Parisa Arbab, Ankita Mishra, Abe Berkley-Vigil

The problem you are trying to solve

We are trying to forecast 1 period ahead of stock returns for AAPL, IBM, KO, WMT. We are using commonly used economic indicators, financial metrics, and our own generated technical analysis metrics. These indicators are meant to cover a broad range of themes to provide a diverse set of factors to explain monthly stock returns.

The data for the economic factors came from the St. Louis Fed database. These factors are meant to cover broad swaths of the US and global economy that could potentially influence monthly stock returns. The factors used were CPI, Housing Prices, WTI Crude Price, $/Yen, Fed Funds Rate, Fed Government Expenditures, GDP, Global Corn Prices, Industrial Production, Gold, Leading Economic Indicator, M1 Money Supply, M1 Money Velocity, Inflation Expectations, 30 Year Mortgage rates, PPI All Commodities, Michigan Inflation Expectation, PPI Soybeans, Real Gross Private Domestic Investment, 10Y2Y yield curve, Unemployment Rate, VIX, and The Coincident Economic Indicator. These are the input variables for which we are using to build models to predict our target variable, next period returns. The variables mostly occurred on a monthly basis to match our target variable's timeframe, but some factors were on a quarterly basis. For these factors we forward filled the quarterly data for the missing months.

Backfilling was required for many of these time series in order to have enough data to meet the requirements of this project. To have around 500 data points, our data needed to go back to ~ 1983-1984. We decided the Coincident Economic Indicator was a good factor to use for backfilling. It is often used as a proxy for current economic conditions. Thus, we made an assumption of a relationship between our various factors and the current economic conditions. This is not a perfect assumption as some variables likely lead or lag current economic conditions, but this would have required a level of refinement and work outside the scope of this project. Additionally, the Coincident Economic Indicator had a long enough history to be used for backfilling. Thus, for a time series that had an insufficient number of data points, we regressed the factor's time series, as our y, against the Coincident Economic Indicator, as our X, and used the Coincident Economic Indicators values in a regression for the periods that factor did not have data. Again, we acknowledge this is an overly broad assumption given that there might not be a linear relationship between the factor and Coincident Economic Indicator. Additionally, the relationship might be very weak and thus not an appropriate proxy. This analysis again was beyond the scope

of this project, as we were mostly looking for a means to achieve a sufficient amount of data.

We also set out to determine the optimal lag for each economic factor relative to the return series we were looking at. We thus built a function to iterate over different lead periods and find which time period for a given factor most correlates with the target variable's returns. We applied these to all our economic factors and shifted each column by its highest correlated lead time. This was done for all four of our stocks, thus, the leads even for the same factor may have been different from stock to stock.

The next set of factors we looked at were financial factors. These were common financial statement-based factors that were pulled from Bloomberg. The same factors were pulled for all four stocks. The factors used were P/E Ratio, Enterprise Value/EBITDA, Price to Sales Ratio, and Price to Free Cash Flow. These metrics were selected as being commonly used metrics that offered reasonably long time series and ease of access on a Bloomberg terminal. We backfilled most time series if they had a certain threshold amount of data, and if it didn't, we dropped that factor. The backfilling was achieved by making a regression with the financial metric against the stock's returns. We deemed this to be proper because ultimately our target variables are the next period's returns not the current period's returns. Again, we did consider the fact that the relationship may not be linear or particularly sufficient as a proxy but finding something better was out of scope.

Finally, we created our own technical analysis of factors for prediction. Our first factor looked at a 24-month lookback period, calculated a rolling mean, and then took the difference between the current price and the 24-month rolling moving average. The next factors looked at current returns, and the derivative of current returns. This was done to capture acceleration or decelerations in price changes. The next factor was a 24-month rolling volatility of a stock. Another factor was the difference in a short term, 6 month moving average, and a long term 24 month moving average. The final factor was taking a 24-month rolling average and taking the difference in the number of days above and below the moving average. Some of these factors are common factors while others were created for this project but seemed to reflect potential information regarding momentum or price reversals. In a more in-depth project, we could have back tested the efficacy of our different lookback windows, but determined that would be out of scope for this project.

Pre-Exploratory Analysis

We set up our data by aggregating common factors for return prediction. For the economic data we iterated over different, commonly used lead time periods and correlated them to

our target returns time series. We did this to determine the optimal lead time periods for predicting returns in a given period. We also plotted our data to see the distribution of our data.

