New tendencies in linear prediction of events

Carmen ROTUNĂ, Antonio COHAL, Ionuț SANDU, Mihail DUMITRACHE

National Institute for Research and Development in Informatics - ICI Bucharest

carmen.rotuna@rotld.ro, antonio.cohal@rotld.ro, ionut@rotld.ro, mihaildu@rotld.ro

Abstract: The continuous development in the field of data science has greatly transformed big data analysis. Data is the foundation of innovation, but their value comes from the information that data experts can collect and then interpret. The volume of data is constantly increasing nowadays, thus businesses could benefit from analyzing existing data to make valuable predictions about the future and to develop a coherent business plan. Time series analysis enables companies to analyze data in order to extract meaningful characteristics and generate useful timely predictions. Mainly, time-series data consists of sequences of chronologically stored observations and are generated by recording, business metrics, monitoring sensors, observing network traffic, etc. In this study, we set out to implement a forecast model for the .ro Registry and, therefore, chose the Prophet FB, because it offers an open source software tool that supports the business area and has been successfully tested in different scenarios. The results showed that Prophet can generate accurate forecasts which can be used to optimize Registry services.

Keywords: time-series, quantitative forecasting, linear prediction, FB Prophet, data model.

Noi tendințe în predicția liniară a evenimentelor

Rezumat: Dezvoltarea continuă în domeniul științei datelor a transformat considerabil analiza datelor de dimensiuni mari. Acestea reprezintă fundamentul inovării, însă valoarea lor provine din informaţiile pe care experţii le pot colecta despre date şi le pot utiliza ulterior. Volumul informațiilor este în continuă creștere în zilele noastre, astfel încât întreprinderile ar putea obține numeroase beneficii în urma analizei datelor existente, pentru a face predicții despre viitor și pentru a dezvolta un plan de afaceri coerent. Analiza datelor de tip serii de timpi permite companiilor să extragă caracteristici semnificative și să genereze previziuni utile bazate pe timp. În principal, datele de tip serii de timpi reprezintă secvențe de observații stocate cronologic și sunt generate prin înregistrarea, de exemplu, a metricilor din domeniul afacerilor, monitorizarea senzorilor, observarea traficului de rețea, etc.

În acest studiu, ne-am propus să implementăm un model de previziune pentru Registrul .ro și, prin urmare, am ales Prophet FB, deoarece oferă un instrument software open source dedicat zonei business și a fost testat cu succes în diferite scenarii. Rezultatele au arătat că Prophet poate genera previziuni exacte care pot fi utilizate pentru a optimiza serviciile Registrului.

Cuvinte cheie: serii de timp, predicție cantitativă, predicție liniară, Profet FB, model de date.

1. Introduction

The volume of generated data has increased exponentially in recent years. The world's technological per-capita capacity to store information has roughly doubled every 40 months since the 1980s. (Hilbert, 2011). A large portion of these data comes as time series data, generated by IoT, monitoring or mobile applications and devices (Vevera, 2019) and is used in machine learning through means of data mining to generate data intelligence (Petre, 2019).

Nowadays companies and governments are in a position to use existing data to forecast annually or for a shorter period of time the demand for specific services or products in order to create a realistic business plan. Statistical software facilitates rapid forecasting using quantitative forecasting models. In order to discover patterns in the evolution of economic phenomena, it is necessary to know the evolution of their past.

Any data that can be observed sequentially over time is a time series. Examples of time series include: daily company stock prices, weather information, company sales results, device monitoring, etc.

https://doi.org/10.33436/v29i3y201902

Forecasting defines the process of collecting and analyzing data from the past and present in order to make trustworthy predictions about the future. It enables organizations to anticipate, plan, set goals and detect abnormal application or user behavior.

Qualitative forecasting is a statistical technique that enables predictions about the future through expert analysis instead of numerical analysis. This method of forecasting is mainly used when historical data is not available. The data is usually collected using tools such as questionnaires or specific measuring devices.

