UNIVERSITY OF TARTU

Faculty of Social Sciences

School of Economics and Business administration

Roza Mikeliani

Nino Kavlashvili

# Evaluation and Comparison of Machine Learning and Classical Econometric AR Model on Financial Time Series Data

Master's thesis

Supervisor: Mustafa Hakan Eratalay

Tartu 2020

Name and signature of supervisor . . . . . .

Allowed for defence on . . . . . (date)

I have written this master's thesis independently. All viewpoints of other authors, literary sources and data from elsewhere used for writing this paper have been referenced

. . . . . (signature of authors)

## Acknowledgements

We would have never written this thesis together if it had not been for the genuine curiosity and enthusiasm, we had found from the very first sight of the topic and our supervisor, Mustafa Hakan Eratalay. We stay wholeheartedly grateful from the first meeting with him, when our efforts were united, until the last remarks and support throughout the path. We would like to thank him for believing in our work even when we lacked the motivation.

We are thankful to Jaan Masso and Ricardo Alfredo Mendes Pereira Vicente for valuable comments which kept us in the right direction. Special gratitude goes to Luca Alfieri for providing needed thorough feedback and possibilities of further development in the pre-defence; the thesis has improved significantly after considering his suggestions.

Many sincere thanks to the Republic of Estonia for supporting our master's studies in the University of Tartu.

While there are many who deserve our gratitude in this journey, we would like to express our appreciation to our families and friends in Georgia and in Estonia, who have been always there for us.

## Abstract

This paper examines the effects of time series data behaviour on the predictive performance of classical econometric univariate autoregressive and machine learning autoregressive models. The research aims to understand which forecasting approach would perform better in extreme scenarios. Even though some empirical studies demonstrate the superiority of machine learning methods relative to classical econometric methods, it is still arguable under what conditions one method can be constantly better than the other. And if there are any cases when econometric models are preferable than machine learning. Data is derived from simulation, ensuring the presence of different outlier and error distributions in small and relatively larger samples. The simulation results show that the machine learning approach outperforms econometric models in most of the cases. However, the existence of outliers worsens the performance of machine learning on small datasets. Even with the presence of outliers, as the sample size grows, the result improves so much for machine learning that it dominates the econometric model. The same models were used to forecast with rolling sample approaches on real financial data.

**Keywords**: univariate time series forecasting, comparative analysis, machine learning, econometric model, simulation, outliers, misspecification, RMSE, MAE

## Contents

# 1. Introduction

Machine Learning (ML) techniques have gained considerable popularity in recent decades. As the demand for the data analysis continues to grow, the comparative performance of ML and classical econometric models has become even more relevant than ever. There have been great advancements in computational capabilities of machine learning algorithms, some of them are specifically designed to perform forecasting on time series data since it turned out to be a challenge due to specification of it. Over hundreds of papers propose new ML algorithms based on advanced methodology and accuracy improvements. Because the number of machine learning models is rapidly increasing, the efforts have been made on comparing them in terms of performance against each other, rather than against statistical models.

Machine learning and Econometrics differ in purpose, focus and methodologies; however, they tend to solve the same problem (Yong et al. 2017). Lately, there have been discussions about the importance of integrating machine learning and econometrics to bring more value to the field.

Noteworthy, that prior to the algorithms in machine learning, the classical econometric models were applied to serve the purpose of forecasting time series data. The reality unfolded that these two approaches of forecasting developed in parallel, while they have a similar purpose (Charpentier, Flachaire, and Ly 2018).

Because of the difference in technical approach, the research has also followed the two methods, existing in parallel simultaneously until recently, after the realization that a huge number of papers accumulated about the time series forecasting both in econometrics and machine learning separately. This may bring a bit of uncertainty in deciding which approach is best to use when forecasting on time series data; It became clear that common ground is needed to consider both approaches and compare the ways of solving the same problem.

Therefore, in recent years number of comparative analyses were conducted between the two approaches of forecasting. Nevertheless, most of the comparative analyses were usually focused on overall performance metrics, obtained from empirical studies. In these cases, it happens quite often that generalizations cannot be made, because the underlying data or the methods differ greatly. Research has shown that the answer is not ubiquitous as different empirical studies have proved contradictory results, one cannot decisively state when or if any of the methodology is consistently better than the other.

Additionally, the overall performance metric from a specific scenario or the dataset can change as the data behaviour changes, so the conclusions cannot be constantly true even within the same approach. Meaning, if a single comparative analysis showed the superiority of one model above another, this might not be true for different data specifications, for example, when the distribution of outliers or errors is changed. This aspect of comparative analysis has been recently addressed, when few empirical studies observed that the sample size matters in the predictive performance of ML models (Cerqueira, Torgo, and Soares 2019).

This paper aims to fill the gap and extend the existing comparative analysis by including different scenarios and aspects of financial time series data behaviour. The contribution of this study will be the possibility of generalization of obtained results since simulations of two approaches will be performed on the same predefined synthetic data. There are four main questions and respective simulations covered.

The first question of this study is the effects of changing the sample size. The question focuses on understanding whether and how the performance of statistical and econometric models change as the sample size grows from 200 to 3000 observations when errors are normally distributed and there are no outliers.

The second question is to understand what happens when the distribution of outlier changes. There are two data samples, with outliers evenly and unevenly distributed, while everything else is kept fixed in both datasets.

In the third question, the objective is to compare the predictive performance of both approaches, when the parameter is reaching the stationarity border. So, there are two cases when the parameter is changed, with the sample size and everything else kept the same.

The fourth question is about the change in the distribution of errors. Like the second question, the error distribution is changed, while everything is the same. There will be three cases considered: 1) when the errors are normally distributed; 2) The errors are drawn from a highly positively skewed distribution, for simplicity. The implications would be similar to the negatively skewed distribution case. 3) And finally, when the errors are unevenly distributed, resulting in the fat-tailed distribution. The case when the true distribution is symmetric but fat-tailed.

The last part of the study tries to gather observations from all experiments, and systematically review to understand if any general conclusion could be drawn.

The first section of the paper is the introduction, which will be followed by the literature review in the II section. The literature review will be summarized within the scope of the simulation study, it will not consider general literature of comparative analysis of models outside of the scope of this paper. The III section will focus on data simulation, specification and methodology of simulation processes. The IV section will be the estimation of models using different statistical and ML algorithms on each specific case. This section will demonstrate the results and compare the performance of methods. In section V the findings will be concluded, and the generalizations made. The appendix and additional tables or Material will be attached to the bottom. Lastly, all the references will be listed at the end.

## 2. Literature Review

Nowadays, with the excess of data and a variety of methods that could be applied for prediction purposes, it is important to strictly define the task and be able to apply the most efficient forecasting methods. To accomplish this task, one needs to be informed about which models perform better for what kind of problem and underlying data. With the increased popularity and interest in the machine learning approach and attempts of performance improvement, the choice among machine learning models is even bigger than ever. This works like an overload of information, the more the information, the harder it gets to find what one is looking for. There is a need for empirical studies and comparative analysis to bring some clarity and valid guidelines in the proper choices and usage of technological advancements.

Time series forecasting itself has been a classic research topic for econometrics, starting from simple methods such as NAÏVE moving to more complex approaches like ARIMA (Box et al. 2015) and ETS (Hyndman et al. 2008). For reference, we should outline that ML approaches originate from statistical methods. Pioneering simple ML algorithms dates to 1950. More about origins are discussed in the paper (Hastie, Tibshirani, and Friedman 2001). Since then the machine learning approaches have been widely used to tackle problems of time series prediction.

Besides the model development and analysis an enormous effort must be made to empirically validate the performance of different models. Therefore, providing insight into the strong and weak points of available algorithms is essential, all of this results in immersing the value of the research around the topic.

Even though machine learning as such, was actively developed only for some decades, there is impressive research in the model performance and comparative analysis of different machine learning models as well as a comparative analysis of machine learning algorithms against the classical econometric models. However, the research has not fully covered the comparative predictive performance of the above-mentioned two approaches in different scenarios of data behaviour. Because of huge developments in machine learning, resulting in an increasing number of new algorithms, studies were mostly focused on catching up with the advancements made in this regard. The concept of machine learning is very broad, it is concisely summarized by (Athey 2018), emphasizing that *„ML literature does not frame itself as solving estimation problems*", emphasizing that even if the traditional and ML methods both deal with the forecasting - predictive task, they differ conceptually, and therefore, in reality, they face different problems. Machine learning performs prediction using the identified patterns in the dataset, while the traditional forecasting methods have the parameter estimation, explaining the relationship between the y and x. Because of this and some other differences stated in the same paper, it gets quite complicated to understand or make an informed choice between the two options.

Due to the vast literature and broader perspective around the topic, it is necessary to narrow down the scope and only consider research papers which are highly relevant to the aims of this study. Therefore, in this part we will present papers focused on comparative analysis on univariate time series forecasting between the traditional econometric AR and simple autoregressive machine learning model.

Looking in a timeline view, the table below is a short summarization of the most relevant papers to our study. The earliest paper comparing econometric and ML approaches in terms of performance, was by (Hill, O'Connor, and Remus 1996), according to which the MLP machine learning model outperformed the statistical models.

