## Executive Summary

### Background

Low volatility can be a very effective predictor for future returns in China A-shares market. According to MSCI, shares with a low volatility has produced strong cumulative returns over the past five years and the MSCI China A Minimum Volatility Index returns outperformed the MSCI China A Index returns by 6.7% per year. As such, AXA Investment Managers wants to build a low volatility signal that is indicative of China A-shares market stock's current and forecasted volatility.

### Problem Objective

Our objective is to build a company-level quantitative predictive model to identify stocks with low future volatility using fundamental data that consist of features of Chinese A-shares companies. Our model should strive to outperform a naive guesstimate of the future volatility.

### Methodology and Results

Data Cleaning & Scoping: Columns that had many missing or little unique values were removed. Considering that we are forecasting stock volatility, we imperatively took note of the number of data points each company had. Companies that were tracked earlier had more data points. Upon inspection, we learnt that 632 of all 1,800 companies, had a full set of data - data points beginning from January 2001. We approached our model building with these *full-data companies*.

Defining Volatility: This project defines volatility, the main dependent variable, as a measure of the variance of a company's returns. The volatility of a company is calculated by taking the standard deviation of the last 12 values of the company's returns. This rolling window approach gives us an accurate indicator of a company's volatility by month.

Regression model attempt: Having established our dependent variable, we then attempted to find any correlation between the volatility and the other features. Iterating through each company, we found that the correlations calculated with each variable varies, making model building tricky.

ARIMA model attempt: Upon studying the autocorrelation of the volatility of the companies, we learnt that a company's volatility is highly dependent on its previous values. Hence, we leaned towards building a time series forecasting model using the ARIMA (auto regressive integrated moving average) model. Despite having visually assuring plots of our forecast, we needed a way to analyse and benchmark our model.

### Model Analysis

Building a naive model that took the mean of the last 12 variance data of a company allowed us to analyse if our model was indeed better than a guesstimate. We found that the mean squared error of our model significantly outperformed the naive model by 52%.

### Limitations

Firstly, training our model with full-data companies meant that companies with less data points may suffer data inaccuracy. The ARIMA model we built on also requires that data points are stationary. However, this is compromised due to the spikes in market volatility in 2008 and 2015.

### Value Proposition

AXA IM's go-to approach has always been towards industry-level analysis. However, our company level approach was able to bring about fresh insights to the analysis of China A-shares. The company level approach was able to bring out tangible features that the industry-level approach may lack, giving us greater depth into our insights.

## Company background

AXA Investment Managers (IM) operates as an investment management company. According to AXA's investment philosophy, low volatility can be an effective predictor for future returns in China A-shares market. Taking an active, long-term approach, it is vital that AXA IM constantly find new ways to analyse data to help clients secure a better investment experience.

## Problem definition

In this project, we are required to establish our understanding of the volatility of a stock and use that definition to forecast future volatilities. We approach this by combining company fundamental data (from the income statement and balance sheet), establishing its current volatility to generate a quantitative signal that is indicative of a company's future volatility over the next 12 months. A description of the data provided can be found in Appendix A.

## Motivation Methodology

### Data Cleaning

Before using the data for prediction modelling, we need to first define which columns (e.g. asst_risk, total_ror) have sufficient data that may become a possible variable for our model. We removed columns that had less than 10% of non-null data or unique values. Additionally, companies had varying number of data points; companies tracked earlier had more data, with the earliest data point at January 2001. We found that 632 out of 1,800 companies had a full dataset - a sizeable number we went forward to build our model with.

## Defining Volatility

We have defined the main dependent variable, volatility, as the standard deviation of the rate of return; and we have explored two approaches in finding it.

### Year-by-year volatility

Our first approach was to try to find a standard deviation value for one entire year. Our description of a year by year volatility can be found in Appendix B. However, this approach has several limitations:

- We may miss periods within a year (e.g. 3-month or 6-month) where the volatility spiked or plummeted or both, which would not give a good way of prediction.
- Aggregating the volatility by year transforms 12 monthly values into 1 yearly value, making feature engineering for regression techniques difficult as we are comparing 12 values to 1.

Therefore, we decided to find a better way to calculate volatility as this approach data points, leading to an inaccurate model.

### Rolling window volatility

The second approach finds the rolling standard deviation, enabling us to understand how volatility has changed over time or behaved in different market conditions. We computed the standard deviation from a rolling window of 12.

A representation on how the past rate of return data is used to calculate the present volatility can be found in Appendix C. In addition, in the case of missing values, we set the window to accept a minimum of 4 values.