Quantitative forecasting is used to forecast future data based on the analysis of present and past data. These methods can be applied only when three conditions are met: there is sufficient past numerical information available, the information can be quantified, and when it is possible to presume that the patterns in the information will continue in the future. Several time series forecasting models have been developed and deployed in various use cases such as: The random walk, autoregressive (AR), moving average (MA), ETS, TBATS and ARIMA. These are widely recognized statistical forecasting models which predict future observations of a time series on the basis of some linear function of past values (Adhikari, 2013).

Having the aim to develop forecasts for .ro Registry commands usage, several methods and tools have been analyzed. These methods, when applied, lead to different levels of accuracy in forecasting. For example, GMDH neural network was proved to have better forecasting performance than the classical algorithms such as Double Exponential Smooth, ARIMA and back-propagation neural network. (Li, 2017). Also, FB Prophet forecasts had substantially lower prediction error than other automated forecast methods such as Naive forecasting methods, ETS methods, and TBATS model (Taylor, 2017). Each method has its own properties, accuracy and costs that must be taken into consideration when choosing a specific method.

2. Time-series data analysis for forecasting

Time-series analysis is a statistical method of analyzing data from repeated observations on a single unit or individual at regular intervals over a large number of observations (Velicer, 2003). It is an approach which enables the analysis of time series data with the goal to extract useful characteristics from data and generate business relevant information. Time-series data is a sequence of information stored in timely manner. When forecasting time series data, the aim is to estimate how the sequence of observations will continue in the future.

There are 11 different traditional time series forecasting methods:

- Autoregression (AR) - uses timely observations from the past as input to a regression equation to predict future values and it is suitable for time series without trend and seasonal components;
- Moving Average (MA) - specifies that the output variable depends linearly on the current and various past values of a stochastic term. The method is adequate for univariate time series without trend and seasonal components;
- Autoregressive Moving Average (ARMA) - combines both Autoregression (AR) and Moving Average (MA) models and assumes that the time series fluctuates more or less uniformly around a time-invariant mean;
- Autoregressive Integrated Moving Average (ARIMA) - combines both Autoregression (AR) Moving Average (MA) models, while adding integration, a pre-processing step of the sequence to make the sequence stationary. The method is suitable for univariate time series with trend and without seasonal components;
- Seasonal Autoregressive Integrated Moving-Average (SARIMA) – is an extension of ARIMA model to support the seasonal component of the series. It is used for univariate time series with trend and seasonal components;
- Seasonal Autoregressive Integrated Moving-Average with Exogenous Regressors (SARIMAX) - is an extension of the SARIMA model that includes the modeling of exogenous variables;

- Vector Autoregression (VAR) - is the generalization of Autoregression to multiple parallel time series, for example, multivariate time series;
- Vector Autoregression Moving-Average (VARMA) - It is the generalization of ARMA to multiple parallel time series and it is suitable for multivariate time series without trend and seasonal components;
- Vector Autoregression Moving-Average with Exogenous Regressors (VARMAX) - extends VARMA model by adding the modeling of exogenous variables;
- Simple Exponential Smoothing (SES) - models the next time step as an exponentially weighted linear function of observations at prior time steps;
- Holt Winter's Exponential Smoothing (HWES) - models the next time step as an exponentially weighted linear function of observations at prior time steps, taking trends and seasonality into account (Brownlee, 2018).

A time series consists in observing a variable *Y* in relation to time. *Y* measurements Y_1, Y_2, ..., Y_t..., Y_T are made at equal time intervals *1, 2, ..., t, ..., T* and a time series will be presented as follows:

$$Y = \begin{pmatrix} 1 & 2 & ...t & ...T \\ Y_1 & Y_2 & ...Y_t & ...Y_T \end{pmatrix}$$

In order for the time series analysis to generate a prediction as close to reality as possible, the length of the analyzed period must be sufficiently large to make it possible to forecast adequate data. The cyclicity of the measurements is given by the type of activity of a company. For example, when it comes to passengers traveling with a particular airline or the products sold by a store, when developing forecasts based on statistical methods, we start from the hypothesis that the phenomenon will continue to have the same behavior as in the past.

