UNIVERZITA KOMENSKÉHO V BRATISLAVE FAKULTA MATEMATIKY, FYZIKY A INFORMATIKY

APLIKÁCIA PRE VIRTUÁLNE OBCHODOVANIE S CENNÝMI PAPIERMI

(Diplomová práca)

2021 **Bc. Eduard Kriváne**k

Introduction

What is the fair value for a company? With the rise of the internet, plenty of information became available for the public and we stated to manage our life from our convenience. One of the conveniences is investing. Within a few minutes we can create an online brokage account, buy some stock and put our money to work. Simple as that, but then a question arises, why the majority so called investors lose money?
We could bring up multiple points which would justify our announcement, but I would like to highlight two of them, which, in the end, were the main motivation of writing this thesis.

The first group are psychology factors. We, as humans, are emotional creatures, that means, many times we tend to make decision by pure impulse without any knowledge of the subject. In world of finance, we refer to this phenomenon as fear of missing out. In simple term, we want what everybody wants. Watching an increasing price of a ticker symbol may result a need of jumping to the trade and also the opposite side, seeing our position decreasing by its value can motivate sell our assets. This impulsive behaviour may be translated into stock market by high buying or selling pressure of a given company, ignoring its long-term value.

The second point is lack of knowledge. If we see a ticker symbol trading for a specific value, how can we say that it is over or under valued? Obviously, we need to analyse the company to make a judgment. There are two types of analysis. Technical, which tries to make prediction of future the movement by evaluating historical daily prices from charts and fundamental, which examines multiple financial metrics of a given company.

To address these two problems, we will need to make our application to fulfill the following roles. First of all, we want to give our users the opportunity of trying to trade shares of the company without any necessity of money deposit. One of the functions of our application will be the ability of simulating real time trading, in other words paper trading (TODO chapter name). Users will start with a predetermined amount of money and can buy or sell stock. The application will record all transactions and generate overview of people's performance.
To solve the valuation part, we will need to obtain financial metrics for every single company (TODO chapter name) and compute price prediction models using neural networks (TODO chapter name) and time series analysis (TODO chapter name).

In the following chapters (TODO chapter name) we will first describe the most common time series analysis and try to predict the 2008 and 2020 market crash. Next, we will use neural network for the same purpose (TODO chapter name) and compare its result with the precious models (TODO chapter name). Simultaneously we will be building parts of our application starting with a content delivery server (CDS) (TODO chapter name) which will retrieve and modify financial metrics from multiple external APIs. To follow microservice architecture, which restricts mixing different module functionalities into one unit, we will create another server which will provide our modified financial data from CDS to the client and also persisting into database (TODO chapter name). Last but not least, out solution should be publicly accessible by anyone, so we will talk about dockerization and deployment to one the cloud (TODO chapter name).

Time series analysis

When we observe behaviour of a single entity measured in data points and we would like to forecast its future value, we can use interpolation or extrapolation. Interpolation is a type of estimation where we construct a new data point within the range of know values obtained by sampling or experimentation. On the other hand, when we try to estimate some values beyond the original observation range, we use extrapolation, but the further we estimate from the known values the greater uncertainty and a higher risk of producing meaningless results arise.

Time series analysis is a type of extrapolation, where a set of numerical measurements of the same entity is taken at equally spaced intervals over time and its goal is to make a forecast for the future. It has four aspects of behaviour. Trend what represents the overall long-term direction of the data series. Seasonality occurs when there is repeated behaviour in the data which occurs at regular intervals. Cycles happens as a result of up and down pattern that is not seasonal and can vary in length, which makes them more difficult to detect than seasonality. Lastly, random variations which can be found in all data. Some time series will be very regular with random variations while others may consist of not much else. The general equation for time series model is as follows

$$y(t) = x(t)\beta + \varepsilon(t)$$

Where $y(t) = \{y_t; t \in N\}$ is a sequence, indexed by the time subscript *t*, which is a combination of an observable signal sequence $x(t) = \{x_t\}$ and an unobservable independent random variation $\varepsilon(t) = \{\varepsilon_t\}$, also called white noise.

