DECEMBER 20, 2020

IAN YU

FORECAST THE STOCK MARKET UP TO A MONTH WITH LSTM RECURRENT NEURAL NETWORK

STOCK MARKET FORECAST

OBJECTIVE

The objective of this project is to design a predictive model that captures the overall macro environment and forecasts the Standard and Poor's 500 Index with relatively good precision and, more importantly, capturing the **rate of change in a longer timeframe**. The model will be part of the Long-Term Asset Allocator that allocates portfolio resources between major markets, which requires the long-term forecast of rate of change to build. We will be creating a Weekly Predictor, Semi-Monthly Predictor, and Monthly Predictor to explore how far in timeframe can we forecast.

CONTEXT

A large number of stock market predictive models on the web are *confined within the equity world*, turning a blind eye to the interaction between major markets, namely stock, bond, commodity, and currency. As a result, they predict only up to 5-trading days or a binary classification of price up or down in 30 days. These models are inherently risky and often not actionable. Our model instead would be using data from the four major markets and forecasts up to 20 trading days, based on Intermarket Analysis methodology, and will be replicated to different major markets to build an Asset Allocator.

Asset allocation is based on the concept of Modern Portfolio Theory, which seeks Efficient Frontier. By forecasting long-term rate of change, one could capture Expected Return and Expected Volatility of different major markets, which is further complicated by covariance between the markets, one would be able to find optimal weightings to long-term allocation of resources to different markets that minimizes risk while maximizing return.

DATA, FEATURE ENGINEERING, AND PROBLEM TRANSFORMATION

For validation purposes, we sourced open data only. The scope of the project is confined within the US only, so our dataset is comprised of market data proxy to major assets in the US. We have S&P 500 as the target, 10 Year US Government Bond Yields to proxy the bond market, Dollar Index to the currency market, WTI Spot Price and Gold to the commodity market, and Annual change to Consumer Price Index to represent inflation environment. We sourced our data from Yahoo! Finance, DataHub, Federal Reserve Economic Data of St. Louis Fed, and Desjardins.

Through exploratory data analysis and data cleaning, we dropped duplicated features, realigned date range mismatch, expanded weekly and monthly values of the bond market

and commodity market to daily values, interpolated values for statutory holidays on weekdays, and combined the data into one single dataset.

SPX Open	SPX High	SPX Low	SPX Close	SPX Volume	US10Y Rate	DXY Open	DXY High	DXY Low	DXY Close	WTI Price	GOLD Price	CPI Annual Rate
237.300003	240.110001	237.300003	238.970001	178100000	7.78	117.220001	117.849998	117.220001	117.680000	12.00	346.095	1.898048
252.699997	252.940002	251.229996	251.789993	108300000	7.30	111.349998	111.500000	111.010002	111.080002	12.23	348.554	1.898048

We further engineered indicators that capture cross-market relative performance, momentum, and volatility of different timeframes. We created 3 datasets for 3 different predictors to train on. We then transformed the time series analysis into a structured learning problem, setting the S&P 500 value of the day as target, and lagging other market values and indicators for each dataset relative to its predictor. In the end, we have 3 datasets of 161 features and 1 target, with a date range of 1986-06-19 to 2020-07-31.

PREPROCESSING, ARCHITECTURE, HYPERTUNING, AND TRAINING

For every dataset, we split the data from the last 250 trading days, representing one year without training. That is, our predictors are not trained beyond 2019-08-18. We normalized our features with MinMaxScaler, a common scaler for time series problems. Through trial and error, we designed a single hidden layer of Bidirectional Long Short-Term Memory Recurrent Neural Network with Time Distributed Layer as the output layer. Recurrent Neural Network is a type of neural network that allows us to learn how past days affect the price today. Long Short-Term Memory is a special type of RNN that also learns default behaviour overtime, which decomposes seasonality, trends, and other patterns. Bidirectional element makes the layer learn both backwards and forward, accounting for the future context, as the anticipation of the future also affects the price today. Time Distributed Layer ensures timesteps are kept during training.

