MINISTERUL EDUCAŢIEI ŞI CERCETĂRII AL REPUBLICII MOLDOVA

Universitatea Tehnică a Moldovei

Facultatea Calculatoare Informatică și Microelectronică

Departamentul Ingineria Software și Automatică

Admis la susținere
Șef departament: Fiodorov I. dr., conf.univ.

„___” _____ 2022

Crearea modelelor bazate pe algoritmi de învățare automată pentru predicția seriilor de timp

Teză de master

Student: _____ **Stamatin Alexandru, gr. IS-211M**

Coordonator: _____ **Beșliu Corina, lect. univ., dr.**

Consultant: _____ **Catruc Mariana, lect. univ.**

Chişinău, 2023

Rezumat

Teza de master cu titlul „Crearea modelelor bazate pe algoritmi de învățare automată pentru predicția seriilor de timp” a fost realizată de masterandul Stamatin Alexandru.

În cadrul tezei de master, au fost analizați algoritmii existenți pentru predicția seriilor de timp, a fost explorat și curățat un set de date din lumea reală, după care în baza algoritmilor selectați au fost create și evaluate modele de predicție. Au fost determinate caracteristicile de bază ale seriilor de timp prezente în setul de date cum sunt existența datelor lipsă și a multiplelor componente de sezonalitate. În rezultat, au fost selectate instrumentele de predicție pentru crearea modelelor. Modelele create au fost evaluate după diferite criterii pentru a determina cât mai exact calitatea predicțiilor. În baza rezultatelor obținute, au fost formulate concluzii cu privire la viabilitatea utilizării instrumentelor de predicție alese cu setul de date selectat, precum și ajustările necesare a parametrilor modelelor pentru a îmbunătăți calitatea predicțiilor.

Cuvintele cheie sunt: serie de timp, predicție, analiză exploratorie, modelare

În continuare este descris pe scurt fiecare capitol al proiectului de licență

- În primul capitol sunt analizate caracteristicile, modul de utilizare și avantajele algoritmilor existenți pentru predicția seriilor de timp, sunt determinați factorii ce trebuie luați în considerație la selectarea instrumentelor de predicție
- În al doilea capitol este realizată analiza exploratorie a setului de date din lumea reală, sunt identificate componentele de sezonalitate prezente precum și examinați pașii necesari pentru curățarea datelor
- În capitolul al treilea sunt prezentate transformările efectuate pentru curățarea setului de date, sunt aplicate instrumentele de predicție, este analizat impactul modificării valorilor unor parametri ai modelelor utilizate precum și comparate strategiile de evaluare a performanței modelelor.

Abstract

The Master Thesis with the title “Creating models based on machine learning algorithms for time series forecasting” was completed by the graduate student Stamatin Alexandru.

Within the master thesis, an analysis of the existing algorithms for time series forecasting was conducted, a real-world dataset was explored and cleaned, followed by the creation and evaluation of forecasting models based on the selected algorithms. The main characteristics of the time series present in the dataset, such as the existence of missing values and multiple seasonal patterns, were identified. As a result, a set of forecasting tools was selected for creating the models. The created models were evaluated using various metrics to determine as exactly as possible the quality of the forecasts. Based on the obtained results, conclusions were made about the viability of using the selected forecasting tools with the chosen dataset, as well as the tuning steps necessary to improve the quality of the forecasts.

The key words are: time series, forecast, exploratory analysis, modeling

Below is a brief description of the chapters present in the thesis:

- The first chapter offers an analysis of the characteristics, usage and advantages of the existing algorithms for time series forecasting, as well as the circumstances that have to be taken into consideration when selecting the forecasting tools
- The second chapter presents the exploratory analysis for the real-world dataset; the seasonal patterns that exist in the data are identified followed by the examination of the steps needed for data cleaning
- The third chapter describes the transformations that were needed to clean the dataset, the results of using the forecasting tools, as well as an analysis of the impact of hyperparameter tuning along with a comparison of various evaluation strategies for model performance

Contents

Introduction

The rapid development and wide adoption of modern information technology tools by private enterprises and public institutions has resulted in a significant increase in the amount of available data. This creates certain challenges such as ensuring the security and integrity of the stored data. At the same time, it presents important opportunities for institutions, as they can extract valuable knowledge from the data and use it for various activities such as resource allocation or cost optimization. As a result, the way in which institutions manage their data is having an increasing impact on their ability to achieve their stated goals.

In order to effectively monitor and manage the way in which a system is functioning, it is often important to analyze how the values of the system's key parameters change with time. Because of this, the efficient collection and processing of time series data is of special importance. As the number of devices connected to the internet is increasing and more and more processes are moving into the digital realm, the amount of data of this type is also growing significantly. A larger amount of data for training increases the likelihood of building a performant model. It is expected that in the near future the role of time-series data will expand considerably [1]. A logical consequence of this phenomenon is the growing demand for tools and algorithms that can efficiently process time series data.

One of the key reasons for the growing interest in time series data is the fact that it can be used in forecasting events or values for various parameters. Time series forecasting can provide real value to institutions in a wide variety of fields [2]. Because of that, there exists a significant interest in the research and development of tools and algorithms that would allow one to effectively analyze and extract knowledge from time series data. Modern hardware equipment and software modules enable the solving of tasks deemed impossible a few decades ago.

