Stock Price Forecasting

Team: Harper, Daniel, Roy


Agenda
01 Problem Statement
02 Data & Transformation
03 Assumption
04 Approaches
05 Proposed Solution
06 Results
07 Future Work

01 Problem Statement

Problem Statement

In this project, we aim to build time series models
to forecast day-to-day Google stock price
with the stock prices of four other companies
in the communication service sector
and improve the predictive accuracy of the models.

02 Data & Transformation

Data Description

- **Target Variable:** Google daily closing price
- **Predictor Variables:** News Corp, Walt Disney Co, Netflix, and AT&T daily closing prices
- **Data Source:**

```
library(quantmod)
getSymbols('GOOGL', from='2015-1-2', to='2020-1-3')  #from Yahoo Financials
head(GOOGL)
```

	GOOGL.Open	GOOGL.High	GOOGL.Low	GOOGL.Close	GOOGL.Volume	GOOGL.Adjusted
2015-01-02	532.60	535.80	527.88	529.55	1324000	529.55
2015-01-05	527.15	527.99	517.75	519.46	2059100	519.46
2015-01-06	520.50	521.21	505.55	506.64	2722800	506.64
2015-01-07	510.95	511.49	503.65	505.15	2345900	505.15
2015-01-08	501.51	507.50	495.02	506.91	3652700	506.91
2015-01-09	508.18	508.60	498.65	500.72	2100000	500.72

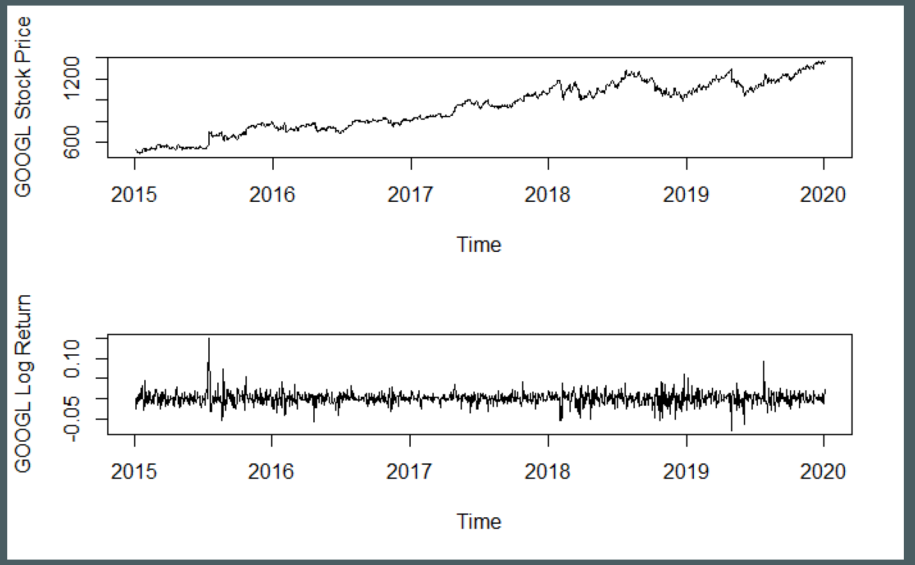
- **Train/test data:**
  Train: from 2015-1-2 to 2019-12-31; test: 2020-1-3 to 2020-1-31

Data Transformation

- Log Returns Calculation:

$$r_t = \ln\left(\frac{P_t}{P_{t-1}}\right) = \ln(P_t) - \ln(P_{t-1})$$

- Benefit of Log Return in Finance:
  - **Log-normality**:
    When assume that prices are distributed log normally, then log(1 + r_i) is conveniently normally distributed
  - **Approximate raw-log equality**:
    When returns are very small, log(1 + r_i) becomes very close to r_i.
  - **Time-additivity**:
    The compound return over n periods is the difference in log between initial and final periods



$$\sum_i \log(1+r_i) = \log(1+r_1)+\log(1+r_2)+\cdots+\log(1+r_n) = \log(p_n)-\log(p_0)$$

03 Assumption

Assumptions about Stock Price Prediction

01 Auto-Correlation

A stock's future price is highly correlated to its own history, and thus is predictable to some extent with auto-regression

02 Cross-Correlation

A stock's price is also correlated to prices of related stocks, which are in the same or close category, and with either a similar or an opposite business trend

