Stock Prediction

Using Autoregression

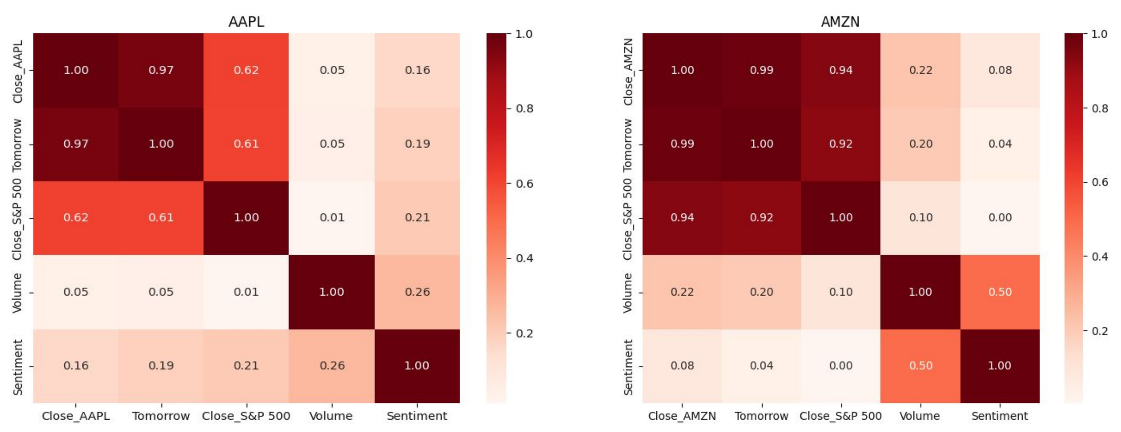
Emil Westling

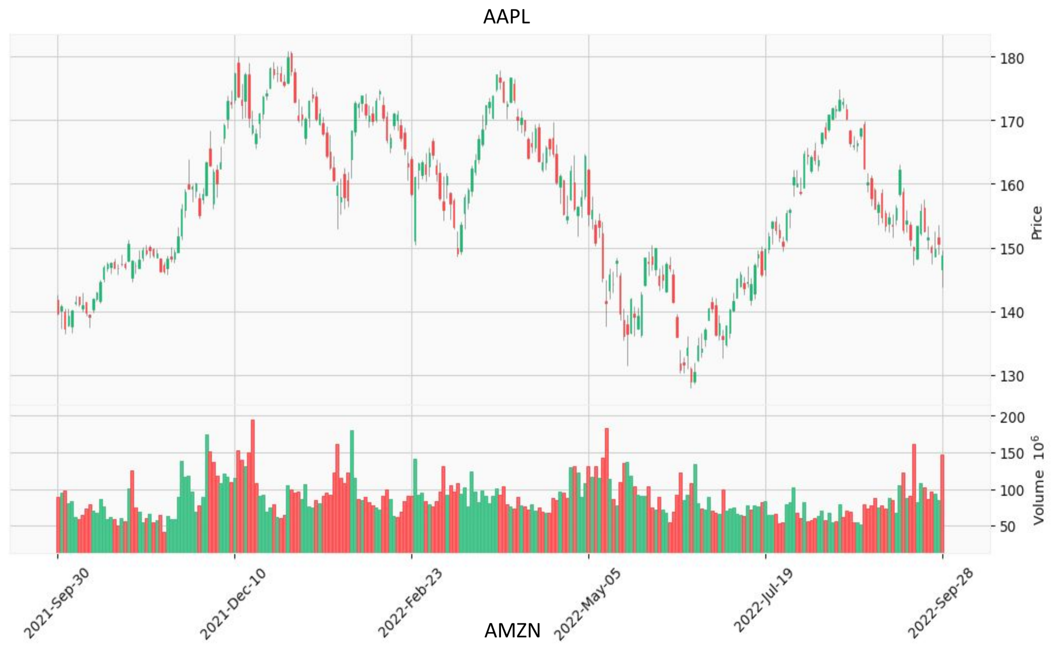
INTRODUCTION

- This report offers a comprehensive exploration of stock prediction through autoregressive models.
- Concentrates on two companies, *Apple* and *Amazon*, which constitute 7.29% and 3.51% of the S&P 500 index, respectively.
- The forecasting will be done by not only using past pricing data but also incorporating a variety of exogenous variables.
- The exogenous variables include S&P 500 pricing data, sentiment analysis, and daily trading volumes.
- Two forecasting models, one for price and one for direction prediction, are used.