For the economic data we plotted the time series to see if they were stationary or not. We chose to ensure the data was stationary as this is a common requirement for making models more stable and robust for predictions.

We transformed the data to make it more stationary if it didn't appear to be so. After the transformations were applied, we replotted the data to see if there was roughly visual stationarity. The transformations applied were either year over year change or percent change. Economic factors that appeared stationary we kept as they were. In terms of improvements we could have made, we were to do an ADF test to verify if the plots were truly stationary but decided against that for this project.

We did not apply transformations to the financial data. It seemed more stationary than the economic data as a lot of the metrics are some forms of a ratio, and not level statistics. Had we had more time we could have explored transforming these factors more to ensure stationarity.

We scaled our data for the models we fed them into. This was because many of our factors were on vastly different orders of magnitudes. By not scaling the data, differences of magnitude could have led certain factors to "explain" disproportionate amounts of variation or have undue influence on the final model results. We used a min-max scaler because we did not want to force normality on some of our factors when visually this appeared to sometimes not to be the case, nor did we test for normality. Additionally, a min-max scaler was used because it is bounded from 0-1, and for a while we did entertain certain transformations, like log transformations, that require input values to be $x > 0$, and a standard normal distribution would have allowed our valued to go below 0. For all the datasets we made our data exist on a common timeframe and removed nans. We also backfilled some of our data to give us sufficiently long time series.

Given that we were building regressions, we plotted the factors against each stock's returns as a scatterplot. We also included the correlation. The factors ranged between 0-0.2 correlation.

Most factors had a correlation less than 0.1. This seemed a little low to us, because most variables did not seem particularly explanatory for each stock's returns. Additionally, many

variables did not appear to have a linear relationship with returns. For AAPL, we plotted the distributions of the variables themselves. We applied transformations to make the data look more normal, like log transforms and power transforms. Then for AAPL, we applied these transformations to some of the most skewed factors and tried replotting the transformed variables against the returns. The relationship still did not appear much more linear and still had a low correlation. Thus, we did not feel it was worth the time to further these explorations for the other stocks, but this is an area we could have investigated further in a longer project.

Models and Results

Given all the data processing, we decided to use 4 models that we believed might achieve effective results at predicting stock returns. We used a K-Nearest Neighbors regression based on a 10-factor exhaustive search of the 10 most correlated variables to the stock's returns. The next model we used was a PCA based regression using the top 18-20 principal components as the factors based on a criteria requiring at least 95% variance captured, and principal components built from all our economic, financial, and technical data. Our third model was a lasso-based regression. Our final model was a random forest. We used an ensemble model to aggregate KNN, Lasso, and Random Forest.

The models were all built with a training set, a testing set, and a validation set. We did this because our model being stock data was sequential thus preventing us from doing cross validation due to lookahead bias. Thus, we thought a more robust check on the data would be to have essentially two sets of data to test the model against. We chose our models by finding the models with the minimum difference between the training and testing set. We did this because a wide difference could mean we are overfitting. Too small a difference could theoretically mean we are under fitting our model, but for the sake of this project, we ignored that because that is much harder to calibrate and account for.

In our analysis, we applied Principal Component Analysis (PCA) to each stock dataset (AAPL, IBM, WMT, KO) to reduce dimensionality while retaining 95% of the variance. By doing so, we identified the optimal number of factors that capture the majority of the data's information. For AAPL, 19 factors explained 95.73% of the variance; for IBM, 20 factors explained 95.56%; for WMT, 18 factors explained 95.05%; and for KO, 19 factors explained 95.50%. These PCA components were then utilized to evaluate the performance of our regression models, yielding impressive validation Mean Squared Errors (MSE) for each stock: AAPL had a validation MSE of 0.00899, IBM achieved 0.00679, WMT resulted in 0.01031, and KO exhibited a RMSE of 0.00228.

For our 10 Factor Exhaustive Search KNN (10-FES-KNN), we iterated over all possible 2^10 models and iterated over K from 2-25. We chose 10 factors because it seemed reasonably diverse while being small enough to still enable us to perform an exhaustive search. Originally, we used both stepwise forward and backward regression for all our factors to determine the optimal ones, but these produced either 1 factor models or all factor models. We wanted to see if the limitations in the algorithm were preventing us from getting broader models and thus decided to try and a filtered by correlation exhaustive search.