Anticipation of business phenomena is a relatively difficult task due to the complexity of the business environment. Thus, there are differences, called forecast errors, between the forecast values for a certain time range and the real values for the same time range. For example, at time t the prediction error is the difference between the actual value Y_t and the forecast value $\hat{Y}_t$ both associated with the same time *t*.

$$e_t = Y_t - \hat{Y}_t$$

When the statistical model generates the forecasts corresponding to the observations $\hat{Y}_1$, $\hat{Y}_2$, ..., $\hat{Y}_t$, in order to measure its quality to generate adequate forecast, a series of synthetic indicators errors are used, the most common being mean squared error,

$$MSE = \frac{1}{S}\sum_{k=1}^{S}(Y_k - \hat{Y}_k)^2$$

the mean absolute error

$$MAE = \frac{1}{S}\sum_{k=1}^{S}\left|Y_k - \hat{Y}_k\right|$$

and the mean absolute percentage error

$$MAPE = \frac{1}{S}\sum_{k=1}^{S}\left|\frac{Y_k - \hat{Y}_k}{Y_k}\right|$$

The indicators above are used to see the model's ability to make forecasts as close to the reality at time *t*.

When a forecast model does not work as expected, consistent human resources are needed to adjust the parameters of the methods to certain aspects. In general, adjusting the methods requires a clear understanding of how the model works. To achieve this goal, organizations need data science teams in order to obtain coherent forecasting.

3. Facebook Prophet prediction model

When applying most of the existing algorithms into practice, researchers identified several challenges, such as data availability, missing values, forecast discontinuity, associated with producing reliable forecasts especially when large amounts of data are involved.

To address these challenges, in early 2017, Taylor and Letham developed a forecasting model "at scale" that combines configurable models with analyst-in-the-loop performance analysis.

They used a modular regression model with configurable parameters that can be intuitively adjusted by analysts with domain knowledge about the time series (Taylor, 2017).

Prophet forecasting is based on an additive model which enables saturating forecasts, trend changepoints, seasonality, holiday effects and regressors, multiplicative seasonality and uncertainty intervals. It involves a decomposable time series model with three main model components: trend, seasonality, and holidays which are combined in the following equation:

$$y(t) = g(t) + s(t) + h(t) + er$$

where *g(t)* is the trend function which models non-periodic changes in the value of the time series, *s(t)* represents periodic changes (weekly and yearly seasonality), and *h(t)* represents the effects of holidays which occur on potentially irregular schedules over one or more days. The *er* error represents any changes which are not accommodated by the model (Taylor, 2017).

Prophet model allows analysts to intervene in several places in the model without having to understand the statistics behind the model. It is used in many Facebook applications for forecasts within different areas. It can be used as an automatic tool without any human intervention and at the same time allows analysts to manually adjust parameters for increased accuracy.

The most relevant features of FB Prophet are (Facebook Prophet Official Documentation, 2019):

- Saturating Forecasts - Prophet uses a linear model which enables forecasts using a logistic growth trend model, with a specified carrying capacity. The maximum achievable point is called carrying capacity and represents the saturation level;
- Trend Changepoints - Prophet enables automatic detection of changepoints and a fine-tuning of this process providing several input arguments which can be used such as the number of potential changepoints, the location of the changepoints and trend flexibility;
- Seasonality – Prophet uses a Fourier order of 10 for yearly seasonality and 3 for weekly seasonality and it will interpret by default annual and weekly seasonality, when the time series has more than two cycles. Prophet has a dedicated function which allows adding other seasonality monthly, quarterly, hourly, etc.;
- Holidays and Special Events – Prophet incorporates a list of holidays for each country into the model composed of major holidays and allows adding new events through a predefined method. Analysts usually have experience concerning holidays impact on the forecast and can add the relevant ones;
- Uncertainty Intervals - the forecast is affected by three sources of uncertainty: the potential for future trend changes, the seasonality estimates, and supplementary observation noise.

The forecast error with Prophet is much lower compared to other models such as Naive forecasting methods, ETS methods, and TBATS model as shown in Fig 1. Each method has its own properties, accuracies, and costs that must be taken into consideration when a choice is made.