By definition, a time series is called white noise if the variables are independent and identically distributed with a zero mean and each value has no correlation with all other values in the data set. It has two meaningful concepts in forecasting. We cannot reasonably model a prediction if it consists of only white noise, because by definition, it is random and ideally the error between our prediction and actual data should be white noise, which would indicate the possibility of further improvements to the forecasted model. As an example, we could use gaussian distribution with a mean of zero and a standard deviation of one. The result would be that the majority of data points will be located near our predefined mean value zero which represents a white noise.

In the following section we look at the most commonly used approaches to model time series analysis with the combination of observed signal and error. Before that, we need to discuss about stationarity, because it can navigate us which forecasting models should be used. A time series is called stationary if it satisfies the following criteria.

- Mean is constant
- Standard deviation (also referred as volatility) is constant
- There is no seasonality

Difference between stationarity and white noise is that the white noise has a constant zero mean, so if it is satisfied then also stationarity is satisfied, but not vice versa. A non-stationary time series data produces unreliable and spurious results and leads to poor forecasting. One feature which violates the constant mean criteria is trend. It is slow, long-run evolution of preferences, technologies and demographics what is translated to always changing mean value. A non-stationary process with a deterministic trend becomes stationary after removing its trend.

Autoregressive model – AR(p)

The AR model specifies that the output variable y_t at some points *t* in time depends linearly on its own previous historical values y_{t-p} of order *p*. It can be only applied for stationary data with the following equation.

$$y_t = c + \sum_{t=1}^{p} \phi_i y_{t-i} + \varepsilon_t$$

Where c is constant, ε_t is white noise and *p* is the order of the process (i.e., how many lags to model). Lags are where results from one time period affect the following periods. ϕ is the corresponding coefficients for each lag up to order *p* which must meet the following criteria.

$$\phi_1 + \phi_2 + \cdots + \phi_k < 1$$

In section (TODO chapter name) we will discuss how to properly choose the order *p* to get the most accurate results. However, since autoregressive models base their predictions only on past information, they implicitly assume that the fundamental forces that influenced the past prices will not change over time, what is not always the case in real life.

Moving average model – MA(q)

In moving average, the difference from autoregressive model is that rather than using past values of the forecast variable in a regression, we will use past forecast errors.

$$y_t = \mu + \sum_{t=1}^{q} \phi_i \varepsilon_{t-i} + \varepsilon_t$$

μ is the mean of the series (also referred as expected value at y_t), ε_t is white noise error term at time *t* and ϕ is a coefficient multiplayer of that error. Fitting the MA estimation is more complicated than in AR model. An iterative solution is required because the lagged error terms are not observable. By using the lag operator, which operates on an element of a time series to produce the previous element, we can rewrite our formula into

$$y_t = \mu + \left(1 + \phi_1 B + \cdots + \phi_p B^p\right)\varepsilon_t$$

Moving average model should not be confused with moving average smoothing which simply states that the next observation is the mean of all past observations. It is used to smooth out short term fluctuations, remove seasonal patterns and highlight long term trends. It requires a specific window size called window width which represents the number of raw observations used to calculate the new series.

Autoregressive moving average model – ARMA(p, q)

It's natural to predict the future values of some equity based on its previous values adjusted with some errors. This is where we can use the combination of AR part, which involves regressing the variable on its own lagged values and MA part, modelling the error term as a linear combination of error terms in the past.

$$y_t = c + \sum_{t=1}^{p} \varphi_i y_{t-i} + \varepsilon_t + \sum_{t=1}^{q} \phi_i \varepsilon_{t-i}$$

In situations like the stock market, where we will try to predict the future value of the company's share, we also have to account trend to our computation. Using simple ARMA model will not be efficient, because it expects the time series to be stationary and as we mentioned before, one of its criteria is to have a constant mean.
When a stock follows a specific trend, its mean is either increasing or decreasing. If this is the case then we can use the ARIMA(q, d, p) model, where the letter *I* stands for integrated, and *d* is its order (i.e. number of transformations to make the time series stationary). Integrated means that instead of predicting the time series itself, we will transform our time series and calculate the difference of one timestamp to previous timestamp.

$$z_t = y_t - y_{t-1}$$

The usefulness of this transformation lies in the linearity (trend) of the original function. When we have an upward or downward trend and we take the difference of two side by side values, we expect the result to hover over some constant. This constant will be the new mean value for our transformed time series where we can use the ARMA model.

$$z_t = c + \sum_{t=1}^{d} \varphi_i z_{t-i} + \varepsilon_t + \sum_{t=1}^{d} \phi_i \varepsilon_{t-i}$$

To choose the right candidates for our ARMA model we need to determine what order of AR and MA terms should be used. To solve this problem, we introduce two functions called Autocorrelation function (ACF), which will answer the AR part, and Partial autocorrelation function (PACF) for MA. Though ACF and PACF do not directly dictate the order of the ARMA model, their plots can facilitate understanding the order and provide an idea of which model can be a good fit for the time-series data.