**Table 1** - Research Papers on Comparative Performance of Econometric and ML approaches

| **Comparative performance of Econometric and ML approaches** | |
|---|---|
| (Hill, O'Connor, and Remus 1996) | MLP outperformed statistical methods. |
| (S. B. Kotsiantis, Kanellopoulos, and Pintelas 2006) | Big sample size is necessary for ML performance, this can be a challenge, some ML algorithms may not run at all. |
| (Ahmed et al. 2010) | Examined comparative performance for 8 major ML models - MLP and Gaussian processes outperform the rest. |

| | |
|---|---|
| (Pritzsche 2015) | ARIMA based models are competitive to Machine Learning models for investigated time series forecasting situations. |
| (Makridakis, Spiliotis, and Assimakopoulos 2018) | Traditional (econometric) methods have been dominated by ML approaches. Nevertheless, most of the statistical methods systematically outperform some of the ML methods for univariate time series forecasting. |
| (Cerqueira, Torgo, and Soares 2019) | Stated that the results of (Makridakis, 2018) were biased due to the sample size. ML methods improve their relative predictive performance as the sample size grows. |

Later, (Kotsiantis, Kanellopoulos, and Pintelas 2006) noticed that a big sample size would be necessary to guarantee the superior performance of ML, however, they also mentioned that running algorithms on big samples could be a challenge and some of the algorithms might not even run. This brought forward the importance of data specifications on top of the algorithms and logic embedded. The advantages and drawbacks of big samples were discussed from the machine learning perspective.

Later, (Ahmed et al. 2010) aimed to examine comparative performance for 8 major ML models, for this task using time series from the M3 competition. The ML models compared are following: The ML models compared are following: multilayer perceptron (MLP), Bayesian neural networks, generalized regression neural networks (GRNN), radial basis functions (RBF), K-nearest neighbor regression (KNN), support vector regression (SVR), CART regression trees, Gaussian processes (GP).

The paper concludes that MLP and Gaussian processes outperform the rest.

Ahmed's (2010) findings have been further explored in (Makridakis, Spiliotis, and Assimakopoulos 2018) for the same M3 competition and 1045 time series with additional eight traditional statistical methods including ARIMA, exponential smoothing, naive, and theta, among others. Surprisingly, some contradictory results were obtained stating that CART and RBF have the best results. But the overall conclusion remains that traditional methods have been dominated by ML approaches for all forecasting horizons examined. Nevertheless, their results suggest that most of the statistical methods systematically outperform some of the ML methods for univariate time series forecasting. Therefore, they concluded the paper stating that the reasons why ML algorithms fail to outperform classical econometric methods for univariate time series forecasting need to be further investigated.

Following the topic, this question was addressed later by (Cerqueira, Torgo, and Soares 2019), in which it is stated that they believe the results of Makridakis were biased due to the sample size. They commented that samples used to draw those conclusions had "*average, minimum, and the maximum number of observations of 118, 66, and 144, respectively.*" The authors also claim that ML methods improve their relative predictive performance as the sample size grows. In their empirical analysis, they used 90 univariate time series from different domains of application, within that setup the results have shown that conclusions by (Makridakis et al. 2018) are only valid with the small sample size.

As per these papers, one of the observed limitations of ML models is that algorithms are unable to perform well with the limited data input, which is considered as a disadvantage of ML methods. Alternatively, the same be an advantage of the traditional methods since they still perform better with minimal data available.

As discussed in the paper by (S. B. Kotsiantis, Kanellopoulos, and Pintelas 2006) big sample size being a necessity for ML performance, can also be a challenge: '*In addition, when a data set is too huge, it may not be possible to run an ML algorithm. In this case, instance selection reduces data and enables the ML algorithm to function and work effectively with huge data.'*

Furthermore, the computational requirements of ML methods are significantly higher since superior performance would require huge data, while the same or somewhat acceptable accuracy, in cases of data limitations, might be achieved using the statistical methods on a much smaller dataset.

On this note, (Cerqueira et al. 2019) leaves a remark that even with large amounts of data, it is not obvious that the machine learning method would always outperform an econometric method. Referencing back to (Wolpert 1996) who states that the learning algorithm cannot be appropriate in all the scenarios according to „*No Free Lunch theorem* ".

This leaves an open question for the research – Is it possible that ML models could not outperform classical econometric methods with the big sample size? If yes then what are the conditions, other than the sample size, which could explain the relatively poor performance of ML algorithms against statistical methods?! This is the gap we will try to concentrate on during the study.

As shown by (Pritzsche 2015) ARIMA based models are competitive to machine learning models for the investigated classical time series forecasting situations, meaning in cases with zero

exogenous covariation. However, these traditional univariate techniques lack a few key requirements for complex predictive tasks. This again stresses the importance of determining the necessity of using either of the methods based on the complexity of the task. In order to be able to make an informed choice, a relevant study should be available considering different scenarios.

Based on the available literature, we believe that the presence of outliers and its implications in ML algorithm performance needs to be studied thoroughly. For example, the problem of misspecification is important because we face the prerequisite of ML model, which is to remove or replace the outliers, affects its forecasting performance, whilst this practice is not common in traditional econometric approaches.

(Li et al. 2015) state that outlier detection and removal affected the variance of the training data and therefore test accuracy has been significantly increased by 13 percent. On the other hand, the results provided by (Maniruzzaman et al. 2018) show that replacing the missing values and outliers for ML model (in this case random forest) by group median values yields an accuracy of 92,26% and AUC of 0.93. Therefore, we aim to explore if this will be true for our simulated data and extreme case scenarios, with outliers distributed evenly or unevenly. On the econometrics side (Hendry and Santos 2005) it seems beneficial to include dummies in a model when the data suggests so; besides including dummies when they do not change anything seems relatively harmless, although there is a small efficiency loss risk. However, if there are outliers and they are ignored, the coefficient estimates will become biased.

The latest discussion presents the relevance of questioning data preparation in machine learning. For many practitioners it is natural to perform exploratory tasks which might be followed by changes in the dataset, considered as the data preparation stage of the forecasting. However, it does not always result in the best performance in the light of the research by (Dingli and Sant Fournier 2017). According to these authors '*methods or models, that best fitted available data, did not necessarily result in more accurate post sample predictions (a common belief until then).*' While another paper addressing the topic of data preparation for ML models published 2 years earlier, suggests specific Trend Deterministic Data Preparation Layer. (Patel et al. 2015). Proposing to convert each of the indicator's continuous value to discrete, which then has been used to predict fluctuations of prices in time series. This data preparation layer proved to increase prediction accuracy for 3 ML models (SVM, random forest and naive-Bayes (Multivariate Bernoulli Process) out of 4 models tested. However, the accuracy of the model ANN was even reduced slightly after applying the above-mentioned data preparation layer. Therefore, again, there is no unique recipe

whether one should use a data preparation step or not. It needs to be rather discussed and reviewed case by case and model by model.

It is also important to emphasize that the latter paper (Patel et al. 2015) suggested a specific data preparation layer, which is not the same as **data preprocessing** examined in the paper published by Makridakis in 2018. In that paper the author clarified that the original data might be changed trying to achieve either one or all of the three goals:

1. **Transforming** (power transformation is applied to the original data to achieve stationarity in the variance)
2. **Deseasonalizing**
3. **Detrending**

While trying to find out which type of the change in original data would result in the best accuracy or the performance, the papers tested different combinations of mentioned changes.

Since there are different problems with time series data, each issue needs to be handled separately using different approaches and methods. For example, detecting outliers is such a complex problem that some practitioners might even disregard them, meaning remove the outliers in order to achieve the balanced dataset. Detecting outliers has been an interesting and challenging topic for data mining enthusiasts who have been trying to detect outliers in large datasets by a distance-based calculation yet in 1996 (Knox and Ng 1996). The problem of outliers was addressed as "*event change detection*" in the data mining community (Guralnik and Srivastava 1999),(Ralanamahatana et al. 2005).

A relatively recent paper on this topic by (Takeuchi and Yamanishi 2006) showed that change point detection and outliers are directly related, which was not explicitly demonstrated by related work around this topic earlier. They also criticized existing approaches for being computationally expensive and suggested a new "*two-stage time series learning scheme*". One of its features is being the learning process which is repeated twice, the outlier detection happens in the first stage and change point detection is done using the learned model from the second stage. As it seems from the description, handling the outliers is not a straightforward task. The authors have contributed largely to this area, since now there exists a computationally better performing approach which they named "*ChangeFinder*". They have listed ideas for further analysis and research. It is clear that machine learning methods must somehow incorporate the issue of outliers instead of neglecting them completely.

Choosing a performance metric is yet another topic of discussion, which will not be covered in detail in this paper. Even though we decided not to enter the topic of performance measurement, it is still very important to mention here that even if we would try to rank and analyze existing models by so far commonly used accuracy metric, there are still cases when it would mislead general objective of financial forecasting. The accuracy is evaluating whether the direction change was correctly forecasted, regardless of the profit/price values.

"*While accuracy might be a good approximation of an algorithm's general ability, it technically does not convey any information on profitability. Taking an extreme example, an algorithm with high accuracy might correctly forecast many comparably insignificant profit opportunities while missing a small number of large profit opportunities*. "(Ryll and Seidens 2019). The performance metrics are to be chosen case by case and data by data, it was shown that specific metrics are more appropriate to be used when working on yearly time series data for example.