## Exploration of data

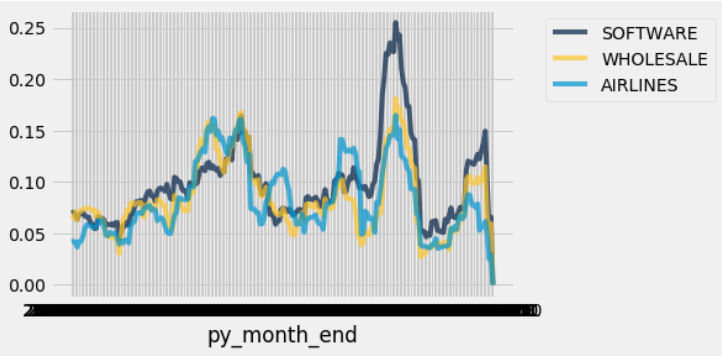



*Figure 1: Rolling standard deviation across different industries*

In order to identify specific industries that are more sensitive to volatility changes, we plotted the rolling standard deviation across different industries, from year 2001 to 2019. From *Figure 1*, software has higher volatility than other industries. Our client can use this information to justify their investment decisions while using our model for more informed investment experiences. In addition, we can see the sudden spikes in rolling standard deviation for year 2008 and 2015, which are caused by the financial crisis and Chinese stock market boom respectively.[1][2]

[1] Amadeo, K. (2019). When and Why Did the Stock Market Crash in 2008?. [online] The Balance. Available at: https://www.thebalance.com/stock-market-crash-of-2008-3305535 [Accessed 11 Dec. 2019].
[2] Nytimes.com. (2019). Opinion | China's Unsettling Stock Market Boom. [online] Available at: https://www.nytimes.com/2015/06/15/opinion/chinas-unsettling-stock-market-boom.html [Accessed 11 Dec. 2019].

## Modelling Approach 1: Regression

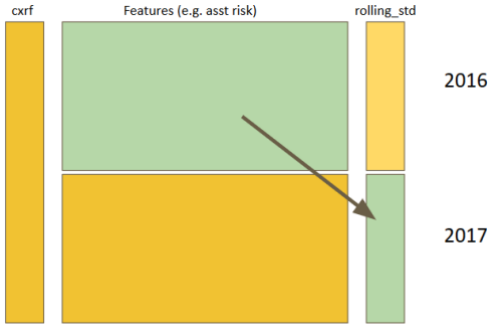



*Figure 2: Initial regression modelling approach*

Having established our dependent variable, our first approach was to identify similarly-behaving-variables by finding the correlation of volatility with other features *(Figure 2).* After iterating through each company, we created a correlation matrix heat map *(Figure 3).* In the table, green and red indicates a positive and negative relationship respectively, and the intensity of a colour's concentration depends on the correlation magnitude. We observed that the correlation varies across each company for each variable, making model building complicated as there are multiple correlation factors to consider. Since we wanted to create a predictive model that is easily replicable, we decided not to delve deeper into regression models and considered other modelling approaches.

*Figure 3: Correlation matrix heat map*

## Modelling Approach 2 : Time Series

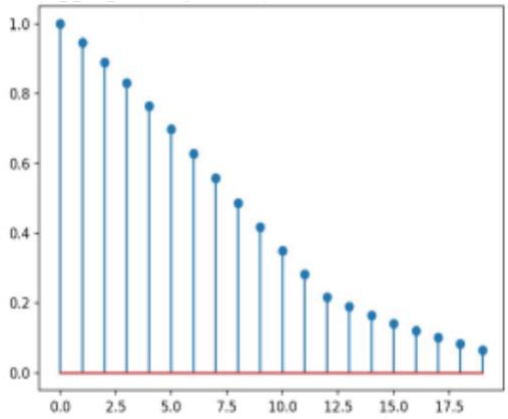



*Figure 4: Combined ACF plots of volatility of all companies (aggregated by mean)*

After moving out of the regression approach, we did an ACF analysis on the variance of all companies. The result of the ACF Plot *(Figure 4)* showed that a company's volatility is highly dependent on its previous values. Hence, we decided to build an ARIMA (Auto Regressive Integrated Moving Average) time series forecasting model.[3]

ARIMA is a forecasting algorithm that uses past values of the time series to predict the future values. The ARIMA model is able to take into account the lags of forecast errors between consecutive data and patterns in growth/decline in data, including the rate of the changes. This is achieved through the use of the autocorrelation function, which is easily controlled by differences, auto-regression and moving averages, without performing complicated transformations or using extra surrogate variables.