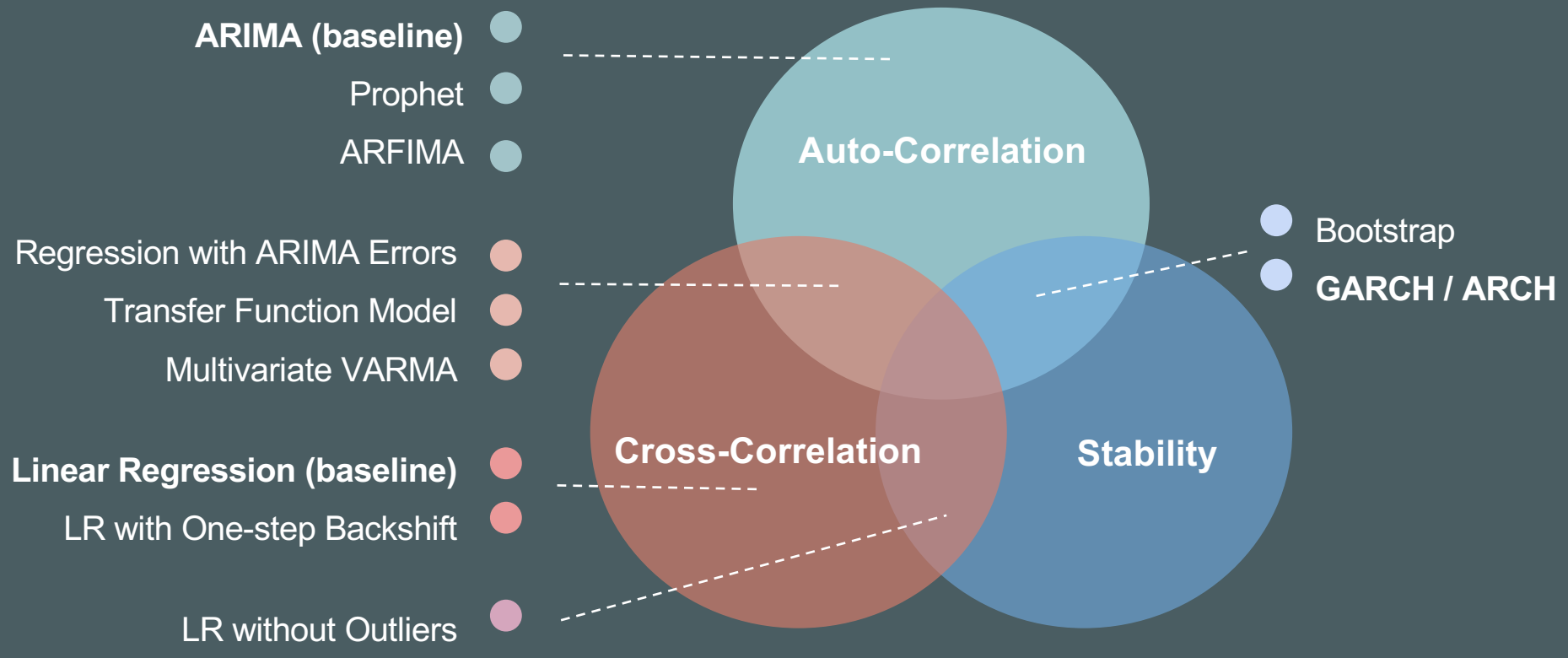
Category: "Communication Service"
'NWSA' : traditional media business
'DIS' : content producer business
'NFLX' : emergent online business
'T' : infrastructure provider business

03 Stability vs. Variance

Too many exogenous factors influence the prices of a stock. Stabilizing our predictions would be important as we are unable to define all the unpredictable factors

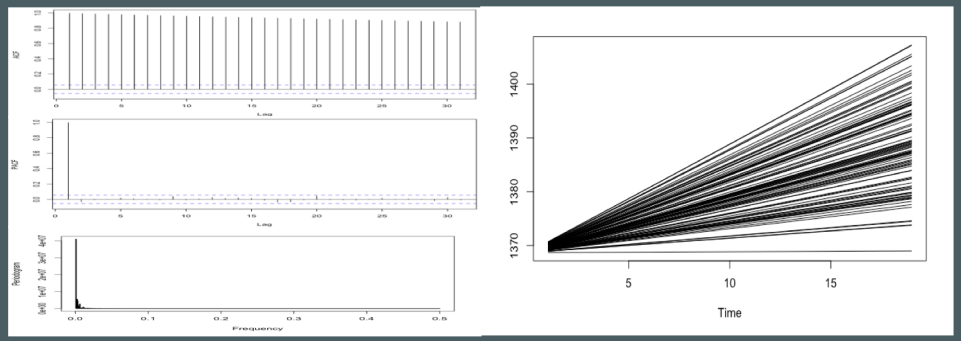
04 Approaches

Methodology



ARIMA Modeling Approaches

- **Stationarity of original TS**
  - High Auto-Correlation
  - PACF lag=1
  - No Seasonality