A comprehensive survey on evaluating machine learning performance in financial market forecasting by authors Ryll and Seidens (2019), who analyzed over 150 papers, classified result metrics in three main categories: Error-based, Return-based and Accuracy-based. In conclusion, they rejected the parametric approach due to the heterogeneity of the literature sample they have covered. Still, the statistics within their sample proved that accuracy is the most popular metric, followed by Root Mean Squared Error **(**RMSE), (Ryll and Seidens 2019). Despite all the uncertainty, specific metrics have become the common choice of many papers, RMSE is usually one of the commonly used performance measurements among the practitioners and researchers.

For a reference, a very recent paper by (Bou-Hamad and Jamali 2020) used Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE) as base performance metrics within the context of a simulation study of time series financial forecasting. This proves that the well-established RMSE and MAE are relevant for financial time series forecasting purposes.

Additionally, the mentioned paper by (Makridakis et al. 2018) concludes that it is wrong to believe that forecasting methods can be considered of superior accuracy, simply because of their mathematical elegance or sophistication, but the accuracy should be rather transparently and empirically proven in an indisputable manner. That is what was supposedly lacking when it comes to ML methods or effective ways to use them. Therefore, even though there are high expectations and demand towards the accuracy of ML methods, the empirical studies or rather their

comparability against benchmarks would require the data to be made available along with the articles for those who want to replicate results.

Another issue not yet studied thoroughly is the dataset balancing and splitting, which is even more complex when talking about financial time series data. According to (Dingli and Fournier 2017), having balanced and adequately split training and test datasets is crucial in achieving unbiased models. The same authors discuss that shuffling and randomly splitting the dataset might end up with populating the '*future instances*' in the training set and '*past instances*' in the test set. Provided that each record would be treated as separate, it would be still a more suitable approach to consider the nature of time series data already at the stage of splitting the original data into training and test sets.

When talking about machine learning and time series forecasting, it is always in the best interest of prediction to provide training samples which have an equal amount of examples for each class, meaning if predicting price moves up or down, it must be ensured that the training dataset is not biased towards any of the two classes (Dingli and Fournier 2017). This leads to other issues with ML methods, trying to achieve a balanced dataset. Imbalanced data can be misleading for the learning algorithm, even though in the real world there are plenty of cases when one needs to work with the imbalanced data. To ensure the reliability of ML models, some measures have been suggested. This judgment applies also to the presence of outliers distributed unevenly.

(S. Kotsiantis, Kanellopoulos, and Pintelas 2006) reviewed options of handling imbalanced datasets. The options were on the data level, algorithm level, or the mixture of both. The authors concluded that imbalanced datasets lead to improper classification performance since the „*In small imbalanced datasets the minority class is poorly represented by an excessively reduced number of examples that might not be sufficient for learning*", This effect is reduced for larger samples.

What is not covered in this paper, are the other properties of data. Some of the characteristics might not be evenly distributed, which can lead to different traits in train and test sets. For example, the outlier, or the error term distribution. The literature does not cover these aspects in great detail so far.

All in all, it ends up with the clarity that the machine learning performance is largely defined by the sample size, the relevant pre-processing approaches to achieve a balanced dataset which again is more promising with the large datasets.

For this reason, it needs to be ensured that there is sufficient data at all, which then can be split in a well-structured manner, achieving a balanced training dataset.

So, how much data would one need to perform reliable predictions using machine learning methods? With a small dataset, is it reasonable to still use ML over statistical methods? What other conditions could hinder the performance of machine learning algorithms?

Obviously, there is no clear answer to this question nor the literature that would give the best guideline for it, since it depends on various factors, such as the complexity of the problem or the complexity of the learning algorithm (Cerqueira et al. 2019).

Summarizing the literature available around the topic of our interest, the reality unfolded that there is not sufficient or aligned research available. Especially little research has been made around the univariate time series forecasting methods. For more advanced machine learning approaches, conflicting results were obtained from different empirical studies. The reasons behind haven't been discovered in most cases, since it requires considerations of different data behaviour and the presence of extreme scenarios. Most of the authors conclude their papers with a discussion about the relevance of further research in the remaining areas they found a reason to be studied. While some of them still challenge ML models stating that statistical econometric models might be still able to outperform sophisticated ML methods in specific cases. With our efforts, we will try to pick the presence of outliers and other specifications, to run the simulations and observe what conclusions can be made. This would be the novelty proposed by this paper.

## 3. Methods and Data

Provided the objective of the paper, in order to answer questions in very specific scenarios of data behaviour, it is needed to ensure desired characteristics of time series data exist. To create relevant scenarios, data simulations were performed.

Before simulating data for the study, real financial data characteristics and specifications have been considered. Thus, making sure the simulations can be as close to reality as possible. On the other hand, we do not intend to limit ourselves by taking a single-time data series and draw conclusions based on the empirical analysis only. Therefore, the simulation of data intends to bring

the advantage of independence and flexibility in adjusting properties and achieving predefined extreme scenarios.

The data is simulated from an autoregressive model of order 1, namely AR(1), which is given by the equation below:

$$r_t = \alpha + \beta r_{t-1} + \varepsilon_t \quad (1)$$

where $\beta$ is called the autoregressive parameter and $\varepsilon_t$ is a white noise process with variance $\sigma^2$.

The idea is to study the effect of different scenarios and misspecifications on the performance of the maximum likelihood estimation method and machine learning.

After running the simulations, the results specifying the parameters of different datasets were created and attached in Appendix B. The combination of parameters changed gives a unique dataset. The parameters adjusted to create scenarios are as follows:

- sample size (200, 500, 1000, 3000)
- outliers (no, evenly distributed, unevenly distributed)
- beta (0.5, 0.99)
- distribution of error terms (normal, skewed normal, Students' t)

The structure of the table is reused to briefly describe the specifications of each dataset and to link them to the scenario question, with the Case ID.

Because the characteristics and parameters of data are largely defined by the question addressed, specific tables were linked to the relevant question. A detailed overview of the simulated datasets is attached in Appendix A.

Research questions are directly linked to specific datasets to incorporate desired scenarios. Four main questions are corresponding to 14 datasets. The first four datasets have a vanilla case, which refers to AR(1) with the simplest scenario (no outliers) when there are no complications such as beta close to stationarity, outliers having skew, or fat-tailed distribution. These four samples of vanilla cases changing only in the sample size (200, 500, 1000, and 3000). In addition, for the rest of the datasets the data size will be changing only from 200 to 3000. These scenarios and datasets

will be used to answer the question of whether the performance would improve by increasing the size of the dataset.

The next four datasets are designed for evenly and unevenly distributed outliers. Evenly distributed means that the outlier may occur in both the training and testing sets. Unevenly distributed means that the training set does not have outliers, while the testing set has. There are two sample sizes for each case. The smallest with 200 observations and relatively larger with 3000 observations. With these datasets we aim to answer the question: how does the change in the outlier distribution affect prediction accuracy of ML algorithm (AR) and classic econometric model AR(1), cases when outliers are evenly and unevenly distributed?

The consecutive two datasets are generated to check what happens in the dataset when the $\beta$ parameter is reaching the stationarity border. For this reason we have a case, when the data is simulated via AR(1) with $\alpha = 0.008, \beta = 0.99, \sigma^2 = 0.011$. These parameter values yield very similar unconditional mean and variance as the vanilla case. The datasets sizes considered are with numbers of observations of 200 and 3000.

The last datasets are simulated to check the effect of different error distributions on the prediction performance of different models. Specifically, when the error distribution is normal, skewed normal and Student's t. For each of the cases discussed, we consider sample sizes 200 and 3000. Unless skewed or fat-tailed, the errors are drawn from normal distribution assuming Gaussian errors. When skewness is considered, the errors are drawn from a skew-normal distribution. Skew normal errors are generated from Hansen's skew t-distribution (Hansen 1994). The degrees of freedom 300 (that makes the t-distribution very close to the normal) and lambda=0.9. This results in a skewness value being around 0.7. And when the fat tail is considered, the errors are drawn from a Students' t distribution with 5 degrees of freedom.

In short, we question if the machine learning algorithms are drastically affected by changes in sample size, since the empirical analysis (Cerqueira et al. 2019) have drawn some conclusions based on their studies, this time we would be able to generalize the results for these specific simulated scenarios.

Our questions addressed hypothetical problems of ML models especially in small samples, from misspecification of distribution, from approaching to stationarity border, from skewness and fat

tails of the errors and the outliers. We seek to find situations where the maximum likelihood estimation approach produces better forecasts.

## 3.1 Econometrics

Moving to the analysis part. A little bit on maximum likelihood estimation (MLE): it is a method used for estimating the parameters of a probability distribution by maximizing a likelihood function so that for the current model observed data has the highest probability. In addition, a specific function that maximizes the value is called the maximum likelihood estimate.

In order to estimate the performance of the econometric model Gaussian loglikelihood is used, which needs to be maximized with respect to the model parameter.

The Gaussian loglikelihood is presented below:

$$\mathcal{LL} = -\frac{T}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{t=1}^{T}(r_t - \mu_t)^2 \quad (2)$$

where $\mu_t$ is a conditional mean, expressed by an equation: $\mu_t = \alpha + \beta * r_{t-1}$.

