01 to 2020:12 out-of-sample period. Shifts to the right (left) imply an improvement (deterioration) in forecasting performance relative to the prevailing mean benchmark. The figure also shows the two top (bottom) contributors to the improvement (deterioration) in forecasting performance for all non-overlapping 24-month subsamples in the out-of-sample period; a green (red) color for the predictor indicates that the 24-month subsample is associated with an improvement (deterioration) in performance. Horizontal gray bars indicate 24-month subsamples that contain an NBER-dated recession.
measures for the PCR forecast for non-overlapping 24-month rolling subsamples. Figure A.4 shows the two predictors that contribute the most to outperformance during a given subsample, as well as the two that contribute the most to underperformance. Predictors in green (red) to the right (left) of the curve indicate that the PCR forecast delivers a lower (higher) MSE than the prevailing mean benchmark for the subsample.

Starting with the one-month horizon in the left panel of Figure A.4, a well-known result from the literature is evident: out-of-sample equity premium predictability tends to be concentrated around business-cycle recessions (e.g., Rapach, Strauss, and Zhou 2010; Henkel, Martin, and Nardari 2011; Dangl and Halling 2012). This is especially the case for the relatively severe and long-lived recessions of the mid 1970s, early 1980s, and late 2000s, as the CDSE curve tends to move substantially to the right during those episodes. Recall from the top panel of Figure A.1 that $dy$ and $dp$ are the two most important predictors according to the in-sample iShapley-VI, but that they adversely affect overall out-of-sample accuracy based on PBSV. The CDSE curve in the left panel of Figure A.4 during the mid-to-late 1990s helps to explain the deleterious contributions of $dy$ and $dp$ to out-of-sample performance. During that period, the CDSE curve moves substantively to the left, indicating that the PCR forecast considerably underperforms the prevailing mean benchmark; furthermore, the predictors in red show that $dy$ and $dp$ are primarily responsible for the underperformance of the PCR forecast during the mid-to-late 1990s according to PBSV. The detrimental influence of $dy$ and $dp$ on out-of-sample performance during the mid-to-late 1990s is consistent with findings in Lettau and Van Nieuwerburgh (2008).

For the six-month horizon in the middle panel of Figure A.4, $tbl$ and $de$ appear in green on multiple occasions, helping to explain their sizable positive contributions to out-of-sample forecasting accuracy in the third panel of Figure A.1. We again see in the middle panel of Figure A.4 that the underperformance of the PCR forecast during the mid-to-late 1990s is chiefly due to $dy$ and $dp$, in line with their adverse contributions to out-of-sample performance in the third panel of Figure A.1. The same pattern is evident during the mid-
to-late 1990s for the twelve-month horizon in the right panel of Figure A.4. Observe that "de" appears in green at multiple places in the right panel, which helps explain its sizable contribution to improving out-of-sample performance in the bottom panel of Figure A.1.

### A.3.4. Simulations

Finally, we use simulations to investigate potential reasons for the discrepancies in rankings between iShaply-VI$_p$ and PBSV$_p$ in Section A.3.3. The first potential source of differences in rankings that we consider is overfitting, which is clearly relevant for equity premium prediction, since stock returns have an inherently small predictable component and thus a small signal-to-noise ratio. We employ the following data-generating process (DGP):

$$r_{m,t+1} = \alpha + x_t^c' \beta_c + x_t^d' \beta_d + \varepsilon_{t+1}, \quad (A.22)$$

where $\beta_c$ and $\beta_d$ are $P_c$- and $P_d$-dimensional vectors of slope coefficients, respectively; $x_t^c$ and $x_t^d$ are $P_c$- and $P_d$-dimensional vectors of continuous and binary predictors, respectively; and $x_t^c \sim N(0, \Sigma_c)$. Matching the empirical application in Section A.3.3, we set $P_c = P_d = 14$, so that the total number of predictors is $P = P_c + P_d = 28$. To replicate the correlation structure of the data, we set $\Sigma_c$ to the sample covariance matrix for the continuous predictors in the application. To ensure that $x_t^d$ has the same mean and covariance structure as the binary variables in the application, we set

$$x_{i,t}^d = 1(u_{i,t} > 0) \quad \text{for } i = 1, \ldots, P_d, \quad (A.23)$$