DATA

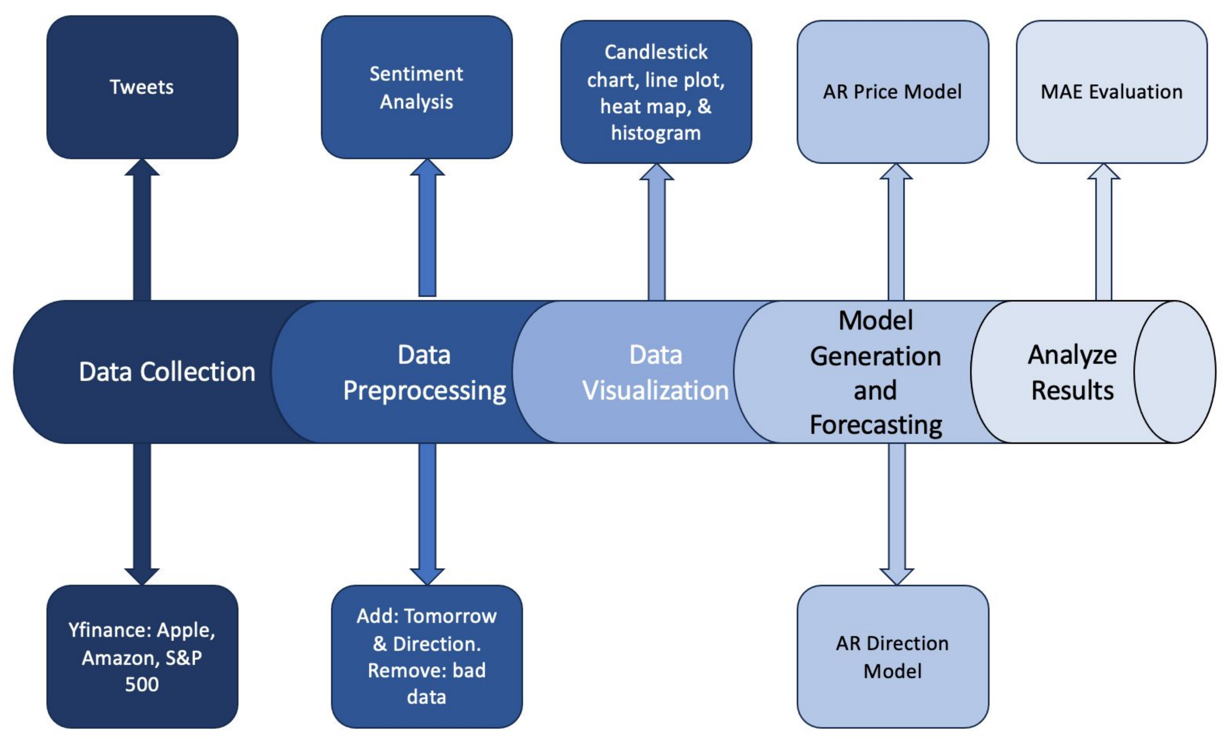
- The API *yfinance* was used to fetch pricing data for *Apple*, *Amazon*, and the *S&P 500* index.
- Tweet time series data set was found on Kaggle. All data spanned from 2021-09-30 to 2022-09-29.
- Sentiment scores were calculated for each tweet calculated using a NLP language library.
- After preprocessing, the resulting data frames consisted of a *closing price, trading volume, tomorrow's closing price, direction of tomorrow's price, and a sentiment score* for each date.
- The closing price of the S&P 500 index was also stored.





https://github.com/emilwestling/AlgorithmTrading

PIPELINE



- Data collection: fetch pricing data from yfinance and a tweet dataset from Kaggle.
- Data preprocessing: generating a sentiment scores, adding columns to the company data frames and removing bad and unnecessary data.
- Data visualization: generate various plots to get to know the data.
- Model generation: generate AR models using Statsmodels' AutoReg library.