- **Baseline Modeling**
  - Stationarity check
  - auto.arima() gives (2,1,1)
  - Arima(4,0,4) after order tuning
  - Residual diagnosis
  - Forecasting & evaluation

- **ARIMA model based Bootstrap**
  - Simulating 99 samples with ARIMA baseline model parameters plus bootstrapped residuals
  - Fitting the 99 samples and forecasting
  - Averaging all the forecasts

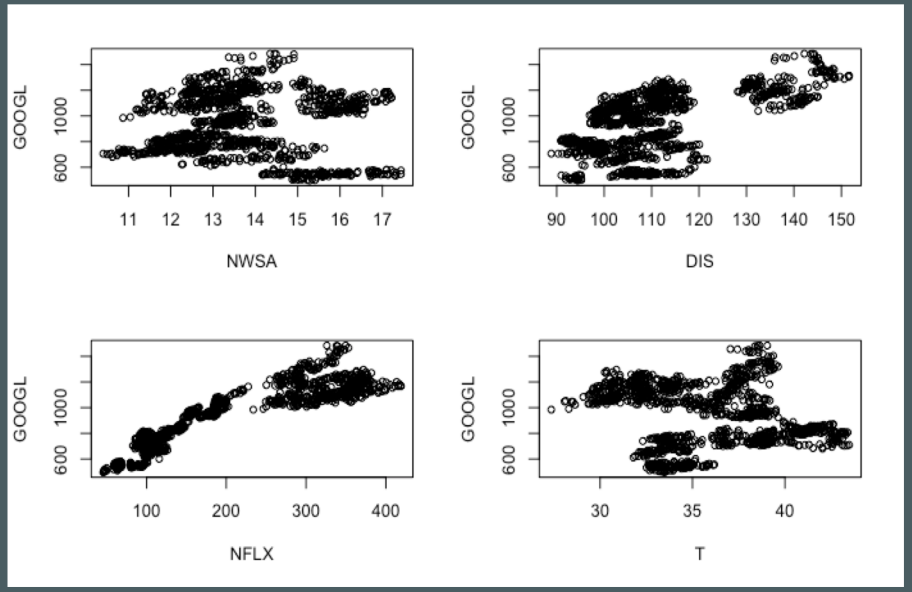
Linear Regression

Reasons for Linear Regression

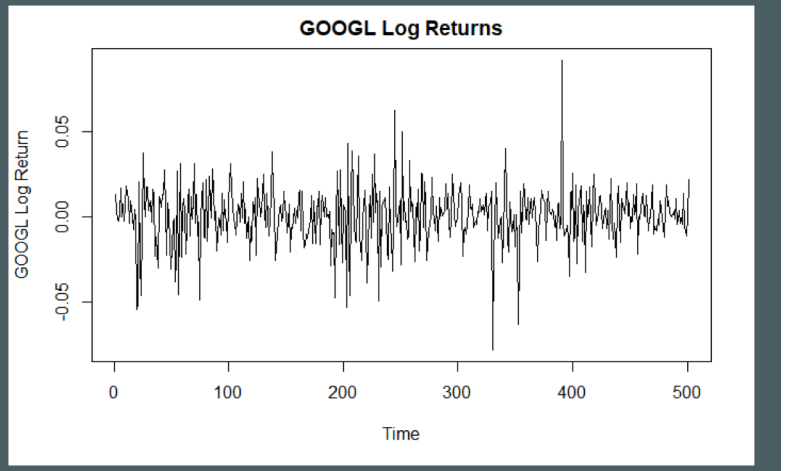
- Stocks are correlated
- Returns are autocorrelated, momentum strategies make money
- Returns and volatility of returns are correlated

Basic Approach

- Simple linear regression and regression with arima errors
- Must shift information available to the model so you don't let it "see the future" when predicting
- Fit based on previous days T, NWSA, NFLX, DIS, and GOOGL log returns and the previous log open-close range of GOOGL