where $u_t \sim N(\mu_d, \Sigma_d)$, $u_t = [u_{1,t} \ldots u_{P_d,t}]'$, and $\mu_d$ and $\Sigma_d$ are set using the procedure in Leisch, Weingessel, and Hornik (1998). For each iteration, we generate the vector $\beta = [\beta_c' \beta_d']$ by drawing each slope coefficient independently from the uniform distribution $U(-P/2, P/2)$. Regarding the disturbance term in Equation (A.22), we assume that $\varepsilon_t \sim U(-P/2, P/2)$. \[22\]
where STN is the signal-to-noise ratio.\footnote{We center and scale the right-hand-side of Equation (A.22) so that the generated data for $r_{m,t}$ have the same mean and variance as the actual equity premium over the out-of-sample period. To introduce some sparsity, we set the five slope coefficients with the smallest magnitudes to zero.} The sizes of the in-sample and out-of-sample periods are set to those for the application. For each iteration of simulated time-series data (and using 10,000 iterations), we compute PCR, ENet, and combination forecasts using an expanding estimation window, as well as predictor rankings based on iShapley-VI and $\hat{\phi}_{out}(W, h, R_{OS}^2)$.

The top panel of Figure A.5 shows how the standardized mean squared deviation in rankings (MSDR) varies with the signal strength as measured by the $R^2$ statistic for the PCR, ENet, and combination forecasts. The figure includes horizontal dotted lines corresponding to standardized MSDRs based on the actual data for $h = 1$ in Figures A.1 to A.3. For each forecast, as expected, the MSDR decreases nearly monotonically as the signal strength increases and risk of overfitting is alleviated. In line with the results in Figures A.1 to A.3, the MSDRs are similar for the PCR and ENet forecasts, while the MSDR is uniformly lower for the combination forecast. The horizontal dotted lines intersect the MSDR curves at approximately 2.5% for the ENet and combination forecasts and 10% for the PCR forecast; in other words, $R^2$ statistics of 2.5%, 2.5%, and 10% for the DGP in Equation (A.22) for the ENet, combination, and PCR forecasts, respectively, are consistent with the differences in predictor rankings in the data.

We next consider structural breaks in slope coefficients as a source of differences in predictor rankings. We introduce structural breaks using the same DGP, except that we fix the $R^2$ statistic at 5% and generate the slope coefficients as follows. For each iteration, we draw three vectors of slope coefficients via the uniform distribution. Each period, with some probability, the DGP can change from the current vector of slope coefficients to one of the
Figure A.5. Standardized MSDR for DGPs based on the empirical application. The figure shows the standardized mean squared deviation in rankings (MSDR) for data-generating processes (DGPs) based on the equity premium application.

The middle panel of Figure A.5 depicts the relation between the standardized MSDR and probability of a structural break. The MSDR curves are typically positively sloped, so when a change occurs, the DGP moves to one of the other two vectors of slope coefficients with equal probability.

\textsuperscript{10}When a change occurs, the DGP moves to one of the other two vectors of slope coefficients with equal probability.
that differences in predictor rankings increase as structural breaks in the slope coefficients occur more frequently, indicating that structural breaks provide a plausible explanation for discrepancies between predictor rankings based on iShapley-VI\textsubscript{p} and PBSV\textsubscript{p}. According to the intersections of the horizontal dotted lines and MSDR curves, structural break probabilities of around 0.1% and 0.2% are consistent with the differences in rankings based on the actual data for the combination and ENet forecasts, respectively. For the PCR forecast, the MSDR curve lies uniformly above the relevant horizontal dotted line, suggesting that we would expect even larger differences in predictor rankings than we see in the data for a structural break probability as small as 0.01%.

We consider evolving predictor volatilities as a final potential reason for differences in predictor rankings. In terms of the DGP, we generate the slope coefficients via the uniform distribution, hold the slope coefficients constant over time, and again set the $R^2$ statistic to 5%. To incorporate evolving volatilities, each period, with some probability, the magnitude of one of the continuous predictors increases by a factor of ten. The predictor experiencing the increase in its magnitude is selected randomly, and the increase in magnitude lasts for twelve periods.

The effects of evolving predictor volatilities can be seen in the bottom panel of Figure A.5, which shows how the standardized MSDR varies with the probability of an increase in the magnitude of a predictor. As anticipated, all of the MSDR curves are positively sloped (and nearly monotonically so), meaning that more frequent changes in the volatilities of predictors lead to greater discrepancies between the predictor rankings. For the combination forecast, a magnitude change probability of approximately 1% appears in line with the differences in predictor rankings in the data. For the PCR (ENet) forecast, the MSDR curve lies uniformly above (below) the relevant horizontal dotted line, so that differences in rankings in the data are less (greater) than those that we expect based on the DGP for probabilities ranging from 0.1% to 10%.
References