For the 10-FES-KNN model for AAPL we use Federal Government Current Expenditures as our X and K = 6. The MSE was .0843. For the IBM model K = 6 and the factor used was $/Yen Exchange Rate. The MSE was .0701. For Coke (KO) K = 2 and the factor used was GDP with an MSE of .06200. For Walmart (WMT) the K = 2 and the factor used was Fed Funds Rate with an MSE of .0674. It seemed odd to us that every stock produced such small models and some with really low K's, but I think this was a function of many of these factors being poorly correlated to the return variables.

The next model was a random forest. To avoid using all of our factors we used the top 15 most correlated factors to each stock's return. We chose 15 instead of 10 because for the KNN regression we were trying to perform an exhaustive search, but the random forest we decided to just feed the factors into the model with the likelihood of the different trees picking up more of the factors more likely. For this model we iterated over the minimum samples per leaf, max depth, and number of estimators. For min sample per leaf, we look from 1-25, max depth 1-8, and number of estimators 5-100 by increments of 5.

The results for the random forests are as follows.

IBM: The most important feature was the VIX.

Factors	**Best min samples/ leaf**	**max depth**	**No. of estimators**	**MSE**
Mortage, First Derivative of Price, Global Corn Price, Inflation Expectatiions, Federal Government Current Expenditures, Real Gross Private Domestic Investment, PPI Soybeans, Fed Funds Rate, Rolling Volatility, M2, Industrial Production, WTI, Rolling Difference, VIX	16	3	5	0.0700

Coke: The most important feature was $/Yen.

Factors	Best min samples/ leaf	max depth	no. of estimators	MSE
WTI, Unemployment Rate, Leading Economic Indicators, $/Yen, Industrial Production, 10 Year Premium, Global Corn Price, GDP, VIX, M1 Velocity, 30 Year Mortgage Rate, PPI All Commodities, CPI, Fed Funds, Inflation Expectation	8	5	10	0.0470

Walmart: The most important feature was $/Yen.

Factors	Best min samples/ leaf	max depth	No. of estimators	MSE
Rolling Difference, PPI Soybeans, Leading Economic Indicators, Price to Sales Ratio, Global Corn Price, VIX, P/E Ratio, Coincident Ratio, Industrial Production, Rolling Volatility, M2, EV/EBITDA, M1 Velocity, GDP, and $/Yen.	8	5	100	0. 0561

AAPL: The most important feature was $/Yen.

Factors	Best min samples/ leaf	max depth	No. of estimators	MSE
Second Derivative, PPI All Commodities, VIX, EV/EBITDA, Global Corn Price, P/E Ratio, WTI, M1 Velocity, Federal Government Expenditures, CPI, M1, PPI Soybeans, Unemployment Rate, $/Yen	4	7	35	0. 0767

For the Lasso Models we iterated over the alpha parameter. This parameter drives the penalty for limiting or allowing less or more factors. A larger value is more restrictive and vice versa. Given that many of these factors were equally bad at predicting our target variable, the best alphas tended to be very small. Thus, many factors were allowed into our models.

	Associated factors and coefficients
AAPL Optimal alpha: 0.0002	AAPL US Equity - Price Earnings Ratio (P/E) 0.002776, AAPL US Equity - Enterprise Value/EBITDA 0.007266, AAPL US Equity - Price to Sales Ratio - 0.001738, Housing Prices 0.028209, WTI -0.001213, $/Yen 0.032244, FEDFUNDS -0.123049, Fed_Gov_Curr_Expenditures 0.029916, GDP 0.052946,Global_Corn_Price 0.004286, M1 Velocity 0.096623, M2SL 0.087226,