Autocorrelation function

Describes how well the present value of the series y_t is related with its past values until y_{t-k}. ACF considers all concepts like trend, seasonality and cyclic and find correlation between data points. The monitored dependency can be caused by correlation of other values between y_t and y_{t-k}.
In case of stationary time series, we express the ACF function as follows. [1]

$$\rho_k = \frac{cov\left(y_t, y_{t-k}\right)}{\sqrt{D\left(y_t\right)}\sqrt{D\left(y_{t-k}\right)}} = \frac{\gamma_k}{\gamma_0}$$

Where γ_k is an autocovaciance function defined as

$$\gamma_k = cov\left(y_t, y_{t-k}\right) = E\left(y_t - \mu\right)\left(y_{t-k} - \mu\right)$$

And it only depends on the time distance of two quantities in a stationary system then it applies that

$$D(y_t) = D(y_{t+k}) = y_0$$

Usually, the parameters μ, y_{t-k} and ρ_k for a stationary system are unknown. The estimation of the central value μ is a selection average

$$\bar{y} = \frac{1}{T}\sum_{t=1}^{T} y_t$$

Where T is a number of time series values and y_t is the observation value at time *t*.
Next, we estimate the variance y_k as a selected variance.

$$\hat{y}_{t-k} = \frac{1}{T}\sum_{t=k+1}^{T} (y_t - \bar{y})\ (y_{t-k} - \bar{y})$$

Lastly the estimation of coefficient autocorrelation $\hat{p}_k$ is defined as

$$\hat{p}_k = \frac{\sum_{t=k+1}^{T}(y_t - \bar{y})(\hat{y} - \bar{y})}{\sum_{t=1}^{T}(y_t - \bar{y})^2}$$

By the graphical representation of the ACF function we can pick only the statistically significant values and by counting them we will obtain the order of AR model.

Partial autocorrelation function

The key assumption behind ACF model is that a variance y_{t-1} captures all values older than itself and is able to explain the correlation between every single data point up until y_{t-k}. That means the end result in y_{t-1} is influenced by all previous correlations in y_{t-2}, y_{t-3} … y_{t-k} from time series.

But what if the assumption is not true and the result in y_{t-1} is not able to explain all the variances contained in y_{t-2} ? We would like to also capture the unexplained portion of y_{t-2} and feed it into y_t. So, in example of the MA model using ACF we could end up with the following formula

$$y_t = \mu + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \varepsilon_t$$

Unfortunately, by using simple ACF it is not possible, because y_{t-2} is already captured in y_{t-1} and by adding it again, it can drastically influence then end result.

Introducing Partial ACF, as the name suggest, by finding the correlation of y_t and y_{t-k}, we eliminate all intermediate values $y_{t-1}, y_{t-2} \ldots . y_{t-k+1}$, so if the value y_{t-k} had some unexpected behaviour, it will be directly captured in y_t without any influence. As an example, the first value of ACF and PACF is the same because there are no intermediate measurements, but the second lag in PACF will measures the correlation between y_t and y_{t-2}. Partial autocorrelation ϕ_{kk} is defined as

$$\phi_{k+1,k+1} = \frac{\rho_{k+1} - \sum_{j=1}^{k-1} \phi_k, \rho_{j,k+1-j}}{1 - \sum_{j=1}^{k-1} \phi_k, \rho_{j,j}}$$

$$\phi_{k+1,j} = \phi_{kj} - \phi_{k+1,k+1}\phi_{k,k+1-j} \text{ , for j = 1, 2, ... k}$$

TODO ARCH, GARCH to model volatility – forecast financial crisis + covid 19

Literatúra

[1] Bc. Milan MACHALEC, Modely časových radov s aplikáciami v ekonómii
[2] Bc. Dávid Slaninka, Analýza vývoja cien vybraných komodít a akcií na medzinárodných trhoch