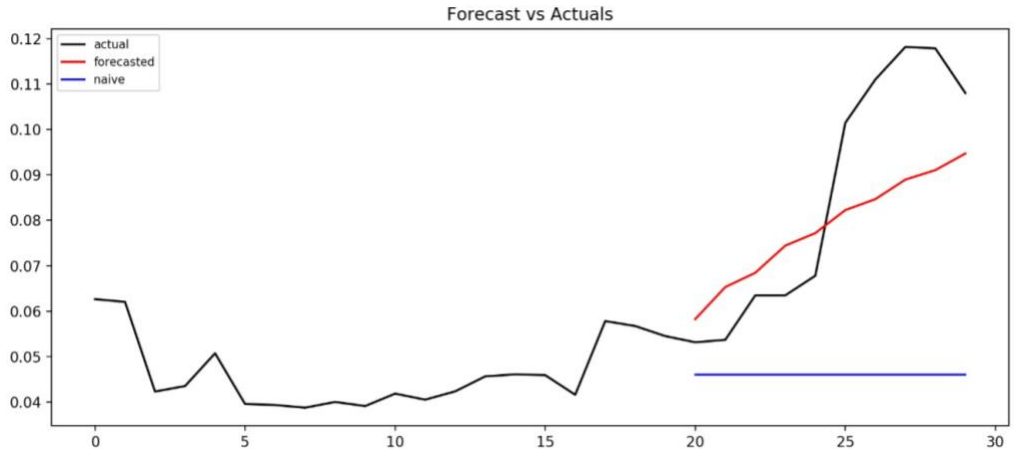
3 An ACF plot gives the values of auto-correlation function of any series with its lagged values, measuring the internal association between observations in a time series

## Tools

In the exploration data analysis phase, we utilized *R* and *Tableau* to get a visual understanding of the data. To manipulate data for specific analysis (e.g. aggregating industries), we used the *data science* and *pandas* library from Python 3.x before exporting it into *Excel* to get a visual sensing of our data. Subsequently, we used *Gretel*, a cross-platform software package for econometric analysis, to experiment with different models for our data. Finally, we used the Python's *statsmodel* and *scikit-learn* to polish up our model for presentation. The scripts used can be found here.

## Main results

To compare the accuracy of the ARIMA model, we created a naive prediction model. This naive model set the future 12 months rolling standard deviation value to be the mean value of the previous 12 months value. *Figure 5* shows the ARIMA and naive model forecast for one company.




*Legend : Red → ARIMA model, Blue → naive model, Black → Actual values*

*Figure 5: ARIMA and naive model forecast for one company.*

After running tests on all full-data companies, the results are as below:

| | |
|---|---|
| Average mean-squared-error of ARIMA model | 0.001637108 |
| Average mean-squared-error of naive model | 0.003174031 |
| Percentage difference | 52% |

*Figure 6: Results of average mean-squared-error of ARIMA model, actual values and naive prediction model*

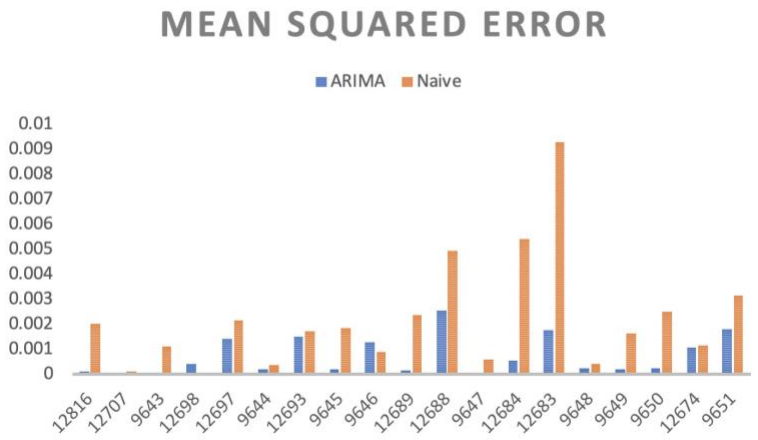



*Figure 7: Mean Squared Error between Arima and Naïve model*

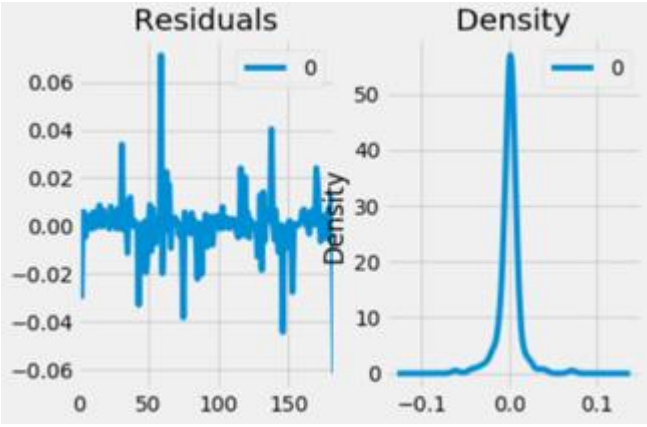



*Figure 8: Residual plot*

*Figure 8* shows a random pattern with density centred at 0, indicating that the residuals are uncorrelated and have zero mean. This indicates that while our model may not be extremely precise, it is accurate.