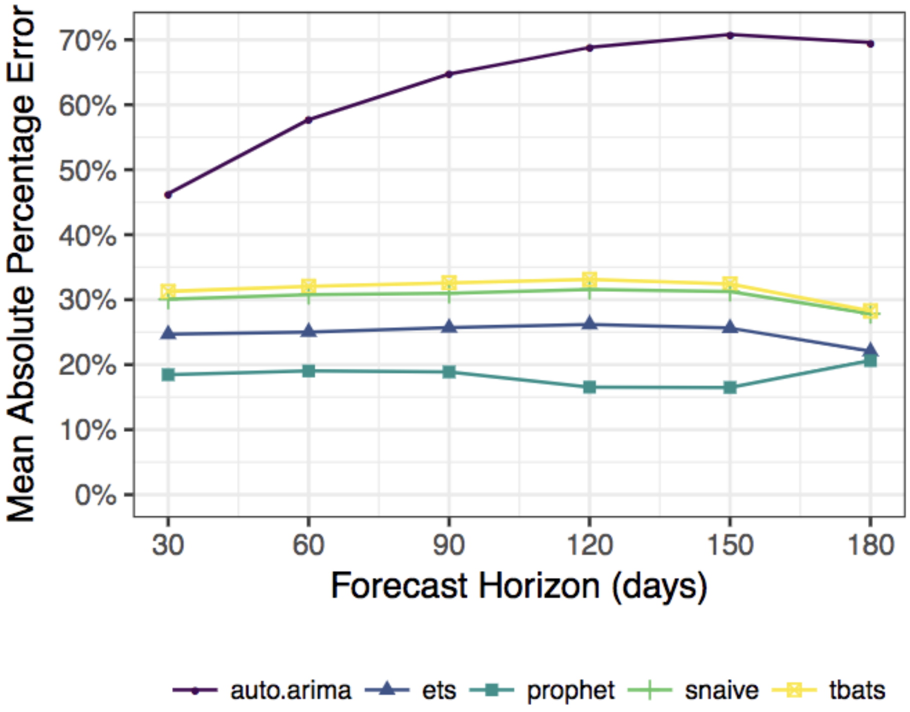


Figure 1. Comparison between forecasting models (Taylor, 2017)

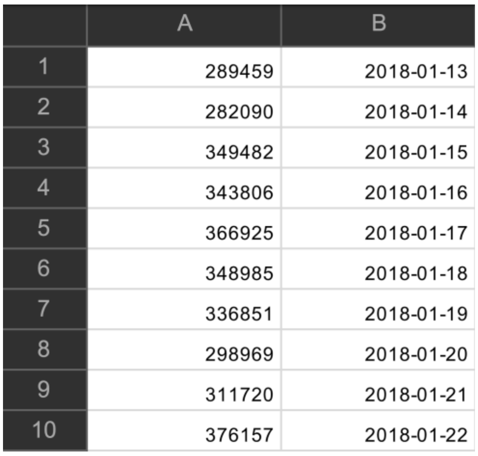
Another important feature of Facebook Prophet is that it does not require any previous knowledge in forecasting time series data, as it automatically detects seasonal trends behind the data and provides configurable parameters. Therefore, it also allows those who are not specialists in statistics to start using it and obtain reasonable results, which are often equal or sometimes even better than those produced by experts.

4. Time Series Analysis of .ro Registry registrar commands

In order to create forecasts concerning the registrar commands usage provided by the .ro domain Registry, we chose FB Prophet because it is a solution that supports the business area and has been successfully tested in various scenarios. This tool was developed with the aim to create high-quality forecasts for the business area, to have a rigorous methodology behind the scenes, but also to provide intuitive parameters which business analysts can adjust.

FB Prophet provides a forecasting library implemented in R and Python. The data input to Prophet is always a data file with two columns: ds and y. The ds (datestamp) column should be formatted, preferably YYYY-MM-DD for a date or YYYY-MM-DD HH:MM:SS for a timestamp and the y column must be numeric, and represents the measurement we wish to forecast (Facebook Prophet Official Documentation, 2019).

A FB Prophet input data sample is provided in the figure bellow which shows the “domain-info” registrar command API hits (column A) by date (column B).