MODEL

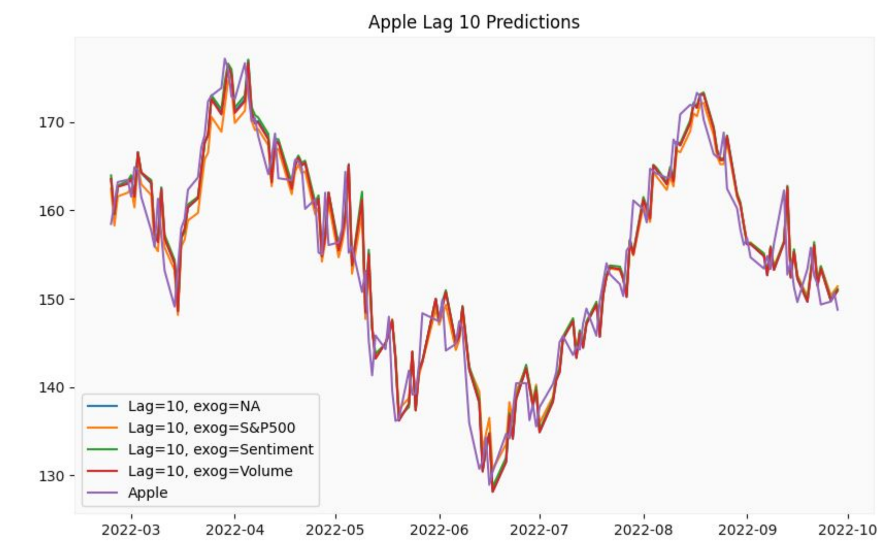
$$y_t = \delta_0 + \delta_1 t + \phi_1 y_{t-1} + \dots + \phi_p y_{t-p} + \sum \kappa_j x_{t,j} + \varepsilon_t$$

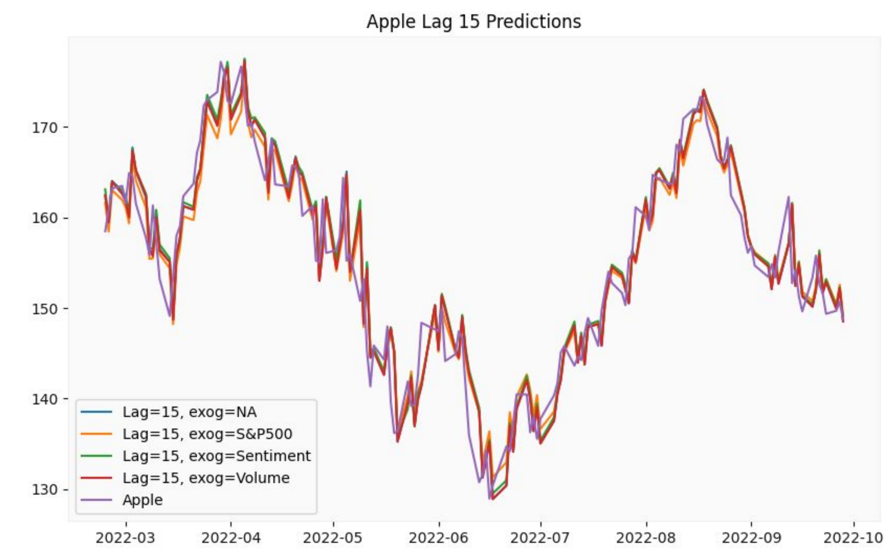
- Models were created using the Statsmodels' *AutoReg* library. The equation above is the underlying forecasting equation.
- ϕ are the weights for each price observation and κ are the weights for the exogenous variables.
- δ_0 is the y-intercept whereas $\delta_1 t$ encodes a linear time trend relationship
- The two models were identical aside from the fact that one model was only concerned with the direction of the predicted price.

