MIE 1628H Final Project – Time Series Analysis of Apple Stock Close

Objective: Apple stock is an important investment vehicle. A reliable stock forecast can inform investor what information can be extracted from historical price and what the future price is likely to be which increase return of investment.

Target Variable: The target variables is **Close** price. It is a continuous non-stationary vector that is generally increasing.

Literature Review: Time series analysis is done using simple moving average, autoregressive integrated moving average, decision tree models and ANN.

Rationale of Model selection: RF and ANN are advance ML model that may be able to provide better prediction than simple models.

Results

Table 1: SMAPE and RMSE of the optimized models

	SMAPE (%)					RMSE				
	1 day	1 week	2 weeks	1 month	4 months	1 day	1 week	2 weeks	1 month	4 months
Ran.For.	16.22	49.23	38.77	58.52	173.8	0.81	1.96	1.58	2.21	4.36
GBT	15.63	20.066	20.435	197.91	189.1112	0.7871	0.986	1.0201	4.62	4.54

Table 2: Data processing pipeline of models

Model	**Used data**	**Feature transformation**	**Train/test split**	**Overfiting?**	**Model Optimization**
Ran.For.	Close, date, Open, High, Low, Volume	Imputer,MA(5,10,20,50,100,200),D,M,Y,DoW,Q,id	60/20/20	No	MaxDepth MaxBin
Dec.Tree	Close, date, Open, High, Low, Volume	MA(5,10,20,50,100,200),D,M,Y,DoW,Q,id	60/20/20	Overfitted	MinInstance

Model Optimization

RF is optimized manual grid search for Max Depth =20, 30 and MaxBins = 32, 128, 256, 512 and 1024. Only MaxBins = 128 and 1024 are performed for all prediction horizon. Figure 1 shows the SMAPE for RF improve as MaxDepth and MaxBins increase and Table 3 shows the SMAPE values. However, the improvement has diminishing returns while the computing time increase linearly. RF model is further optimized by adding moving average for

5,10,20,50,100,200 days and Day, Month, Year, Day of Week, Quarter and number of trading days since IPO using increasing Id and there is a significant drop in SMAPE for validation set.

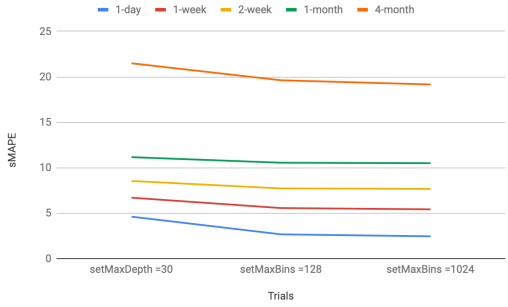


Figure 2 Gini Importance of Features for Random Forrest

Table 3 SMAPE for RF for validation Set

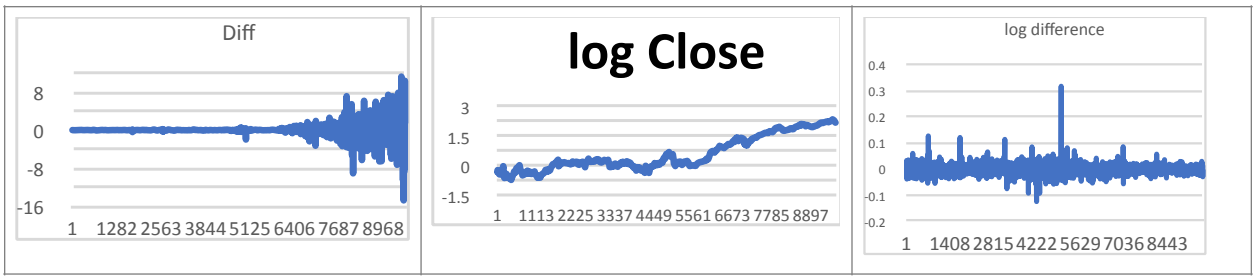
Model	1 day	1 week	2 week	1 month	4 month
MaxDepth30	4.614	6.7	8.54	11.16	21.49
MaxBins128	2.67	5.56	7.72	10.55	19.62
MaxBins1024	2.45	5.42	7.67	10.5	19.16
Sea. & MA	1.9	2.89	3.106	3.138	3.206

RF models are optimized by stationalizing the label using differencing, log and log differencing transformation. Table 4 shows the SMAPE of RF models with various horizons. Table 5 shows differencing can stationalize the series but the variance was larger in the testing set, leading to larger error. Log difference can stationalize the time series and variance is evenly distributed compared to differencing, however, most of the values are in the range of 0.05and -0.05, which means a small prediction error leads to large percentage based SMAPE. Logging strike a balance between stationalizing and maintaining a reasonable range, thus the SMAPE is the lowest for the 3 transformation strategies.

Table 4 SMAPE of RF with different transformation of label

Model	1 day	1 week	2 week	1 month	4 month
Difference	68.157	84.82	104.73	NA	NA
log	16.228	49.23	38.77	58.52	173.8
Log differencing	153.59	159.46	146.86	NA	NA

Table 5: Label Transformation for stationalization



Gradient boosting

Gradient boosting is another type of ensample of trees that minimize a loss function. The SMAPE of the model is affected by the loss function, numIterations, learningRate, MaxDepth, and MaxBins. Increasing MaxDepth and MaxBins improves performance with diminishing returns. Higher number of iterations produces more trees and improves training data accuracy, however, a number too high will overfit the data.

Model	1 day	1 week	2 week	1 month	4 month
GBT log	15.63	20.066	20.435	197.91	189.11
RF log	16.228	49.23	38.77	58.52	173.8

Feature Importance

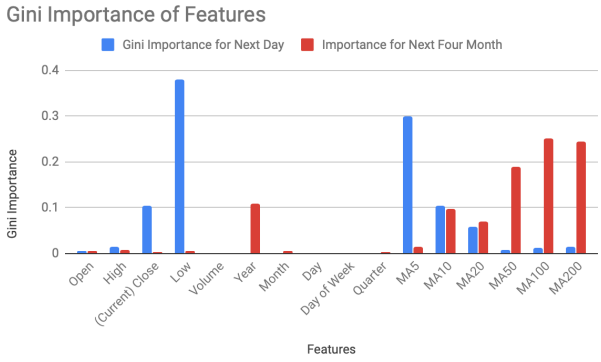


Figure 1 SMAPE for Random Forest For Validation Set

A RF model is built for each prediction horizon. Figure 1 shows the relative importance of features in predicting close price and prediction horizon influence feature importance. For Next day "Close" prediction, feature "Low", "MA5", "Close" and "MA10" stood out and have strong influence in prediction. However, for 4-Month prediction, "MA100","MA200" and "Year" are important features. In this analysis, Volume, Day, Day of week and Quarter do not have a strong influence on model's predictive power.

Mentor's Recommendation

Our mentor is Yania Shevcheenko and the recommendation includes

1. Create a simple RF model that works before adding feature engineering and hyperparameter optimization.
2. Manually adjust hyperparameters and don't use cross validation
3. Try gradient boosting model

Future work

Create transformers for logging so that the prediction is in the unit of close instead of log close. Combine the models to take advantage of the strength of each models. Feature engineering from existing data and include external data such as S&P 500 and financial data of apple such as asset, earnings, revenue and profit.