	MICH_Inflation_Expectation 0.051046, MORTGAGE30US -0.039337, PPI_Soybeans 0.001555, Real_Gross_Private_Dom_Inv -0.021236, 10 Y Premium 0.010268, VIX 0.084076, Difference -0.104808, Second_Derivative 0.060650, Rolling_Volatility -0.054264, Rolling_Difference 0.016850
KO Optimal alpha: 0.0001	KO Equity (USD) - Price Earnings Ratio (P/E) -0.210827, KO Equity (USD) - Enterprise Value/EBITDA 0.143881, KO Equity (USD) - Price to Sales Ratio 0.119036, CPI 0.033988, Housing Prices -0.000414, $/Yen -0.040988, Fed_Gov_Curr_Expenditures -0.119957, Gold Prices 0.032882, Industrial Production -0.110559, Leading_Economic_Indicators -0.066573, M1SL 0.000224, M2SL 0.061296, MICH_Inflation_Expectation 0.009534, PPI_All_Commodities 0.042446, PPI_Soybeans 0.018997, Real_Gross_Private_Dom_Inv 0.059300 22 10 Y Premium 0.010979, VIX -0.013294, Coincident_Economic_Indicator -0.063467, Difference -0.122123, First_Derivative 0.064196, Second_Derivative -0.022841, Rolling_Volatility -0.028087, Rolling_Difference 0.042631
WMT Optimal alpha: 0.0001	WMT US Equity - Price Earnings Ratio (P/E) 0.001230, WMT US Equity - Enterprise Value/EBITDA 0.089064, WMT US Equity - Price to Sales Ratio 0.010009, CPI -0.013499, Housing Prices 0.018555, WTI 0.003249, $/Yen -0.004553, FEDFUNDS -0.056985, Fed_Gov_Curr_Expenditures -0.022341, Global_Corn_Price 0.011855, Gold Prices 0.010132, Industrial Production -0.002966, Leading_Economic_Indicators 0.055213, M1SL -0.031893, M2SL 0.040143,MICH_Inflation_Expectation 0.040737, MORTGAGE30US -0.018880, PPI_Soybeans -0.020862, Real_Gross_Private_Dom_Inv -0.058488, 10 Y Premium 0.023704, VIX 0.000833, Moving Average Diff 0.001189, First_Derivative 0.107727, Rolling_Volatility -0.101655, Rolling_Difference 0.036127
IBM Optimal alpha: 0.0001	IBM US Equity - Price Earnings Ratio (P/E) -0.009531, IBM US Equity - Enterprise Value/EBITDA 0.006550, IBM US Equity - Price to Sales Ratio -0.038961, IBM US Equity - Price to Free Cash Flow 0.084049, CPI -0.030362, Housing Prices 0.053939, WTI 0.051727, $/Yen -0.027967, FEDFUNDS -0.082652, Fed_Gov_Curr_Expenditures 0.049328, Global_Corn_Price 0.018822, Gold Prices -0.044627, Industrial Production -0.172725, Leading_Economic_Indicators -0.105648, M1V 0.105341, M2SL 0.015392, MICH_Inflation_Expectation -0.005257, MORTGAGE30US 0.055908, PPI_Soybeans 0.050556, Real_Gross_Private_Dom_Inv 0.113702 23 10 Y Premium 0.096216, VIX 0.156163, Coincident_Economic_Indicator -0.062122, Moving Average Diff 0.090852, Difference -0.087413, First_Derivative -0.035075, Rolling_Volatility -0.118573, Rolling_Difference 0.004835

Ensemble Model

For our ensemble model we fed in our Lasso regression, Random Forest, and KNN model. We utilized a weighted voting regressor that combines the predictions from each of these models. To find the optimal weights for combining these models, we employed a

RandomSearchCV. The search space consisted of various combinations of weights, ensuring the weights would be normalized so they sum to 1.0. After the random search identified the best weights, the final ensemble model was trained using these weights to predict the validation set returns. The results below summarize the ensemble model for each stock.

	Validation RMSE	**Best Weights**
AAPL	0.120254	[1.0, 0.0, 0.0]
KO	0.120240	[1.0, 0.0, 0.0]
WMT	0.119956	[0.714, 0.286, 0.0]
IBM	0.120533	[1.0, 0.0, 0.0]

For AAPL, KO, and IBM, the ensemble model heavily favored the Lasso Regression, assigning it a weight of 1.0 and zero weights to the other models. This indicates that Lasso Regression alone provides the best predictive performance for these stocks. For WMT, the ensemble model assigned significant weights to both Lasso (0.714) and Random Forest (0.286), indicating a combination of these models yielded the best predictive performance, while KNN was not utilized. The results quantified by the RMSE values suggest that the ensemble approach tailored the best performing combination of models for each stock, though in most cases, Lasso dominated the predictions.

And of course any ideas on how you could have done better if you did not get the results you expected.

1. Your data files (Zipped if they are large).
2. Your final IPython notebook. Give me one and combine if you have multiple notebooks.
3. To get an A, we required a minimum of 4 models plus an ensemble model using your best of the 4 models. (It might be 2,3, or 4) fed into either a voting, weighted voting, or stacking model.