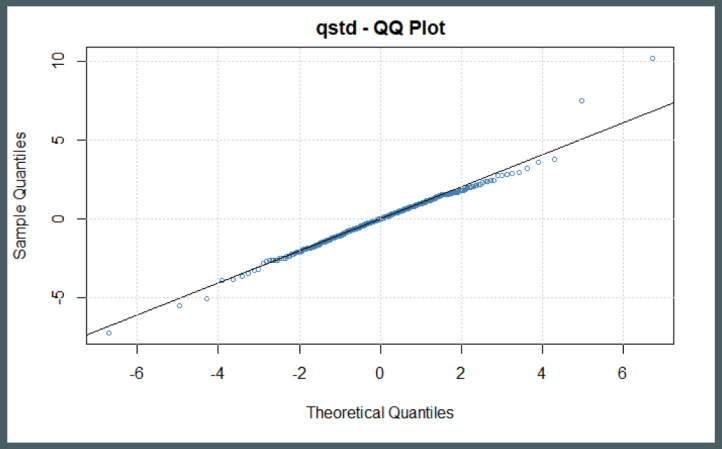
GARCH



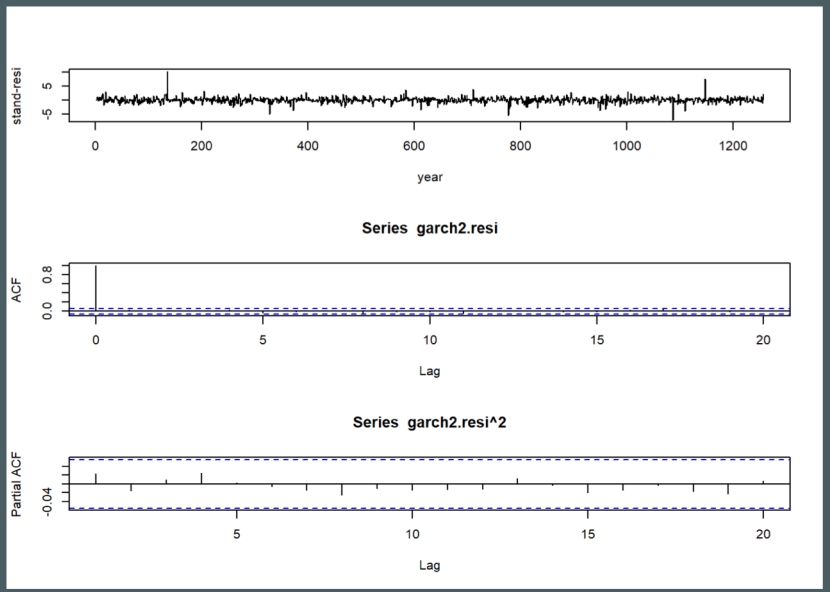
- **Reasons of selecting GARCH model:**
  - The goal of GARCH and ARCH models is to provide a measure of volatility that can be used in financial decision-making
  - Compared to the ARCH model, the GARCH model is usually much more parsimonious
- **Modeling Steps:**
  - Fit an ARMA model for the return series to remove any linear dependence
  - Use the residuals of the ARMA model to test for autocorrelation and ARCH effects
  - Fit several different ARCH/GARCH models
    - Using the PACF of the residual squared to determine the ARCH order
    - Fit several models with different distribution for the conditional variance (Gaussian, Student t, Skew Student t …)
  - Check the significance of parameters for each model
  - Select the best model based on the AIC and BIC
  - Transform the predicted log returns into prices

GARCH

- **Findings:**
  - The best model is GARCH(1,1) with student-t innovations
  - All the parameters are significant
  - Residuals are not autocorrelated
  - Residuals are close to a normal distribution




```
Error Analysis:
         Estimate  Std. Error  t value Pr(>|t|)
mu     3.622e-04   1.327e-04    2.729 0.006349 **
ar1    6.349e-01   8.163e-02    7.778 7.33e-15 ***
ma1   -6.179e-01   8.534e-02   -7.240 4.48e-13 ***
ma2   -1.132e-01   3.307e-02   -3.422 0.000622 ***
omega  3.189e-05   9.791e-06    3.257 0.001125 **
alpha1 1.545e-01   3.426e-02    4.510 6.50e-06 ***
beta1  7.137e-01   6.648e-02   10.735  < 2e-16 ***
```



