Thesis title?

Thesis draft 0

Tristan McKechnie
18976697

January 28, 2022

Table of Contents

1 Introduction

The purpose of this mini-dissertation is to investigate and explore various feature engineering approaches for time series forecasting using machine learning. An important step in forecasting time series data is preprocessing the data with the objectives of reducing noise, increasing the signal-to-noise ratio, removing trends, and reducing the feature space. The methods which can be applied to achieve these objectives are known as featuring engineering approaches. All work is carried out in the specific context of financial time series data.

The remainder of this chapter is structured as follows: provide a background and motivation for the research, provide a problem description, define the scope and objectives of the research, discuss the research methodology used, clarify contributions to the field of data science and to provide an outline of the chapters in this mini-dissertation.

1.1 Background

This mini-dissertation is completed with guidance and direction from NMRQL Research, an investment management firm using and researching machine learning based approaches to construct optimal investment portfolios. The context of the work is therefore feature engineering of financial time series data. More specifically, the focus is aimed at stock market data.

There are many challenges around stock market forecasting. Common and repeatedly noted issues are, stock price data contains high noise, the time series are typically non-stationary, auto-correlation, the data is typically highly non-linear, and the vast quantity of data available for the markets results in complex high-dimensional problems. These challenges are discussed and used as the motivation for the research performed by Kim (2003), Tay and Cao (2001), Fischer and Krauss (2018), Guresen et al. (2011), and Hsieh et al. (2011).

The typical financial time series forecasting model pipeline requires feature engineering of the time series data, which is then provided to a model for training and forecasting. According to Geron (2019) feature engineering is the process of performing feature selection, feature extraction, and generating new features from existing ones. Feature selection refers to selecting a subset of most useful features for optimal model training and forecasting performance. Feature extraction is the process of combining existing features in different ways to select the most useful subset for optimal model training and forecasting performance.

1.2 Problem description

The specific problem this research aims to solve, or contribute good understanding toward, is which feature engineering approaches are useful in the context of financial time series forecasting. The feature engineering approaches should aim to improve model performance by overcoming some of the challenges of stock market forecasting mentioned earlier: noisy data, non-stationarity, and high-dimensionality.

As well as finding methods to over the challenges of stock market forecasting, the work also aims to identify under which conditions or rather, time series characterisations, the

different feature engineering approaches are useful. There are different types of time series forecasting problems: Uni-and multi-variate problems, where the number of input time series can be single (uni-) or multiple; One- or multi-step ahead, predicting one or multiple time steps into the future; The structure of time series data itself varies. Stationary data contains no trends over time and the descriptive statistics remains mainly unchanged. Non-stationary data can trends up or down over time, and the variance of the data may change through time too.

The purpose of investigating different feature engineering methods is to determine when the methods are suitable and useful, depending on the structure of the time series data. If possible, specifically identifying under which time series characteristics each method is applicable and useful. A method is seen as useful if it improves the forecasting performance of a model.

1.3 Scope and objectives

The scope of this research is limited to investigating time series feature engineering approaches, in the context of financial time series forecasting using machine learning methods. Different feature engineering approaches will be identified and investigated. The result of the various feature engineering approaches, the feature engineered training data set, will then be passed to a group of different models for training and forecasting. Through this approach the relative performance of the different methods will be evaluated based on forecasting performance.

The high-level forecasting problem pipeline is ~~something like this~~;

1. Retrieve raw time series data.
2. Feature engineer and transform into a supervised learning data set
3. Train machine learning model.
4. Evaluate forecasting performance.

The scope of this work is limited to investigating different feature engineering approaches only - ie part two is varied based on the feature engineering approach being investigated. The models used for feature engineering are kept constant. Their hyper-parameters are however optimised for each new feature engineering approach, but the same hyper-parameter tuning process is followed each time. The models used for forecasting are identified through a literature review of typical models for this specific problem. Different time series data is fed through this process as well. The reason for this is to testing feature engineering approaches on time series data with different characteristics.

The objectives of this work are summarised as:

1. Identify suitable feature engineering approaches commonly used in literature.
2. Implement these feature engineer approaches, describe the theory behind each method and compare their relative performance to a baseline case with no feature engineering.
3. Identify for under which time series characteristics the various feature engineering approaches are useful.