For each of the cases, we simulate 1000 replications. We then estimate the model with the mentioned method and calculate the out-of-sample forecasts for the next 5 periods. Two error-based performance evaluation metrics will be used for the comparative analysis of models. As Root Mean Squared Error (RMSE) is commonly used for time series forecasting evaluation, it will also be used in this study in addition to the Mean Absolute Error (MAE).

After predictions, we compare the forecast accuracies via the root mean squared forecast error and mean absolute forecast error metrics. In the paper by (Ryll and Seidens 2019), evaluation metrics of financial market forecasting were explored. The study focused on ML methods is listed 150 papers, out of which 62 papers used RMSE, MAE, or both as the evaluation metric. The paper explains that while accuracy and return metrics are important and popular from a machine learning perspective, the results should be benchmarked against „*an ideal classifier*" so that the comparative performance can be provided.

In addition to RMSE, we considered MAE especially because it is robust to outliers (Hansen and Lunde 2005) and we will have scenarios where outliers will be present. The difference is that in

case of MAE all errors, big and small are treated equally, whilst RMSE penalizes large errors more due to the squared term.

RMSE is defined as:

$$RMSE = \sqrt{\frac{1}{B}\frac{1}{5}\sum_{b=1}^{B} \quad \sum_{t=T+1}^{T+5} \quad \left(\hat{r}_{t,b,i} - r_{t,b,i}\right)^2} \tag{3}$$

On the other hand, MAE is defined as:

$$MAE = \frac{1}{B}\frac{1}{5}\sum_{b=1}^{B} \quad \sum_{t=T+1}^{T+5} \quad \left|\hat{r}_{t,b,i} - r_{t,b,i}\right| \tag{4}$$

where B=1000 (number of simulations) and the forecast horizon is 5.

Noteworthy, that the machine learning models use a classical econometric approach to tackle time series forecasting problems with auto-regressive tasks, that is based on the AR(p) model (Cerqueira et al. 2019).

„*According to the AR(p) model, the value of a given time series can be estimated using a linear combination of the p past observations, together with an error term and a constant term* „(Box et al. 2015). The formulation of it for our specific model, was presented above as equation (1).

## 3.2 Machine Learning

As for the machine learning approach, the baseline equation for ML algorithms is the same autoregression used for the AR model. This technique is generally applied for time series forecasting where input variables are taken as observations at a previous timestamp, also called lag variables.

The input dataset for the ML model is the same as for econometrics. One approach is calculating linear dependence manually. An alternative could be to use a built-in autoregression model in statsmodels library, that automatically selects an appropriate lag value using statistical tests and trains a linear regression model, which is provided in the AR class. Since we are interested in AR (1) specifically, the first step is creating the AR model.

Generally, when evaluating a model for time series forecasting, we need to make sure that the same data was not used for training and testing. In ML, this is called out-of-sample data. We can achieve this by splitting up the dataset. As a general approach cannot be used for time series because there is a relationship between the observations and each observation is dependent on previous ones, we must split data and respect the temporal order in which values were observed. After that calling fit() function in order to train it on our dataset. This step returns an ARResult object.

The next step will be performing walk-forward validation that involves moving along the time series one-time-step at a time. Where we define the window based on model fitting parameters. Shortly, the whole process cycle is the following: starting at the beginning of the time series, the minimum number of samples in the window is used to train a model. Then the model predicts the next time step, which is stored or evaluated against the known value. The window is expanded to include the known value and the process is repeated. Additionally, because a sliding or expanding window is used to train a model, this method is also referred to as rolling window analysis or a rolling forecast.

The whole cycle is iterated 1000 times as the number of replications made in the dataset. The loop implemented in the code outputs a thousand rows for estimates for the model: RMSE and MAE. The average of these coefficients (each individually) is calculated to make the comparison of ML and econometrics models feasible. So, RMSE is calculated taking the inputs of the test dataset and predictions made by the machine learning model. The same goes for MAE. Both are defined the same way as for econometric models. Given in the equations (3) and (4). The lower the grade of RMSE and MAE the better the model performs.

## 3.3 Empirical application

The next stage after the simulation would be an application with real data. The empirical study is performed on the S&P 500 historical prices from Yahoo Finance (Anon 2020). The historical daily, weekly or monthly data can be downloaded from the source.

The initial data is for 12 years and consists of 3020 observations from April 14th, 2008 till April 9th, 2020. The dataset has several columns we will be using adjusted close prices for our analysis. And the timeline looks as follows:


**Figure 1.** Time series of S&P 500 historical prices over 12 years. *Note: data retrieved from Yahoo Finance (Anon 2020).*

The overall trend was increasing with a slight fall in 2008 starting from October. One can also easily see a drastic drop in prices due to the Corona situation. The distribution of the data looks in Figure 2 is shown below:

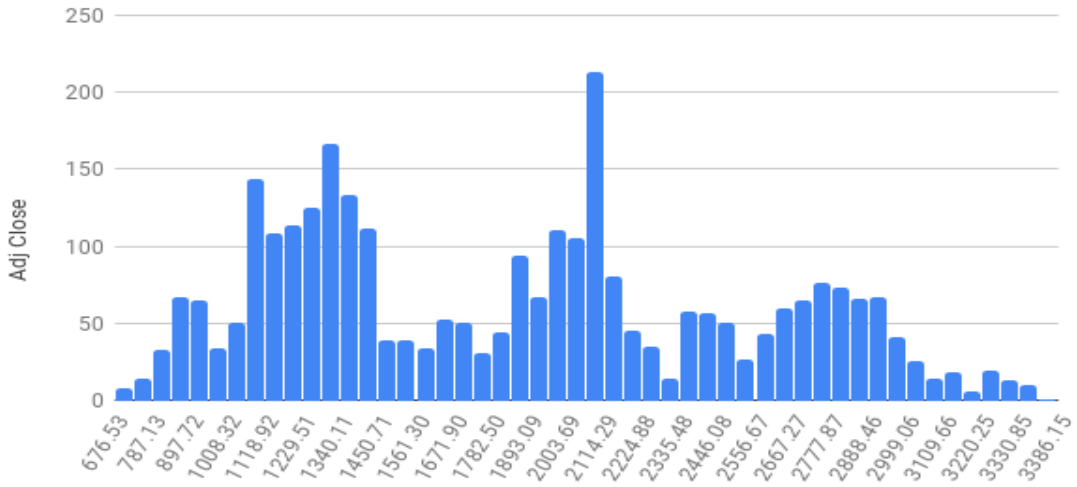


**Figure 2.** Real data distribution. *Note: data retrieved from Yahoo Finance (Anon 2020)*

Simple descriptive statistics for the dataset is presented shortly:

**Table 2 -** Descriptive Statistics of Empirical Data

| Descriptive Stats | Count | Min | 25% | 50% | 75% | Max | Mean | Median | Std.Dev |
|---|---|---|---|---|---|---|---|---|---|
| Adj Close | 3020.000 | 676.530 | 1282.670 | 1867.065 | 2367.520 | 3386.150 | 1853.861 | 1867.065 | 657.834 |

The lowest price in the dataset was 676.53. The highest value reached 3386.15, while the mean - 1853.86 was also very close to the median value of 1867.065. Notice that the standard deviation is smaller relative to the mean, which indicates that more of the data is clustered about the mean.

Rolling window approach will be used for prediction in both econometric and ML cases to ensure the stability of the model over time. The approach for econometrics is similar as discussed in the paper (Nakajima 2012). The idea is to transform the dataset using a rolling sample methodology also described in the machine learning section above. We will be using two different approaches called sliding and expanding windows.

For the sliding approach we will start by defining the lowest and highest indexes, so from t=1 to T=500 observations. After that we make predictions for the next 5 days. Then comparing the results with real data. The new cycle starts with updating indexes and repeating the same for the values from t=6 to T=505. Please notice that the dataset size is always 500. Therefore, it is a fixed approach.

The similar procedure with the increasing the sample size will be applied for expanding window cases. Meaning that the dataset will start from t=1, but the ending point is updated by 5 days: T=500, T=505, T=510. The interval is 5 days always because it corresponds to trading days in a week. So, the sample size is growing with every step.

For performance evaluation metrics the same RMSE and MAE will be applied, just like in the case of the simulation analysis.

**Discussion on how the data fits to achieve the aim of the thesis**

This study aims to specify conditions when econometric or machine learning would perform better by running simulations for predefined scenarios on a synthetic dataset. Simulation analysis is performed for every question specified.

Because the aim is to observe and generalize the findings to the questions, the specific actual dataset would not serve this purpose, since it would be just another empirical study, possibly

conflicting with or supporting some other similar studies. Therefore, to pursue our goal - be able to see the broader picture and draw the conclusion regarding the relative performance of econometric and machine learning algorithms, simulations are used.

In financial econometrics 500 or 1000 replication is a common practice in simulation studies. This practice is followed in this study as well. Simulated data is introduced, and 1000 replications are performed to ensure the validity of the conclusions and results.

# 4. Results and Analysis

## 4.1 Simulation Study

As described in the methods, we present the results for each question based on the datasets simulated incorporating specific properties of the data. We generated simple comparative visualizations and tables for a detailed overview of each question.