05 Proposed Solution

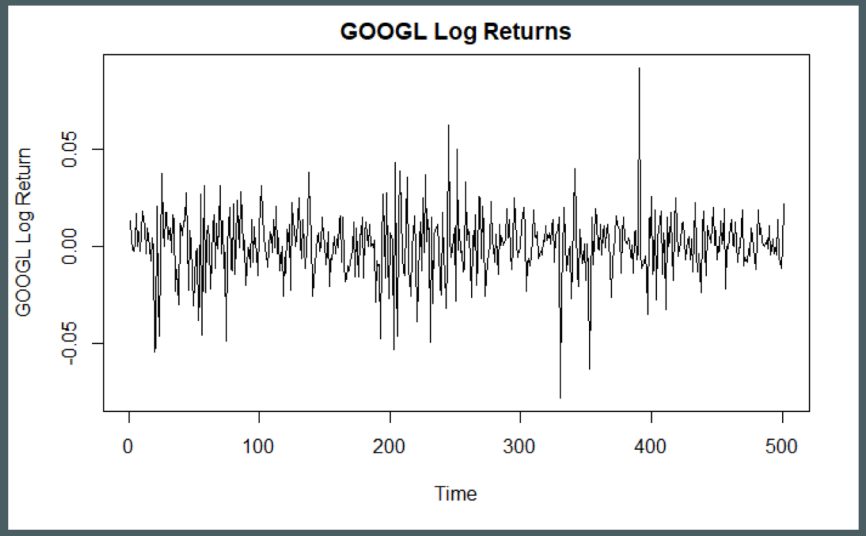
Models Comparison using sMAPE

Categories	Models	Original data	Transformed data
Auto-Correlation	ARIMA (baseline)	0.05053411	0.04362964
	ARFIMA	0.11405243	0.04542452
	Prophet	NA	0.08365312
In-between	Regression with ARIMA errors	0.05146522	0.05482949
	Regression with ARIMA errors, no outliers, n=60	*NA*	0.02625852
	Transfer Function Model	0.05533000	0.04451388
	VARMA	*NA*	0.04265250
Cross-Correlation	Linear Regression 5-year window	*NA*	0.03879357
	Linear Regression, no outliers, n=60	*NA*	0.02938541
Stability	GARCH	0.03927748	0.04862327
	ARIMA Bootstrap	*NA*	0.04321630

Regression with ARIMA errors, outlier removal, n=60

Outlier Analysis and Handling

- Outlier analysis on each individual return time series
- Compare identified outliers to known market events, adjust if reasonable



Varying Fit Window Lengths

- All models fit over 5 years predict some mean return with little variation
- Vary the length of the fit window to choose more appropriate fit time window

Mean sMAPE over train

	20	60	90
arima	0.03894521	0.03539375	0.03483157
regression.arima	0.03892030	0.03537137	0.03481717
regression	0.03892124	0.03537199	0.03480973

Standard deviation sMAPE over train

	20	60	90
arima	0.05099452	0.03153994	0.03365244
regression.arima	0.05098095	0.03153587	0.03364129
regression	0.05098040	0.03153536	0.03364589

06 Result

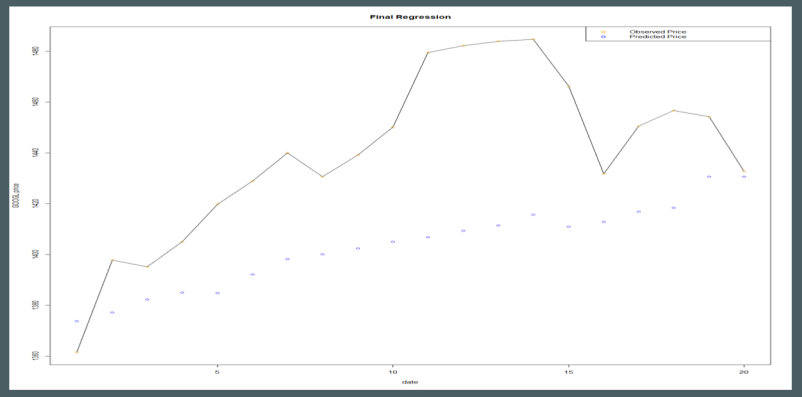
Applying Prediction Results to Trading

Let's suppose:

- We have to set our buys and sells for the whole month of January
- Use the simple strategy:
  - If our prediction indicates a local minimum then buy 100 shares for the close price that day
  - If our prediction indicates a local maximum then sell 100 shares for the close price that day
- We must close our position at the end of the month

How do we perform?