	A	B
1	289459	2018-01-13
2	282090	2018-01-14
3	349482	2018-01-15
4	343806	2018-01-16
5	366925	2018-01-17
6	348985	2018-01-18
7	336851	2018-01-19
8	298969	2018-01-20
9	311720	2018-01-21
10	376157	2018-01-22

Figure 2. FB Prophet data sample for .ro Registry

Facebook Prophet was used to predict the usage of .ro domain Registry registrar commands having the following goals:

- Overview of the collected data and check for possible inconsistencies;
- Analysis per command type and correlational analysis;
- Perform Time Series Analysis and interpret the results;
- Predict commands usage using Facebook Prophet.

The “domain-info” registrar command API hits were selected for analysis. The data from the Registry’s internal database was exported as a CSV file and then it was used as input to FB Prophet for generating a forecast. The model was developed by instantiating a new Prophet object and fetching the input data, consisting in “domain-info” command API hits for 7 months from March to September 2018. By applying the method to the historical data, we aimed to compare predicted data against actual data, to verify the accuracy of the forecast.

In the figure below, the actual data is represented by the black dots while the blue line represents the prediction made by FB Prophet. Although there are a few leaps in the model, Prophet does not take them into consideration.

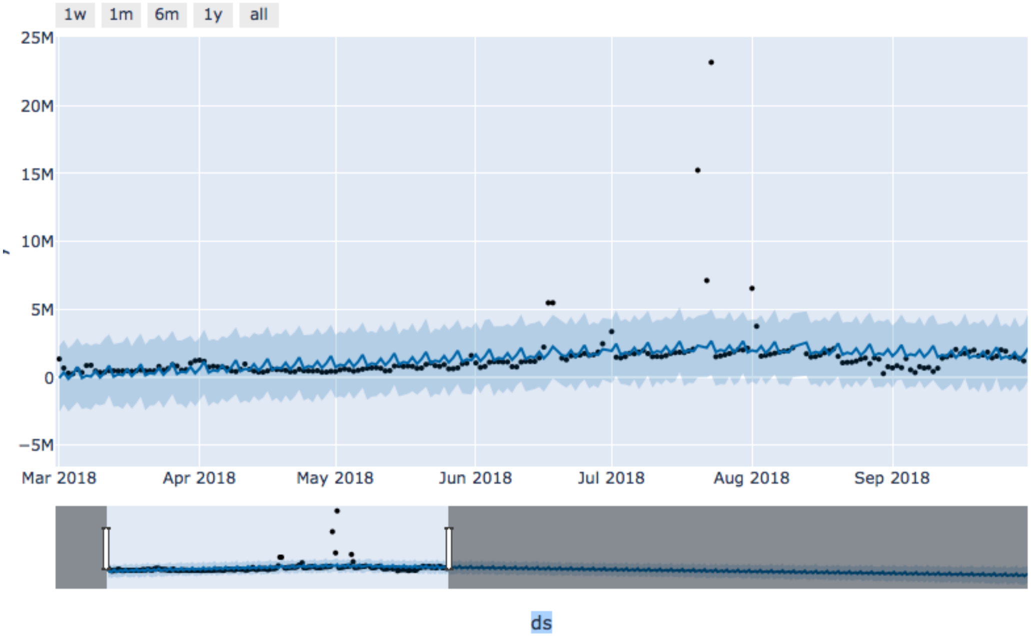


Figure 3. Input data

The input data was then used to predict the command usage for the following two months October and November. The forecasts obtained for the historical data using Prophet are shown in the figure bellow:

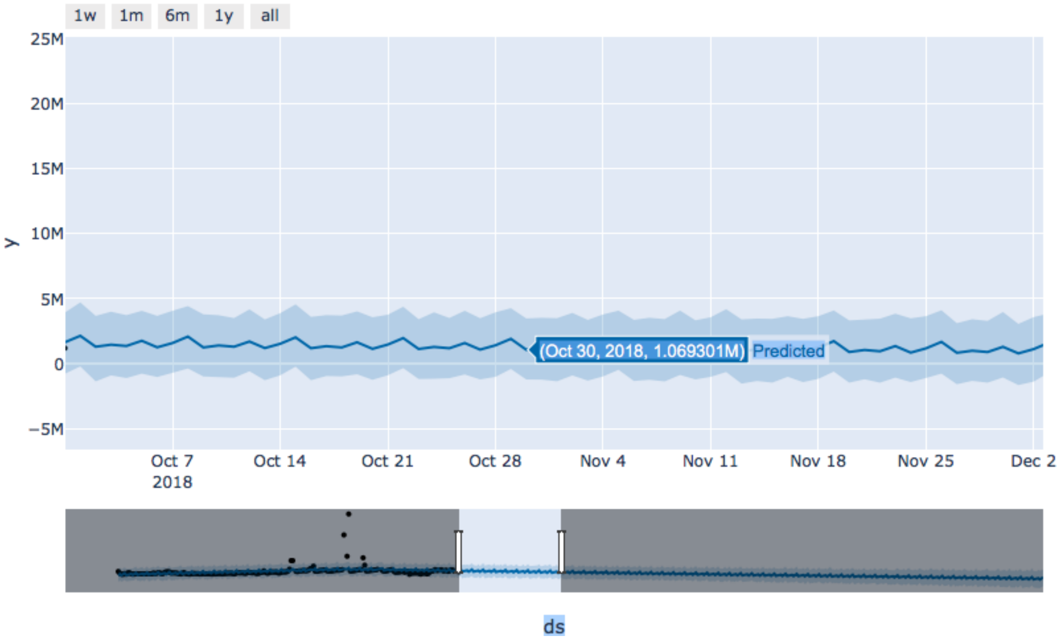


Figure 4. Predicted data

These measures are used to determine how accurate the Prophet method is able to reproduce the time series data that are already available. For an increased precision, the model was refined by adding holidays and known events that have major impacts on our time series and the weekly seasonality parameter was set to “true”. In order to measure the accuracy of the forecast, the real data from September to October was compared to predicted data and the result proved the model can fit the data points with high precision as shown in Figure 5.

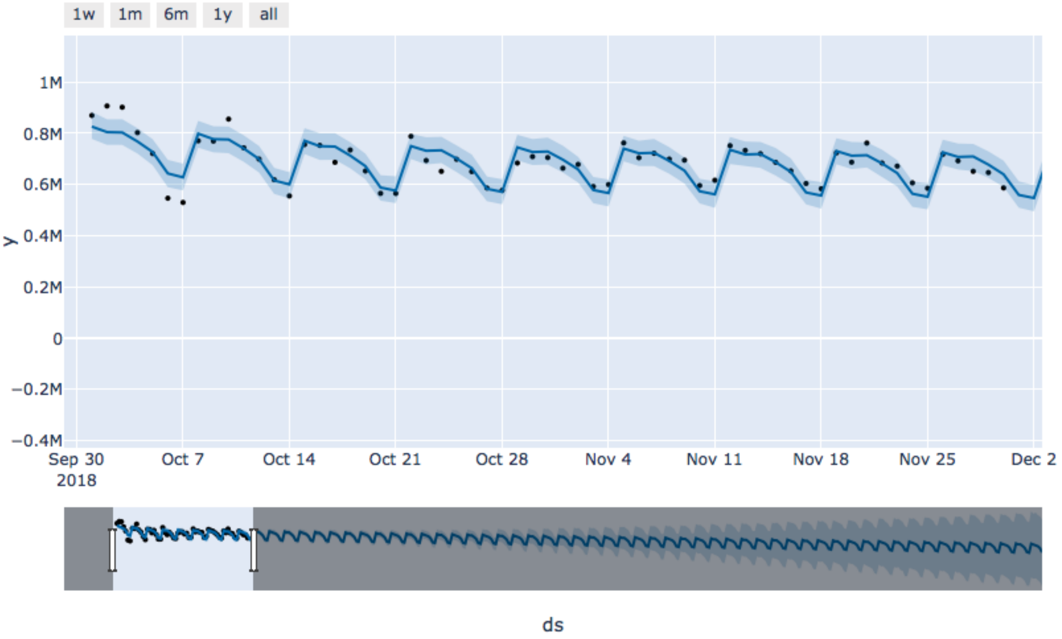


Figure 5. Predicted vs real data

Thus, FB Prophet through its capability to allow multiple adjustments within the model proved to be valuable tool for forecasting within this selected use case.