In all the visualizations presented below, econometric and ML models are depicted in blue and grey colours, respectively. The vertical axis represents the RMSE score and the horizontal varies across the different scenarios.

### 4.1.1 Results - Question 1

The first question is based on the vanilla case, where the sample size grows from 200 to 3000 gradually, the comparative performance of two methods of this study are depicted. As the RMSE for the ML model, coloured in grey, is lower than that of econometric model AR(1) at every step, the first question can be answered with just a glance at the graph above, meaning ML model outperformed econometric model in each case of this question.

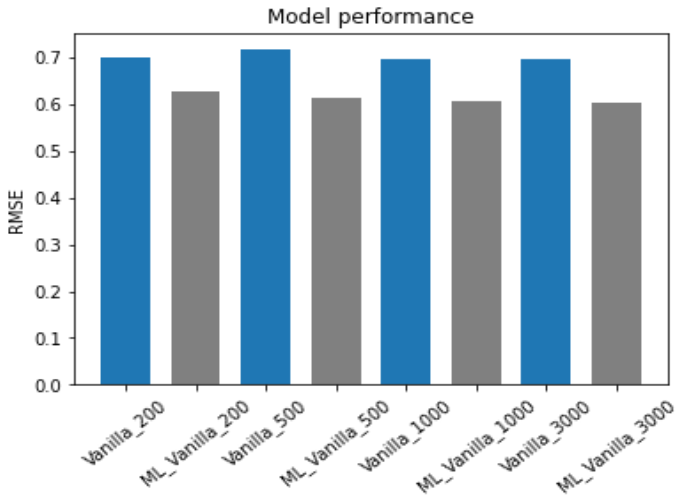


**Figure 3.** The overall performance of econometric and ML models based on RMSE. Case 1

Additionally, to the main answer, further observations can be made. Table 3 presents the RMSE and MAE values for both models and all sample size cases.

**Table 3** - Performance measurements and differences between the two models. Case 1

| Question/Performance Metric | Sample Size | RMSE | | Diff | MAE | | Diff |
|---|---|---|---|---|---|---|---|
| | | AR (1) | ML | | AR (1) | ML | |
| Performance across different sample sizes | 200 | 0.6998 | 0.6260 | -11.8% | 0.5568 | 0.5261 | -5.8% |
| | 500 | 0.7166 | 0.6137 | -16.8% | 0.5670 | 0.5160 | -9.9% |
| | 1000 | 0.6977 | 0.6081 | -14.7% | 0.5538 | 0.5077 | -9.1% |
| | 3000 | 0.6962 | 0.6042 | -15.2% | 0.5567 | 0.5040 | -10.5% |

It is straightforward, that the ML model has lower error values than the econometric model. However, the difference between the measurement values of these models is not increasing at every step. We would expect that the performance improves for both models, and it improves much faster for the ML model and therefore the difference between the two would also be somewhat consistent. But this is not the case.

It is also to be mentioned that RMSE shows higher error values than MAE, still, according to both measurement metrics, the ML model is performing better. According to RMSE values, the percentage difference between the models' metrics can be as high as 16.8%, while for MAE the largest difference between the two models' performance metrics amounts to 10.5%.

The table shows that the RMSE and MAE values decreased as the sample size grew for the ML model at every step, this is not the pattern for the AR(1) model. Instead, metric value fluctuated as

the sample size grew, meaning model performance does not directly depend on the sample size. Therefore, we can see based on our results that the statement - performance improves as the sample size grows - applies only to the ML model and not to the econometric model.

The differences between the same model's measurement values of different sample sizes are calculated and presented below in table 4.

**Table 4** - Performance measurements and the differences within the same model. Case 1

| Question/Performance Metric | Sample Size | RMSE | | | | MAE | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | AR (1) | Diff | ML | Diff | AR (1) | Diff | ML | Diff |
| Performance across different sample sizes | 200 | 0.6998 | | 0.6260 | | 0.5568 | | 0.5261 | |
| | 500 | 0.7166 | -2.3% | 0.6137 | 2.0% | 0.5670 | -1.8% | 0.5160 | 2.0% |
| | 1000 | 0.6977 | 2.7% | 0.6081 | 0.9% | 0.5538 | 2.4% | 0.5077 | 1.6% |
| | 3000 | 0.6962 | 0.2% | 0.6042 | 0.6% | 0.5567 | -0.5% | 0.5040 | 0.7% |

If comparing the two measurements for the AR(1) model, one can see that they stop being consistent, since in the case where the sample size consists of 3000 observations, according to the RMSE the performance changed from the previous step by 0.2%, where the number of observations was 1000. The value of RMSE dropped from 0.6977 to 0.6962, but the change in MAE was -0.5% increasing to the absolute value of 0.5567.

The overall answer is that the size matters for predicting capabilities in both approaches, but ML has a clear tendency that the performance improves as the sample size grows.

The poor results of machine learning algorithms in smaller datasets were a well-known fact by practitioners and were also indicated in some research papers. However, this property was not checked in combination with other specifications. We can conclude that when errors are normally distributed and there are no outliers, ML performance will increase as the sample size grows. The reason behind this is that ML needs a larger amount of training set to learn the patterns of data behaviour and its trends. Even though in this case size did not change the conclusion about model superiority, it decreased the error in the forecasting for ML. On the other hand, the sample size does not have a direct effect on maximum likelihood estimations as they seem to fluctuate as the size increases, and it is true for both RMSE and MAE estimates.

### 4.1.2 Results - Question 2

In this question, we are trying to see the impact of outliers in the dataset.

We estimated both AR (1) and ML model performance on 4 datasets with small and large sample sizes and outliers being evenly and unevenly distributed. So, there are different scenarios to be checked. The graph below shows that the smallest RMSE is achieved when the sample size is 3000 and the outliers are evenly distributed. In all the cases when the sample size is 200, the RMSE is high, but it is the highest when in addition to the small dataset, the outliers are unevenly distributed.

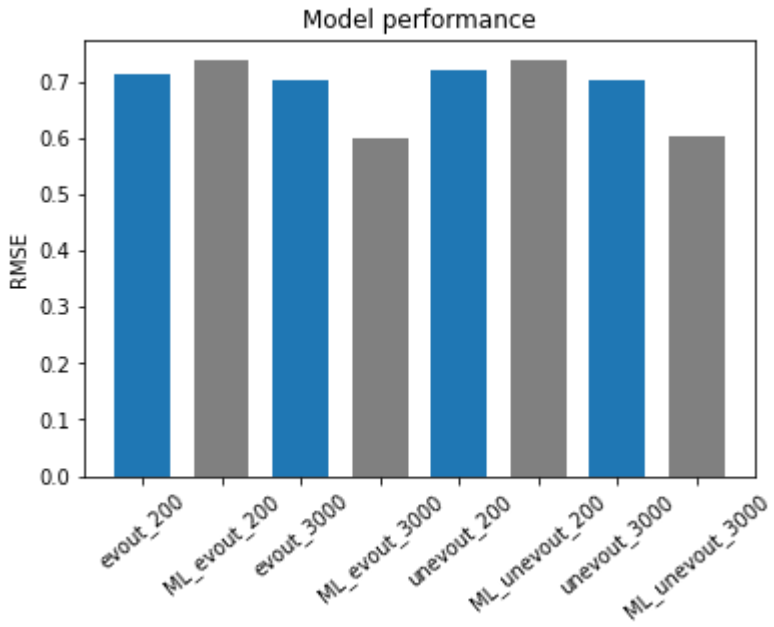


**Figure 4.** The overall performance of econometric and ML models based on RMSE. Case 2

Table 5 indicates that in the presence of outliers, the econometric model can perform better when the dataset is small.

**Table 5** - Performance measurements and the differences between the AR ad ML models. Case 2

| Question/Performance Metric | Sample Size | RMSE | | Diff - models | MAE | | Diff - models |
|---|---|---|---|---|---|---|---|
| | | AR (1) | ML | | AR (1) | ML | |
| 2.a - Outliers evenly distributed | 200 | 0.7156 | 0.7378 | 3.0% | 0.5696 | 0.6022 | 5.4% |
| | 3000 | 0.7050 | 0.6010 | -17.3% | 0.5647 | 0.5019 | -12.5% |
| 2.b - Outliers unevenly distributed | 200 | 0.7220 | 0.7378 | 2.1% | 0.5730 | 0.6022 | 4.8% |
| | 3000 | 0.7050 | 0.6045 | -16.6% | 0.5647 | 0.5048 | -11.9% |

So, when there are two specifications, meaning sample size and outliers, none of the models are consistently better than the other. It seems that both models improve performance when the sample size grows, but ML improves with a larger drop in the error.

When outliers are evenly distributed and the sample size is 200, the RMSE for the AR (1) model is smaller by 3% than the same of ML. When the sample size increases to 3000 observations, the performance of both models improve, but the reduction in RMSE of the ML model is more significant and the difference between the two models reaches 17,3%. The same logic applies to MAE values, reduction in errors as the sample size grows is more significant for machine learning, even with the unevenly distributed outliers. The improvement for ML is so big that when the sample size changes overall picture changes drastically. So, AR (1) is no longer better even if the outliers are present when the sample size is big enough.

Additional details about within the model differences of evaluation metrics are included in table 6.