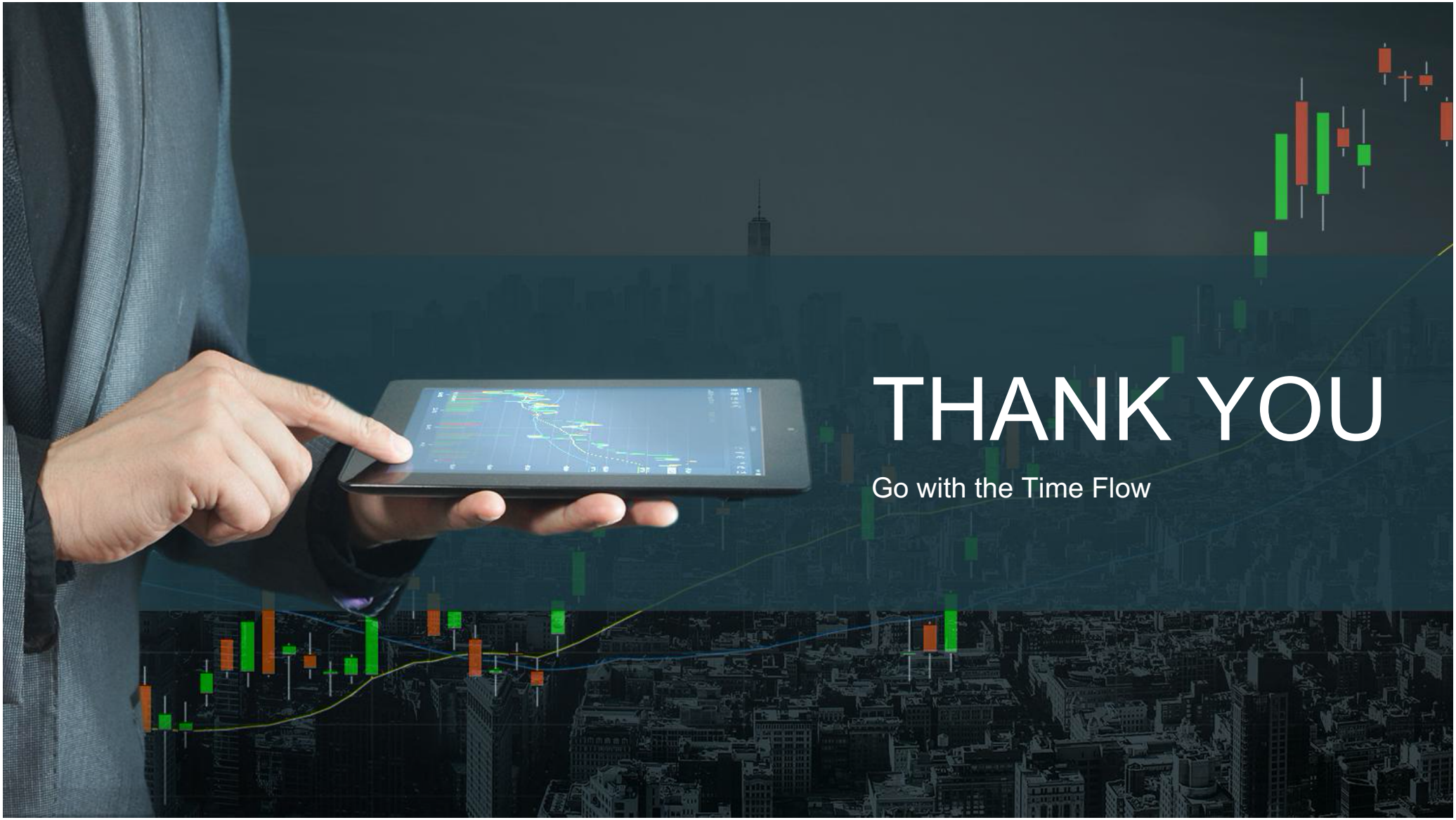
- Regression with ARIMA errors
  - 6 trades
  - 0 losing trades
  - $2,523 in profit



07 Future Work

How can we improve?

- For Using sliding window to optimize predictions
- For Cross-Correlation: Incorporating exogenous factors
  - Macro-economic (inflation, unemployment, recession)
  - Market sentiment (needs metrics to quantify)
- More advanced algorithms
  - Long Short Term Memory networks (LSTM)
  - RNN


THANK YOU
Go with the Time Flow