5. Conclusions

Nowadays public administration and companies can have significant benefits from anticipating the future because it allows them to make better business decisions. In order to develop an efficient business plan, the managers need to forecast the demand from clients for specific periods in order to coordinate their teams accordingly. Achieving reasonably accurate forecasts of a time series is a very important but challenging task.

The main objective of time series analysis is to develop a model that describes the pattern of the time series and could be used for forecasting. FB Prophet is a widely applied and effective forecasting model which is implemented in production environments by several companies and has the advantage that it can be implemented as an automatic tool and, at the same time, allows analysts to edit parameters for increased accuracy. The main advantage of this method is that it makes a trade-off between precision and scaling in the sense it does not add the data when missing, but simply makes an average of the ends. A disadvantage when using the FB Prophet could be, as shown in Figure 1, the MAPE error is increasing when the Prophet FB method is applied for long periods of time.

For this study, FB Prophet was chosen and implemented and the results showed that is a valuable open source tool that can be included into the python data science workflow and can be used to create forecasts with a high degree of precision. The study will continue by investigating means to extend and optimize the model to other timely information in order to increase the accuracy of the forecast data.

Confirmare

Acest articol a fost susținut în cadrul *Simpozionului Slove Muscelene*, ediția a XI-a, desfășurat în perioada 18-19 iulie 2019 la Câmpulung Muscel.

REFERENCES

1. Adhikari, R., Agrawal, R. (2013). *An introductory study on time series modeling and forecasting*, preprint arXiv:1302.6613.
2. *Big Data Facts, How much data is out there?* (2017) https://www.nodegraph.se/big-data-facts/
3. Byrd, R. H., Lu, P., Nocedal, J. (1995). *A limited memory algorithm for bound constrained optimization*. SIAM Journal on Scientic and Statistical Computing 16 (5).
4. *Facebook Prophet Official Documentation* (2019). Github https://facebook.github.io/prophet/
5. Harvey, A., Peters, S. (1990). *Estimation procedures for structural time series models*. Journal of Forecasting 9, 89-108.
6. Hilbert, M., and López, P. (2011). *The World's Technological Capacity to Store, Communicate, and Compute Information*. Science, Vol. 332, Issue 6025, 60-65 DOI: 10.1126/science.1200970
7. Hyndman, R. J, Athanasopoulos, G. (2012). Forecasting: *Principles and Practice* https://www.otexts.org/book/fpp
8. Brownlee, J. (2018). *11 Classical Time Series Forecasting Methods in Python* https://machinelearningmastery.com/time-series-forecasting-methods-in-python-cheat-sheet/
9. Li, Rita, Yi Man; Fong, Simon; Chong, Kyle Weng Sang (2017). *Forecasting the REITs and stock indices: Group Method of Data Handling Neural Network approach*. Pacific Rim Property Research Journal. 23 (2): 123–160. doi:10.1080/14445921.2016.1225149

10. Petre, I., Boncea, R., Zamfiroiu, A., Rădulescu, C. (2019). *A Time-Series Database Analysis Based on a Multi-attribute Maturity Model*. Studies in Informatics and Control, 28, 177-188. 10.24846/v28i2y201906

11. Taylor, S. J., Letham, B. (2017). *Forecasting at scale*. Peer J Preprints 5:e3190v2 https://doi.org/10.7287/peerj.preprints.3190v2

12. Velicer, Wayne & Fava, Joseph. (2003). *Time Series Analysis*.

13. Vevera, A. V., Onofrei-Riza, D. B. (2019). *Investigații mobile – captură, analiză și stocare a datelor sensitive*. Romanian Journal of Information Technology and Automatic Control, 29(1), 45-50. (6)

14. Yang Lyla (2019) *A Quick Start of Time Series Forecasting with a Practical Example using FB Prophet*, https://towardsdatascience.com/a-quick-start-of-time-series-forecasting-with-a-practical-example-using-fb-prophet-31c4447a2274

15. Zhang, G. P., Qi, M. (2005). *Neural network forecasting for seasonal and trend time series*. European Journal of Operational Research 160 (2), 501-514.