**Table 6** - Performance measurements and the differences between the same models. Case 2

| Question/Performance Metric | Sample Size | RMSE | | | | MAE | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | AR (1) | Diff | ML | Diff | AR (1) | Diff | ML | Diff |
| 2.a - Outliers evenly distributed | 200 | 0.7156 | | 0.7378 | | 0.5696 | | 0.6022 | |
| | 3000 | 0.7050 | -1.5% | 0.6010 | -18.5% | 0.5647 | -0.9% | 0.5019 | -20.0% |
| 2.b - Outliers unevenly distributed | 200 | 0.7220 | | 0.7378 | | 0.5730 | | 0.6022 | |
| | 3000 | 0.7050 | -2.4% | 0.6045 | -22.1% | 0.5647 | -1.5% | 0.5048 | -19.3% |

Looking at the differences between the RMSEs within the same models as the size grows, it is noticeable that AR (1) error only changed by -2,4%, while the ML model error changed by -22.1%.

These details are brought forward to emphasize that the predictive capability increases drastically for ML and when the sample size is small the difference is also very small, such that it could be discarded. It would be interesting to check further sample sizes between 200 and 3000 observations and check when the machine learning approach starts to perform better.

This brings our discussion to the relevance of questioning how to choose the best model in case of different situations. Based on the results our suggestion is to use econometric models in smaller datasets with outliers nevertheless of their distribution. However, if the dataset is large enough then a machine learning approach would be performing better even with the outliers. This answer however cannot be generalized, since we haven't checked different amounts of outliers, it might

be that if the number of outliers is increasing as the sample size grows, then the ML would still perform worse than the econometric model. The impact of the number of outliers on the predictive performance of models will be a great possibility to further expand this study.

### 4.1.3 Results - Question 3

While questioning the performance when the parameter is reaching the stationarity border, we observed ML performing drastically better than AR (1) in both small and relatively large dataset cases.

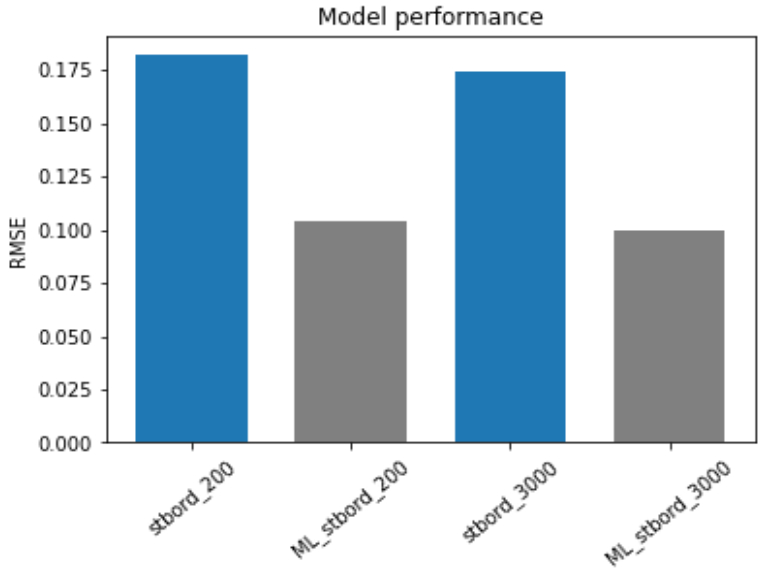


**Figure 5.** The overall performance of econometric and ML models based on RMSE. Case 3

However, it can be observed that the ML performance does not improve significantly as the sample size grows, unlike the case with outliers. Table 7 gives a detailed overview of the error metrics.

**Table 7** - Performance measurements and the differences with stationary data. Case 3

| Question /Performance Metric | Sample Size | RMSE | | | | Diff - models | MAE | | | | Diff - models |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | AR (1) | Diff | ML | Diff | | AR (1) | Diff | ML | Diff | |
| 3 - Parameter reaches stationarity border | 200 | 0.1819 | | 0.1041 | | -74.7% | 0.1401 | | 0.0876 | | -59.9% |
| | 3000 | 0.1747 | -4.1% | 0.1001 | -4.0% | -74.5% | 0.1355 | -3.4% | 0.0834 | -5.0% | -62.5% |

The difference between the performance metrics for the two methods are almost the same for small and large datasets. Based on these results, the stationarity does not hinder ML performance,

however, a bigger dataset does not guarantee a significant improvement in the performance of any of the models as the change in RMSE for AR (1) was 4.1% and for ML 4.0%. This is only interesting in the regard that the sample size change caused a slightly bigger outcome for AR (1), which was not the case when working with outliers. Even if the size effects were minor for each model, comparison between the models makes it clear that machine learning outperformed AR (1) according to both metrics and in both sample size cases. The largest difference, 74,7% of performance according to RMSE was when the sample size was 200, while in MAE results, the largest difference equalled 62.5% for the sample size being 3000.

### 4.1.4 Results - Question 4

To answer the question about the error distribution and its effects on predictive performance, three error distributions were considered: normal, like in the vanilla case, skewed and Student's t, while for each of these cases the datasets tested consisted of 200 and 3000 observations. From the visualization below, the ML model is consistently performing with lower RMSE values for both small and large datasets.

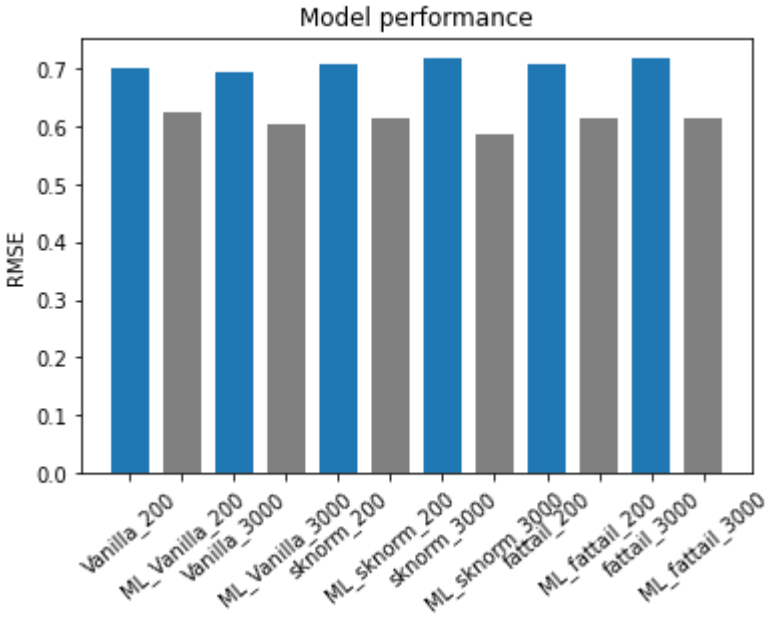


**Figure 6.** The overall performance of econometric and ML models based on RMSE. Case 4

The specific measurements for all these cases are presented in table 8. What's noteworthy here is the fact that as the sample size grows the performance improves for the ML model, but it also consistently worsens for the AR (1) model.

**Table 8** - Performance measurements and differences – error distributions. Case 4

| Question/Performance Metric | Sample Size | RMSE | | | | Diff - models | MAE | | | | Diff - models |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | AR (1) | Diff | ML | Diff | | AR (1) | Diff | ML | Diff | |
| 4.a - Normal errors distribution | 200 | 0.700 | | 0.626 | | -11.8% | 0.557 | | 0.526 | | -5.8% |
| | 3000 | 0.717 | 2.3% | 0.614 | -2.0% | -16.8% | 0.567 | 1.8% | 0.516 | -2.0% | -9.9% |
| 4.b - Skewed error distribution | 200 | 0.707 | | 0.615 | | -15.0% | 0.566 | | 0.518 | | -9.1% |
| | 3000 | 0.718 | 1.6% | 0.587 | -4.6% | -22.3% | 0.567 | 0.3% | 0.497 | -4.3% | -14.1% |
| 4.c Fat-tailed error distribution | 200 | 0.708 | | 0.616 | | -15.0% | 0.530 | | 0.498 | | -6.6% |
| | 3000 | 0.719 | 1.5% | 0.578 | -6.5% | -24.3% | 0.535 | 0.9% | 0.467 | -6.6% | -14.6% |

In earlier cases we did not spot any specific tendency for AR (1) related to the sample size, as the RMSE and MAE have fluctuated across different sample sizes. So, it seems the parameter being close to the stationarity border impacts on AR (1) performance more than it does on the ML model.

It is straightforward that all the percentage changes in error metrics of AR (1) model are positive, meaning the error values increase, while the same changes for ML are all negative constantly reducing the error values.

The difference between the RMSE values of AR (1) and ML changes from 11.8% to 24.3%, the larger differences spotted when the sample size is 3000 and the errors are not normally distributed.

### 4.1.5 Results - Question 5

Table 9 below is a summarization of our simulation analysis for all cases. With the columns for RMSE and MAE the performance of each model is evaluated for small and large sample sizes. The results estimated by RMSE and MAE are tightly interwoven and aligned, meaning in none of the cases are they contradictory. We can immediately notice that in the great majority of the cases ML has performed at smaller RMSE and MAE metrics compared to the AR (1) model. Therefore, in most cases machine learning is outperforming classical econometric approach for maximum likelihood estimations. But there are several exceptions to this rule, which turned into interesting findings.