(1) I am a bit confused about this paragraph. I expect a listing of the types of time series problems. It seems that this is what you start to do, but then you start defining things. Also, the sentences are not full sentences.

(2) What is predicted? What is the target feature?

(3) Avoid pronouns such as "it", "them", "they"

1.4 Research Methodology

The research methodology of this mini-dissertation is outline in this section. The work is broken into X parts: research and review, implementation and empirical investigation and finally conclusion and summary of results.

The first part of the work, research and review, is focused on literature review of relevant literature, followed by research of the methods identified. The literature review specifically covers financial time series forecasting using machine learning. The purpose here is to identify common modelling approaches, feature engineering methods, models used and typically expected forecasting performance. Once this is covered and different feature engineering methods and relevant machine learning models are identified, then specific research into the theory behind these methods and models is completed. This is to gain understanding of the methods, in order to perform effective implementation.

The second part of the mini-dissertation covers the investigation into the various feature engineering approaches. Here the various feature engineering approaches are implemented on different time series data sets to investigate their benefits and under what conditions they can be used. Four models are tested with each feature engineering approach. The models have their https://www.overleaf.com/project/6000ab53dca4161cb2864576 hyperparameters optimised for each different feature engineering approach, but their architecture is kept constant. Before a feature engineering approach is implemented, a high level theoretical overview is provided to describe how the methods works. Feature engineering approaches are grouped into chapters according to their similarities.

The final part of the mini-dissertation is a summary of the finds, which feature engineering approaches work well, and under which conditions they work or don't work.

During this work, wherever possible, unless otherwise stated, all algorithms and models investigated are implemented using open source libraries. Wherever such a library is used, the original authors will be cited. The purpose of this work is not the specifics of implementing the various models and algorithms, but rather investigating their potential benefits or drawback. For this reason, as far as possible, open source libraries are used.

1.5 Contributions to data science

The work in this mini-dissertation contribute to the understanding of different feature engineering approaches for financial time series forecasting using machine learning. Specifically when the different approaches are beneficial. This work can hopefully serve has a useful guide in the future into feature engineering for the specific problem of financial time series prediction.

1.6 Mini-dissertation outline

Provide an overview of each chapter was the contents of each chapter is known.

2 Literature Review

In this chapter we summarise the findings of a literature study around the topics of financial time series forecasting using machine learning. This section starts by reviewing the top 10 most cited articles on the Scopus database under the search: financial time series forecasting.

2.1 Financial time series forecasting with machine learning

In this section we review the most cited articles from Scopus under the search "financial time series forecasting". All articles have more than 100 citations. The purpose of this review is to identify common modelling approaches applied in time series forecasting. Special attention is given to the data preparation / feature engineering approaches employed. Although the search did not specify machine learning as a specific forecasting technique, all top cited articles implemented some machine learning models or paradigms.

Kim (2003) authored a highly cited article perform forecasting of a stock price index. In the paper the a support vector machine and neural network are implemented for the regression task. The purpose of the work is to study the performance of the SVM for the specific forecasting task. The author motivates of the study by stating that SVMs may be more suitable to problem because they have fewer hyperparameters to choose, and are not as susceptible to noise and their training finds a global optimum (rather than a local optimum for neural networks).

The specific forecasting problem set up by Kim (2003) is the daily price change direction of the Korea composite stock price index. The data used for the forecasting problem are technical indicators. Technical indicators are metrics derived from the stock price data. The data set is made up of 2928 days of price data. 80% was used from training and 20% for testing. The data was scaled into a range of [-1.0,1.0]. Twelve technical indicators were used as input features: %K, %D, slow %D, momentum, price rate-of-change, William's %R, A/D oscillator, disparity5, disparity10, OSCP, CCI and relative strength index. Interestingly, the author does not make use of many lag parameters. The technical indicators are calculated from the original data, and used as the input features.

Kim (2003) determined that the Gaussian radial basis functions kernel was most suitable to the problem. The author performed a parametric study of the regularization parameter, with values from 10 to 100 tested. The final results showed that the SVM outperformed the neural network by 3-6%.