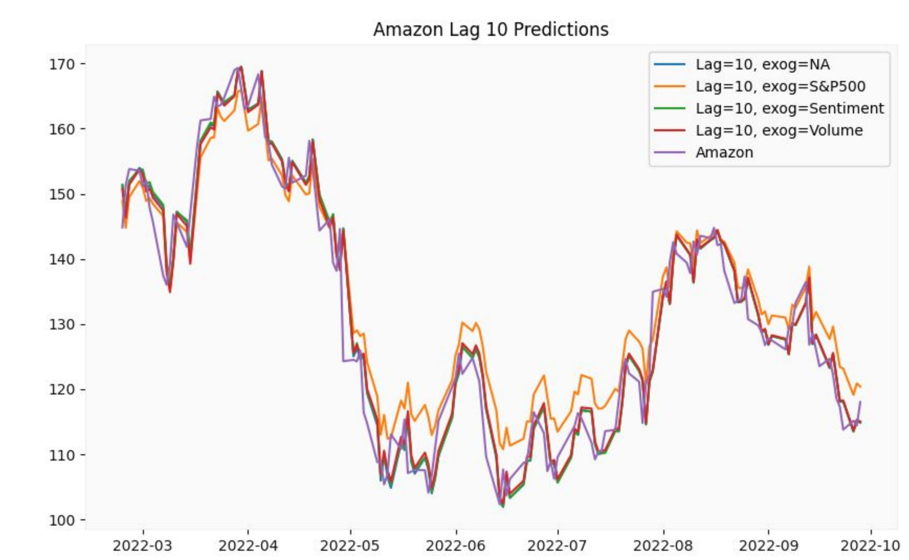
RESULTS

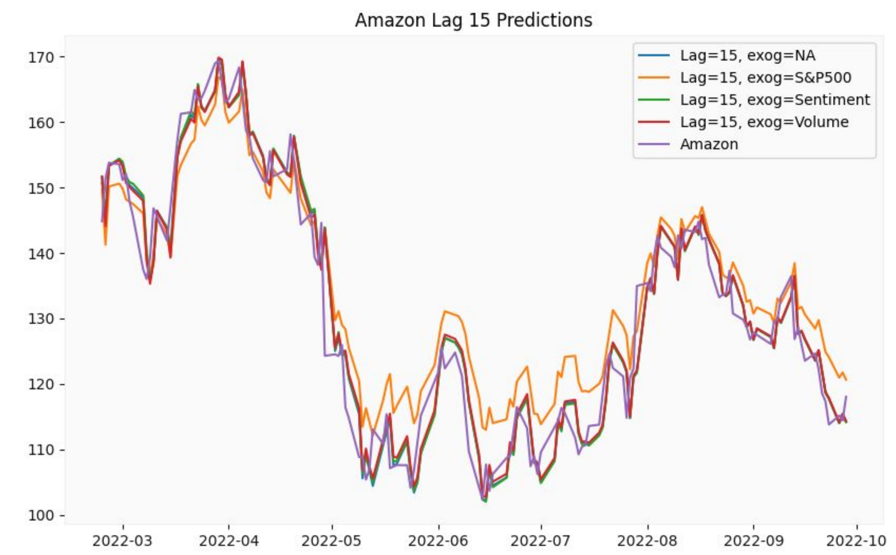
	Company	Amazon		Apple	
		MAE_Direction	MAE_Price	MAE_Direction	MAE_Price
Lag	Exog				
10	NA	1.019868	**3.346440**	0.953642	2.762461
	S&P500	**0.993377**	4.859555	**0.940397**	2.795329
	Sentiment	1.033113	3.356564	0.953642	**2.744114**
	Volume	1.006623	3.360806	0.953642	2.755890
15	NA	1.033113	3.411418	0.966887	2.839545
	S&P500	**0.993377**	5.569160	1.059603	2.787392
	Sentiment	1.006623	3.413289	1.006623	2.829494
	Volume	1.006623	3.443118	0.980132	2.835158
20	NA	1.086093	3.424944	0.993377	2.853466
	S&P500	**0.993377**	5.841501	0.980132	2.768990
	Sentiment	1.033113	3.474168	0.953642	2.837408
	Volume	1.086093	3.447584	0.980132	2.851404

- The bolded scores are the best performing forecasts for that company.
- The model predicts direction much better than price.
- Exogenous variables does not improve the performance significantly.
- The lower the lag, the better performance in general.
- For Amazon, the exogenous variable S&P 500 does poorly predicting price, but particularly well predicting direction.
- The predictions are fairly similar across the board indicating little impact of the exogenous variables.









CONCLUSION

- This project served as a great introduction to the field of computational finance
- To expand on this project, it would be necessary to explore different models, companies, exogenous variables, and prediction horizons.