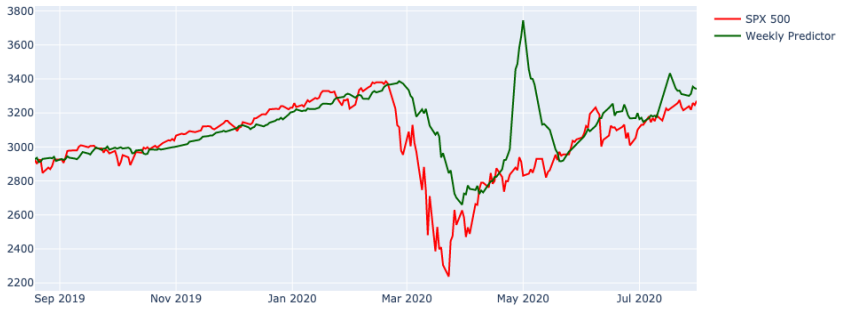
For each of the predictor, we performed hypertuning with Keras Tuner to find the optimal parameters. As our models are computationally expensive, we uploaded our 3 engineered datasets to AWS S3 personal bucket and leveraged AWS EC2 to performed pat of the hyper tuning process. During hypertuning, we performed a 25-split Time Series Split to create 25 validation sets, each with 322 trading days as it is unrealistic to have models not trained for years, and compare the best trial results across validation set to validate the configuration returned by Keras Tuner. After obtaining the configuration for each predictor, we trained the predictors again on the local machine without validation set, applied Early Stopping to prevent overfitting, and saved the model with the lowest loss.

FINAL RESULT

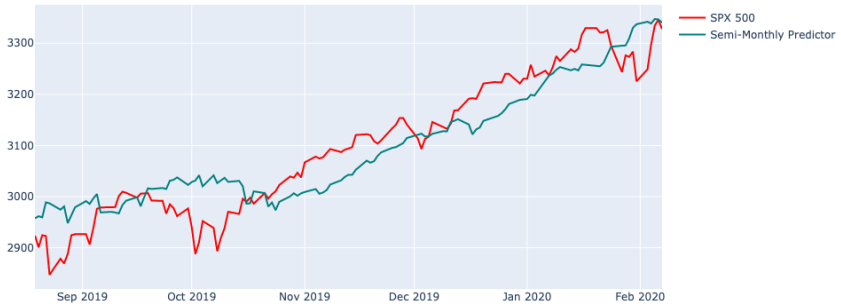
As we tested Weekly, Semi-Monthly, and Monthly against the market after 2019-08-18, our models performed well during regular times, but became more sensitive since the pandemic market crash March 2020. COVID-19 Market Crash is considered the worst market crash since 1929, and the post-crash environment is continues to be volatile. Even our Weekly Predictor, the shortest timeframe, became more sensitive post-crash. Part of the reason is because our predictors did not learn about the recent volatile market environment. But our model was also trained on relatively fewer volatility-related features.

During regular times, however, **all our predictors were able to capture the rate of change of the market** up to February 2020. For Weekly and Semi-Monthly Predictors, the prediction were also relatively precise. Our predictors have returned the element to start building a long-term asset allocator.

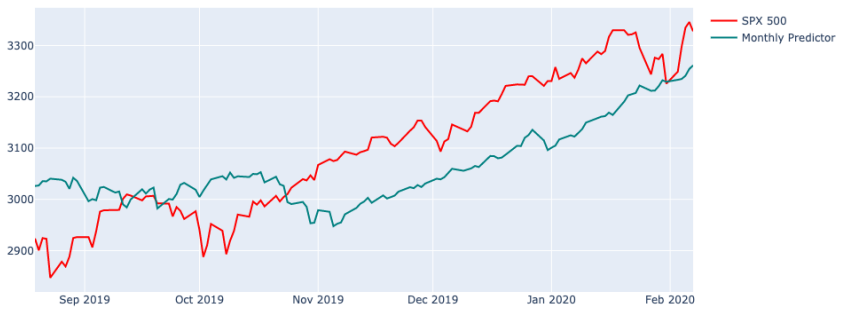
Weekly Prediction, With Pandemic, RMSE:198.75



Semi-Monthly Prediction, Regular Times, RMSE:55.60



Monthly Prediction, Regular Times, RMSE:97.74



FUTURE DIRECTIONS

We will first 1) create pipeline to ensure our current models are trained up to date to learn the most recent environment. We will then 2) replicate the concept to other major markets, leading up to 3) build a long-term asset allocation optimization system. Once the asset allocation system is up and running, we will continue to 4) develop new data and features to improve on precision and extending the forecasting timeframe.