Tay and Cao authored a couple of highly cited papers exploring the use of SVMs for financial time series prediction (Tay and Cao, 2001; Cao and Tay, 2003).

In the first paper, Tay and Cao (2001) compare the performance of an SVM against a neural network. The authors develop a problem of forecasting futures contracts. The motivation of the work is to test the performance against neural networks for financial time series forecasting, and to determine suitable SVM hyperparameters. The models are testings on 5 different data sets, the daily close price of ; Standard&Poor 500 stock index futures, United States 30-year government bond, United States 10-year government bond, German 10-year government bond, and French government stock index futures (MATIF-

(4) Section 2.1 reviews the ..... Spectral methods for noise reduction are discussed in Section 2.2. ....Give rest of chapter outline.

(5) Should you not first define time series and discuss what is meant by financial time series forecasting? And just a short overview of machine learning?

(6) This section reviews...

(7) "Time series feature engineering"?

(8) Machine learning is not a specific technique, but a collective name to refer to different techniques.

(9) SVM does classification, not regression.
For regression ⇒ SVR
Are you considering the problem as a regression problem, i.e. predict next value of stock, or as classification problem to predict an action, e.g. sell, buy, hold.

(10) New sentence. Avoid long sentences.

(11) I suggest that these financial concepts be discussed as background in a separate section/chapter. As part of point (5). All important concepts to be first separately described/defined.

(12) Do not start a sentence with a number.

(13) Wrt point (11), all the used ML concepts to also first be described.

CAC40).

Tay and Cao (2001) performs some basic feature engineering on the data. Four relative difference features are calculated, and one 15 day exponential moving average feature. A , 10, 15 and 20 day relative difference feature is calculated. The author motivates the differencing features use as the transformation is said to give the data a more normal distribution. Another motivation is to remove trend from the data. The moving average feature is kept to preserve information lost through the differencing transformation. The target variable for the problems are a five day ahead prediction of the relative difference in price. However the target variables are also smoothed using a three-day exponential moving average. Outliers for the input features ±2 std deviations are capped. All input data is also normalized to a range of [-0.9,0.9]. The data is split into a testing, validation (for hyperparameter tuning) and testing subsets. The datasets have a total length of 1307. The models prediction performance is measured using: the normalized mean squared error (NMSE), mean absolute error (MAE), directional symmetry (DS) and weighted directional symmetry (WDS). Tay and Cao (2001) use a Gaussian kernel for the SVM. Across all datasets the SVM models outperformed the neural networks, for all evaluation metrics.

Cao and Tay (2003) wrote a second paper about SVMs for financial time series forecasting. This paper was similar to the first. In the second paper the authors compare the SVM to a multilayer perceptron and radial basis neural network. The most signifcant difference between the papers are that the model are trained and evaluated using a walk-forward method.

Cao et al. (2003) build on the modelling approach using SVMs developed by Tay and Cao but investigate the use of principal component analysis (PCA), kernel principal component analysis (KPCA) and independent component analysis (ICA) for feature extraction prior to model training. These feature extraction / selection methods are used to transform the original input features to a lower dimension set. The authors apply the various feature selection methods, followed by SVM modelling, to the same data used by Tay and Cao, futures contracts prices for difference indices. The original input features, prior to feature extraction, are a number of relative differences, differencing using exponential moving averages and three technical indicators. The output variables is the relative difference for five days ahead. Outliers are again capped and all data scaled to a range of [-0.9,0.9]. All three feature extraction methods yielded better results than simply using the original input dataset. KPCA performed best, followed by ICA. However the computational run time was significantly longer for KPCA.

Kaastra and Boyd (1996) investigate the use of neural networks for financial and economic time series forecasting. The focus of their paper was to produce a guide for design neural networks for the specific application. The authors motivate their work due to the large number of hyperparameters and network topology than must be chosen to design effective neural networks. The results of the authors work is an eight-step neural network design process, specifically from financial and economic time series forecasting. On the topic of data used to the authors state that technical and fundamental indicators, and time differencing are typical input features for the forecasting problem. The authors also suggest use related, intermarket data could be useful. For data processing / feature engineering the authors suggest performing first differencing and log transforms. To remove trend