The release of new open-source software packages and the wide availability of performant hardware equipment makes it possible for institutions of all sizes to create and take advantage of advanced forecasting models. Recent algorithms allow solving more complex tasks such as analyzing a large number of time series and processing data that contains various patterns. As a result, it is possible to build forecasting models with an increased accuracy that can take into consideration a large number of data features of the data. Additionally, advances in the field of cloud services enable a higher degree of flexibility in the deployment and usage of forecasting models. New software packages that offer a detailed documentation and an intuitive API increase the accessibility of forecasting tools. These developments lead us to the conclusion that the adoption of forecasting models is becoming vital for entities of all kinds, thus making research in this field more important than ever.

Bibliography

1. Aileen NIELSEN. *Practical Time Series Analysis: Prediction with Statistics and Machine Learning*. O'Reilly Media, Inc., 2019. ISBN 978-1492041658
2. CERQUEIRA, V., TORGO, L. & MOZETIČ, I. Evaluating time series forecasting models: an empirical study on performance estimation methods. In: *Machine Learning* [online]. 2020, volume 109, issue 11, pages 1997 – 2028. ISSN 1573-0565. Available: https://link.springer.com/article/10.1007/s10994-020-05910-7
3. Antti SORJAMAA, Jin HAO, Nima REYHANI, Yongnan JI, Amaury LENDASSE. Methodology for long-term prediction of time series. In: *Neurocomputing* [online]. 2007, Volume 70, Numbers 16-18, pages 2861-2869. ISSN 0925-2312. Available: https://www.academia.edu/8159103/Methodology_for_long_term_prediction_of_time_series?auto=citations&from=cover_page
4. J. Scott ARMSTRONG. *Principles of Forecasting: A Handbook for Researchers and Practitioners*. Dordrecht: Kluwer Academic Publishers, 2001. ISBN 0-7923-7930-6
5. Jan ADAMOWSKI, Christina KARAPATAKI. Comparison of Multivariate Regression and Artificial Neural Networks for Peak Urban Water-Demand Forecasting: Evaluation of Different ANN Learning Algorithms. In: *Journal of Hydrologic Engineering* [online]. 2010, Vol. 15, No. 10. ISSN 1943-5584. Available: https://www.academia.edu/948036/Comparison_of_Multivariate_Regression_and_Artificial_Neural_Networks_for_Peak_Urban_Water_Demand_Forecasting_Evaluation_of_Different_ANN_Learning_Algorithms?from=cover_page
6. Toni TOHARUDIN, Resa Septiani PONTOH, Rezzy Eko CARAKA, Solichatus ZAHROH, Youngjo LEE, Rung Ching CHEN. Employing long short-term memory and Facebook prophet model in air temperature forecasting. In: *Communications in Statistics - Simulation and Computation* [online]. 2021. ISSN 1532-4141. Available: https://doi.org/10.1080/03610918.2020.1854302
7. TAYLOR SJ, LETHAM B. 2017. Forecasting at scale. *PeerJ Preprints* 5:e3190v2 Available: https://doi.org/10.7287/peerj.preprints.3190v2
8. *Prophet | Forecasting at scale.* Facebook Open Source, [cited 02.10.2022]. Available: https://facebook.github.io/prophet/
9. Lorenzo MENCULINI, Andrea MARINI, Massimiliano PROIETTI, Alberto GARINEI, Alessio BOZZA, Cecilia MORETTI, Marcello MARCONI. Comparing Prophet and Deep Learning to ARIMA in Forecasting Wholesale Food Prices [online]. Available: https://doi.org/10.48550/arXiv.2107.12770

10. da S. GOMES, G.S., LUDERMIR, T.B. & LIMA, L.M.M.R. Comparison of new activation functions in neural network for forecasting financial time series. In *Neural Computing and Applications* [online]. 2011, Vol. 20, pages 417–439. ISSN 1433-3058. Available: https://doi.org/10.1007/s00521-010-0407-3
11. Le XUAN-HIEN, Hung Viet HO, Giha LEE, and Sungho JUNG. Application of Long Short-Term Memory (LSTM) Neural Network for Flood Forecasting. In *Water* [online] 2019, Volume 11, Issue 7. ISSN 2073-4441. Available: https://doi.org/10.3390/w11071387
12. Mohamed ABDEL-NASSER, Karar MAHMOUD. Accurate photovoltaic power forecasting models using deep LSTM-RNN. In: *Neural Computing and Applications* [online]. 2019, Volume 31, pages 2727–2740. ISSN 1433-3058. Available: https://doi.org/10.1007/s00521-017-3225-z
13. H.D. NGUYEN, K.P. TRAN, S. THOMASSEY, M. HAMAD. Forecasting and Anomaly Detection approaches using LSTM and LSTM Autoencoder techniques with the applications in supply chain management. In *International Journal of Information Management* [online]. 2021, Volume 57, Article 102282. ISSN 0268-4012. Available: https://doi.org/10.1016/j.ijinfomgt.2020.102282
14. *Optuna – A hyperparameter optimization framework*. Preferred Networks, Inc. [cited 08.10.2022]. Available: https://optuna.org/
15. *Overview of the NeuralProphet Model*. Oskar TRIEBE [cited 08.10.2022]. Available: https://neuralprophet.com/html/model-overview.html
16. KOMOROWSKI, M., MARSHALL, D.C., SALCICCIOLI, J.D., CRUTAIN, Y. (2016). Exploratory Data Analysis. In: *Secondary Analysis of Electronic Health Records* [online]. Springer, Cham. ISBN 978-3-319-43742-2. Available: https://doi.org/10.1007/978-3-319-43742-2_15