**Table 9** - Overall performance measurements and the differences within the same model.

| Question/Performance Metric | Sample Size | RMSE | | Diff - models | MAE | | Diff - models |
|---|---|---|---|---|---|---|---|
| | | AR (1) | ML | | AR (1) | ML | |
| 1 - Performance across different sample sizes | 200 | 0.6998 | 0.626 | -11.8% | 0.5568 | 0.5261 | -5.8% |
| | 500 | 0.7166 | 0.6137 | -16.8% | 0.567 | 0.516 | -9.9% |
| | 1000 | 0.6977 | 0.6081 | -14.7% | 0.5538 | 0.5077 | -9.1% |
| | 3000 | 0.6962 | 0.6042 | -15.2% | 0.5567 | 0.504 | -10.5% |
| 2.a - Outliers evenly distributed | 200 | 0.7156 | 0.7378 | 3.0% | 0.5696 | 0.6022 | 5.4% |
| | 3000 | 0.705 | 0.601 | -17.3% | 0.5647 | 0.5019 | -12.5% |
| 2.b - Outliers unevenly distributed | 200 | 0.722 | 0.7378 | 2.1% | 0.573 | 0.6022 | 4.8% |
| | 3000 | 0.705 | 0.6045 | -16.6% | 0.5647 | 0.5048 | -11.9% |
| 3 - Parameter reaches stationarity border | 200 | 0.1819 | 0.1041 | -74.7% | 0.1401 | 0.0876 | -59.9% |
| | 3000 | 0.1747 | 0.1001 | -74.5% | 0.1355 | 0.0834 | -62.5% |
| 4.a - Normal errors distribution | 200 | 0.6998 | 0.626 | -11.8% | 0.5568 | 0.5261 | -5.8% |
| | 3000 | 0.7166 | 0.6137 | -16.8% | 0.567 | 0.516 | -9.9% |
| 4.b - Skewed error distribution | 200 | 0.7068 | 0.6147 | -15.0% | 0.5655 | 0.5183 | -9.1% |
| | 3000 | 0.7182 | 0.5874 | -22.3% | 0.5673 | 0.497 | -14.1% |
| 4.c Fat-tailed error distribution | 200 | 0.7081 | 0.6158 | -15.0% | 0.5304 | 0.4975 | -6.6% |
| | 3000 | 0.7188 | 0.5781 | -24.3% | 0.5352 | 0.4669 | -14.6% |

As for descriptive statistics: the RMSE range is between 0.1001 to 0.7188, whilst MAE is between the value range of 0.0834 and 0.6022. Out of 14 datasets, 85.71 % of cases ML approach was better than the former. The biggest difference between the models' performance according to MAE was 62.5% when the parameter reaches the stationarity border and the sample size is 3000. This is the case when the ML performs at its best with the MAE of 0.0834, so this could be considered when choosing between the models.

Given the methodology and using the synthetic data with all its replications, we believe that outcome is matching what we have expected. It is consistent with the recent literature, where several papers pressed the importance of sample size for ML methods. Based on our results, we can agree with (Cerqueira, Torgo, and Soares 2019) stating that as the sample size grows the performance of ML improves, this statement applies to all the scenarios of the simulation. We supposed that it is most likely that with the small dataset, classical econometric models would perform better, however, in our simulation study, we could even see an interesting pattern. Even with the small sample size, there are overall more cases when ML is performing better than the econometric model AR (1). Out of 7 scenarios, where the sample size consists of 200 observations, in 5 cases ML still performed at a lower RMSE and MAE values than AR (1). There are only two scenarios in which AR (1) outperformed the ML model and both of those scenarios have the outliers evenly or unevenly distributed, in addition to the small sample size.

This brings us to a realization that the econometric model would perform better than ML when the two conditions are met:

1. The sample size is small
2. There are outliers in the dataset

The rest of the cases when the errors have different types of distribution - normal, skewed, or fat-tailed, or when the parameter is reaching the stationarity border, ML again performs better than the econometric model, regardless of the sample size.

### 4.2 Empirical Study

As discussed in methodology, the empirical study was performed using the S&P 500 historical prices. It is important to emphasize that the data includes the recent economic impact of COVID-19 pandemic, therefore the real data behaviour in these exceptional times is interesting.

As per the result, we observe machine learning outperforms econometric model in both sliding and expanding window approaches as in most cases for the simulation scenarios. We did have the same sample size 3020 observations from the real dataset.

Having the data from April 2008 up until January 2009, represented yet another economic crisis, followed by the increasing trend across the last few years from the crisis of 2009, culminating in a sudden drop due to the ongoing pandemic period of 2020. We are facing a situation, where the specifications we were looking into became secondary since recent major events caused the dramatic change in the data behaviour.

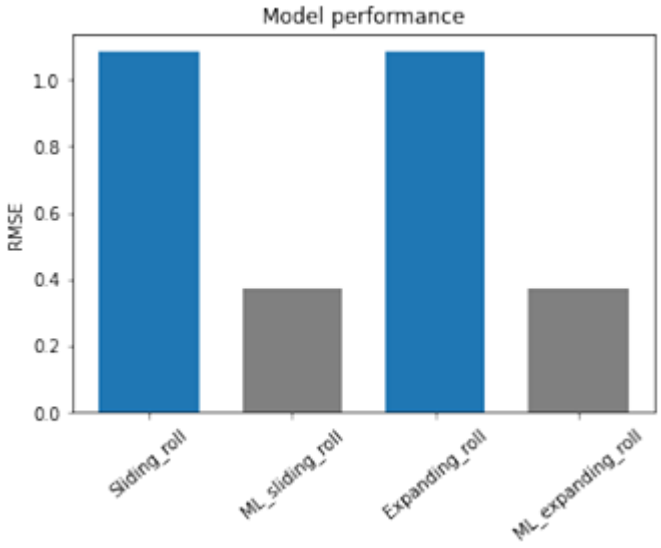


**Figure 7** - Empirical Results

Indeed, for empirical study as well machine learning approach is better based on both RMSE and MAE estimates. When looking at the sliding window case RMSE of ML being 0.3743 that is way lower than it was for AR. Additionally, MAE being 0.2355 versus 0.6823. On the other hand, a quite similar situation is for the expanding window with the lowest RMSE 0.3722 overall. Nevertheless, the difference between the approaches seems to be insignificant when drawing the overall conclusion. Although in the case of ML, there seems to be a preference for expanding window approach as the errors are decreasing in that case. Whilst for econometrics it seems to have contradictory results in RMSE and MAE estimations, as based on the MAE sliding window seems to be better and for RMSE expanding window is slightly better.

**Table 10** - Performance measurements and the differences within the same model for empirical data

| Question/Performance Metric | | RMSE | | | MAE | | |
|---|---|---|---|---|---|---|---|
| | Sample size | AR (1) | ML | Diff - models | AR (1) | ML | Diff - models |
| Sliding rolling window | 3020 | 1.0854 | 0.374 | -190% | 0.6823 | 0.2355 | -190% |
| Expanding rolling window | 3020 | 1.0852 | 0.372 | -192% | 0.6835 | 0.2209 | -209% |

Compared to simulation study there seems to be a drastic difference between the models, as machine learning is performing way better than the former one. Although we cannot claim that for each empirical time series forecasting for financial data, one will conclude the same as we did. But we think that it is important to see that our study, divided into 2 major parts, strongly support each

other. And based on our results in the empirical analysis as well as in majority cases for simulation study machine learning is outperforming based on both estimation methods.

## 5. Conclusions and Discussion

Time series forecasting was performed in parallel by an econometric AR (1) model using maximum likelihood estimation and machine learning method using an autoregressive algorithm. For simulation study five main questions with relevant predefined scenarios were discussed with sample size changing from 200 to 3000. In addition, an empirical analysis was performed on S&P 500 data on 3020 observations.

The main findings are that the ML model performance improves gradually as the sample size grows based on the simulations. The econometric model doesn't always perform better than ML when the sample is relatively small. The only condition when the econometric model outperformed the ML prediction, was when there were outliers in the dataset, in addition to the sample being small. When errors are normally distributed, in the vanilla case, as the sample size grows, the performance of the econometric model fluctuates, while that of the ML model increases for all the scenarios examined. When the parameter is reaching the stationarity border, the improvement of accuracy in the ML model is insignificant as the sample size grows. Still in this case prediction performance of the ML model is drastically better than the same in the econometric model.

The main practical implication of this study would be that informed choices can be made between models based on the specifications of data. As for scientific implications this paper contributes to fill the gap in comparative analysis. Specifically, when the dataset is mis-specified or in the presence of outliers.

Yet it is unclear why the econometric model’s performance does not improve and even worsens when the sample size grows. The limitation of the paper is that we have only considered a fixed set of outliers. We believe that there is still room for extending study in that direction. It will, therefore, be reasonable to check whether the number of outliers could affect the performance of models. What happens when the number of outliers increases as the dataset grows?! Additionally, to be questioned at what sample size does the ML start to outperform the econometric model.