**Assumptions**

- The time series is stationary – the mean and standard deviation is constant
- The data for the ARIMA model has to be univariate (works on a single variable). In this case, we are using rolling standard deviation as our variable.
- We assume that a rolling standard deviation with 12 months sliding window is a good representation of volatility

**Limitations**

One limitation of the model is that the model does not take into account large scale global happenings. Thus, the model is subject to world events like the 2008 financial crisis and the 2015 Chinese stock market boom.

Furthermore, since we are training our model with full-data companies, companies with less data points will have less accurate predicted volatility values.

**Conclusion**

ARIMA model has its advantages in time-series analyses. The trend, autocorrelation is easily controlled by autoregressive and moving average without performing complicated transformations. Rolling standard deviation is a good statistical measurement of market volatility. All in all, the entire methodology is well described and is made to be easily replicable.

## References


1. Wei, Z. (2018). *How the low volatility factor has performed in China A shares - MSCI*. [online] Msci.com. Available at: https://www.msci.com/www/blog-posts/how-the-low-volatility-factor/01132841955 [Accessed 10 Dec. 2019].
2. Machine Learning Plus. (2019). *ARIMA Model - Complete Guide to Time Series Forecasting in Python | ML+*. [online] Available at: https://www.machinelearningplus.com/time-series/arima-model-time-series-forecasting-python/ [Accessed 10 Dec. 2019].
3. Brownlee, J. (2019). *How to Create an ARIMA Model for Time Series Forecasting in Python*. [online] Machine Learning Mastery. Available at: https://machinelearningmastery.com/arima-for-time-series-forecasting-with-python/ [Accessed 10 Dec. 2019].
4. Amadeo, K. (2019). *When and Why Did the Stock Market Crash in 2008?*. [online] The Balance. Available at: https://www.thebalance.com/stock-market-crash-of-2008-3305535 [Accessed 11 Dec. 2019].
5. Nytimes.com. (2019). *Opinion | China's Unsettling Stock Market Boom*. [online] Available at: https://www.nytimes.com/2015/06/15/opinion/chinas-unsettling-stock-market-boom.html [Accessed 11 Dec. 2019].

# Appendix

## Appendix A: Data provided

| Column Name | Description |
|---|---|
| cxrf | Company index number |
| returns | Returns of a company. (e.g. if a company loses half its worth, its return will be 0.5) |
| industry | Industry that company belongs to (e.g. service, F&B) |
| date | Month that data is recorded of the company |
| asset risk | 45 different financial indexes of the company that we could possible use to predict its volatility |
| ... | |
| cnc | |

## Appendix B: Year-by-year volatility

| cxrf | returns | date |
|---|---|---|
| 3544 | 0.5076 | Jan 18 |
| 3544 | 0.8275 | Feb 18 |
| 3544 | 0.3782 | Mar 18 |
| 3544 | 0.4761 | Apr 18 |
| 3544 | 0.9074 | May 18 |
| 3544 | 0.3025 | Jun 18 |
| 3544 | 0.1246 | Jul 18 |
| 3544 | 0.9623 | Aug 18 |
| 3544 | 0.1385 | Sep 18 |
| 3544 | 0.2920 | Oct 18 |
| 3544 | 0.3222 | Nov 18 |
| 3544 | 0.7232 | Dec 18 |



| cxrf | year | std |
|---|---|---|
| 3544 | 2018 | 0.2795 |

Appendix C: Rolling Window Volatility

| cxrf | returns | date | rolling_std |
|---|---|---|---|
| 3544 | 0.5076 | Jan 18 | |
| 3544 | 0.8275 | Feb 18 | |
| 3544 | 0.3782 | Mar 18 | |
| 3544 | 0.4761 | Apr 18 | 0.236268513 |
| 3544 | 0.9074 | May 18 | 0.281560159 |
| 3544 | 0.3025 | Jun 18 | 0.31146446 |
| 3544 | 0.1246 | Jul 18 | 0.410351122 |
| 3544 | 0.9623 | Aug 18 | 0.438084996 |
| 3544 | 0.1385 | Sep 18 | 0.098505482 |
| 3544 | 0.2920 | Oct 18 | 0.438084996 |
| 3544 | 0.3222 | Nov 18 | 0.098505482 |
| 3544 | 0.7232 | Dec 18 | 0.236268513 |