## References


Ahmed, Nesreen K., Amir F. Atiya, Neamat El Gayar, and Hisham El-Shishiny. 2010. 'An Empirical Comparison of Machine Learning Models for Time Series Forecasting'. *Econometric Reviews* 29(5–6):594–621.

Bou-Hamad, Imad, and Ibrahim Jamali. 2020. 'Forecasting Financial Time-Series Using Data Mining Models: A Simulation Study'. *Research in International Business and Finance* 51:101072.

Box, George E. P., Gwilym M. Jenkins, Gregory C. Reinsel, and Greta M. Ljung. 2015. *Time Series Analysis: Forecasting and Control*. John Wiley & Sons.

Cerqueira, Vitor, Luis Torgo, and Carlos Soares. 2019. 'Machine Learning vs Statistical Methods for Time Series Forecasting: Size Matters'. *ArXiv:1909.13316 [Cs, Stat]*.

Dingli, Alexiei, and Karl Sant Fournier. 2017. 'Financial Time Series Forecasting - A Machine Learning Approach'. *Machine Learning and Applications: An International Journal* 4(1/2/3):11–27.

Dingli, Alexiei, and Karl Sant Fournier. 2017. 'Financial Time Series Forecasting – A Deep Learning Approach'. *International Journal of Machine Learning and Computing* 7(5):118–22.

Hastie, Trevor, Robert Tibshirani, and Jerome Friedman. 2001. 'Springer Series in Statistics'. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*.

Hill, Tim, Marcus O'Connor, and William Remus. 1996. 'Neural Network Models for Time Series Forecasts'. *Management Science* 42(7):1082–92.

Hyndman, Rob, Anne B. Koehler, J. Keith Ord, and Ralph D. Snyder. 2008. *Forecasting with Exponential Smoothing: The State Space Approach*. Springer Science & Business Media.

Kotsiantis, S. B., D. Kanellopoulos, and P. E. Pintelas. 2006. 'Data Preprocessing for Supervised Learning'. 1(1):7.

Kotsiantis, Sotiris, Dimitris Kanellopoulos, and Panayiotis Pintelas. 2006. 'Handling Imbalanced Datasets: A Review'. 12.

Li, Weizhi, Weirong Mo, Xu Zhang, John J. Squiers, Yang Lu, Eric W. Sellke, Wensheng Fan, J. Michael DiMaio, and Jeffrey E. Thatcher. 2015. 'Outlier Detection and Removal Improves Accuracy of Machine Learning Approach to Multispectral Burn Diagnostic Imaging'. *Journal of Biomedical Optics* 20(12):121305.

Makridakis, Spyros, Evangelos Spiliotis, and Vassilios Assimakopoulos. 2018. 'Statistical and Machine Learning Forecasting Methods: Concerns and Ways Forward' edited by A. R.

Hernandez Montoya. *PLOS ONE* 13(3):e0194889.

Ord, Keith. 2020. 'Data Adjustments, Overfitting and Representativeness'. *International Journal of Forecasting* 36(1):195–96.

Patel, Jigar, Sahil Shah, Priyank Thakkar, and K. Kotecha. 2015. 'Predicting Stock and Stock Price Index Movement Using Trend Deterministic Data Preparation and Machine Learning Techniques'. *Expert Systems with Applications* 42(1):259–68.

Pritzsche, Uwe. 2015. 'Benchmarking of Classical and Machine-Learning Algorithms (with Special Emphasis on Bagging and Boosting Approaches) for Time Series Forecasting'.

Ryll, Lukas, and Sebastian Seidens. 2019. 'Evaluating the Performance of Machine Learning Algorithms in Financial Market Forecasting: A Comprehensive Survey'. ArXiv:1906.07786 [q-Fin].

Athey, Susan. 2018. 'The Impact of Machine Learning on Economics'. 31.

Cerqueira, Vitor, Luis Torgo, and Carlos Soares. 2019. 'Machine Learning vs Statistical Methods for Time Series Forecasting: Size Matters'. ArXiv:1909.13316 [Cs, Stat].

Yong, Tan, Eric Zheng, Ramnath Chellapa, Michael Shaw, Olivia Sheng, and Alok Gupta. 2017. 'When Econometrics Meets Machine Learning'. Data and Information Management 9.

Hill, Tim, Marcus O'Connor, and William Remus. 1996. 'Neural Network Models for Time Series Forecasts'. Management Science 42(7):1082–92.

Kotsiantis, S. B., D. Kanellopoulos, and P. E. Pintelas. 2006. 'Data Preprocessing for Supervised Leaning'. 1(1):7.

Nakajima, Jouchi. 2012. 'Bayesian Analysis of Multivariate Stochastic Volatility with Skew Distribution'. ArXiv:1212.5090 [Stat].

Hansen, Bruce E. 1994. 'Autoregressive Conditional Density Estimation'. *International Economic Review* 35(3):705–30.

Hansen, Peter R., and Asger Lunde. 2005. 'A Forecast Comparison of Volatility Models: Does Anything Beat a GARCH(1,1)?' *Journal of Applied Econometrics* 20(7):873–89.

Ryll, Lukas, and Sebastian Seidens. 2019. 'Evaluating the Performance of Machine Learning Algorithms in Financial Market Forecasting: A Comprehensive Survey'. *ArXiv:1906.07786 [q-Fin]*.

Anon. 2020. 'S&P 500 (^GSPC) Historical Data - Yahoo Finance'. Retrieved 18 April 2020 (https://finance.yahoo.com/quote/%5EGSPC/history/).

# Appendixes

## Appendix A

Detailed Research Questions

1. Would performance improve by increasing the size of the dataset?
   - To be tested on the vanilla case - when errors are normally distributed and there are no outliers, 4 datasets change only in the sample size (200,500,1000 and 3000). (**Table 1**)
2. How does the change in the outlier distribution affect the prediction accuracy of the ML algorithm (AR) and classic econometric model AR(1)?
   - To be tested on normal error data separately for evenly and unevenly distributed outliers. (**Table 2**)
3. What happens when the beta parameter in the dataset is reaching the stationarity border?
   - In order to check how the maximum likelihood and machine learning forecasting performances compare when the parameter value is closer to the stationarity border, we will be using stationary data for 200 and 3000 observations. (**Table 3**)
4. How does the change in the error distribution affect prediction accuracy of ML algorithm (AR) and classic econometric model AR(1) in cases where:
   - The errors are normally distributed, like in the vanilla case
   - The errors are drawn from a highly positively skewed distribution, for simplicity. The implications would be similar to the negatively skewed distribution case.
   - The errors are unevenly distributed, resulting in the fat-tailed distribution. The case when the true distribution is symmetric but fat-tailed

   The datasets are specified in (**Table 4**).
5. Is machine learning algorithm AR generally outperforming classical econometric model AR (1) in time series prediction?
   - The conclusions are based on the overall results of all 4 data tables mentioned above. **(Simulated Datasets**.)

**Appendix B**
Simulated Datasets

Table 1

| # | Case ID | Model | Sample Size | Outliers | Distribution of Errors | Beta | Table_Name |
|---|---|---|---|---|---|---|---|
| 1 | 1a_1 | AR(1) | 200 | No | Normal | 0.5 | vanillaAR1_200 |
| 2 | 1a_2 | AR(1) | 500 | No | Normal | 0.5 | vanillaAR1_500 |
| 3 | 1a_3 | AR(1) | 1000 | No | Normal | 0.5 | vanillaAR1_1000 |
| 4 | 1a_4 | AR(1) | 3000 | No | Normal | 0.5 | vanillaAR1_3000 |

Table 2

| # | Case ID | Model | Sample Size | Outliers | Distribution of Errors | Table_Name |
|---|---|---|---|---|---|---|
| 5 | 1b_1 | AR(1) | 200 | Evenly | Normal | evoutAR1_200 |
| 6 | 1b_2 | AR(1) | 3000 | Evenly | Normal | evoutAR1_3000 |
| 7 | 2b_1 | AR(1) | 200 | unevenly | Normal | unevoutAR1_200 |
| 8 | 2b_2 | AR(1) | 3000 | unevenly | Normal | unevoutAR1_3000 |

Table 3

| # | Case ID | Model | Sample Size | Beta | Table_Name |
|---|---|---|---|---|---|
| 9 | 1c_1 | AR(1) | 200 | 0.99 | stbordAR(1)_200 |
| 10 | 1c_2 | AR(1) | 3000 | 0.99 | stbordAR(1)_3000 |

Table 4

| # | Case ID | Model | Sample Size | Outliers | Distribution of Errors | Beta | Table_Name |
|---|---|---|---|---|---|---|---|
| 1 | 1d_1 | AR(1) | 200 | No | Normal | 0.5 | vanillaAR1_200 |
| 4 | 1d_4 | AR(1) | 3000 | No | Normal | 0.5 | vanillaAR1_3000 |
| 11 | 2d_1 | AR(1) | 200 | No | Skewed Normal | 0.5 | sknormAR1_200 |
| 12 | 2d_2 | AR(1) | 3000 | No | Skewed Normal | 0.5 | sknormAR1_3000 |
| 13 | 3d_1 | AR(1) | 200 | No | Fat-tailed | 0.5 | fattailAR1_200 |
| 14 | 3d_2 | AR(1) | 3000 | No | Fat-tailed | 0.5 | fattailAR1_3000 |