16. Zhang, G. P. (2003). *Time series forecasting using a hybrid ARIMA and neural network model*. Neurocomputing 50, 159-175. 2.

Carmen ROTUNĂ a absolvit programul de master din cadrul Facultății de Matematică și Informatică, Universitatea București. În prezent, este Cercetător Științific în cadrul Institutului Național de Cercetare-Dezvoltare în Informatică unde desfășoară activități de cercetare în domeniile eGovernment, eServices, Cloud, Big Data și IoT, fiind autor și co-autor al unor articole publicate în reviste de specialitate și volume de conferință recunoscute la nivel național și internațional precum și livrabile de proiect. Totodată a participat în proiecte naționale și europene din aria IT&C: SPOCS – Simple Procedures Online for Cross-border Services (CIP-ICTPSP), eSENS - Electronic Simple European Networked Services (CIP ICT), Cloud for Europe C4E (FP7), iar în prezent participă în proiectul TOOP - The “Once-Only” Principle Project (H2020), unde are rolul de coordonator la nivel național pentru pachetul de lucru WP2 Arhitectură și coordonator național pe zona de pilotare.

Carmen ROTUNĂ graduated the Faculty of Mathematics and Computer Science, University of Bucharest Master program. Currently she is a Scientific Researcher at the National Institute for Research and Development in Informatics where she carries out research activities in eGovernment, eServices, Cloud, Big Data and IoT, being the author and co-author of articles published in specialized journals and conference volumes recognized nationally and internationally as well as project deliverables. She was also a team member in national and European projects in the IT&C area: SPOCS - Simple Procedures Online for Cross-border Services (CIP-ICTPSP), eSENS - Electronic Simple European Networked Services (CIP ICT), Cloud for Europe C4E (FP7), and currently participates in TOOP - The "Once-Only" Principle Project (H2020), where she has the role of national coordinator for the WP2 Architecture work package and national coordinator for Piloting.

Antonio COHAL este în prezent Cercetător Științific în cadrul Institutului Național de Cercetare-Dezvoltare în Informatică unde desfășoară activități de cercetare în domeniile eGovernment, eServices, Cloud, Big Data și IoT, fiind autor și co-autor al unor articole publicate în reviste de specialitate și volume de conferință recunoscute la nivel național precum și livrabile de proiect. Totodată a participat în proiecte naționale și europene din aria IT&C: SPOCS – Simple Procedures Online for Cross-border Services (CIP-ICTPSP), Sistem optic integrat de gestionare a defectelor din industria textilă - TexDef. Participă la elaborarea, dezvoltarea şi întreţinerea aplicaţiilor software şi a bazelor de date.

Antonio COHAL graduated the Faculty of POLYTECHNICS NUCLEAR POWER PLANTS. Currently he is a Scientific Researcher at the National Institute for Research and Development in Informatics where he carries out research activities in eGovernment, eServices, Cloud, Big Data and IoT, being the author and co-author of articles published in specialized journals and conference volumes recognized nationally as well as project deliverables. He was also a team member in national and European projects in the IT&C area: SPOCS - Simple Procedures Online for Cross-border Services (CIP-ICTPSP) Integrated optical defect management system in the textile industry - TexDef. Participate in the development, use and maintenance of software applications and a databases.

Ionuț SANDU este licențiat în Știința Sistemelor și a Calculatoarelor (2006), obține master în Administrație Publică Electronică în anul 2007. Din anul 2010 devine cercetător științific în cadrul departamentului de Administrare Domenii .RO. din ICI – București, iar începând cu anul 2015 devine șef serviciu tehnic RoTLD și Cercetător Științific gradul III în cadrul aceluiași institut. Are responsabilități de administrare sisteme, dezvoltare de noi servicii, dezvoltare și mentenanță a infrastructurii de comunicații, precum și relația cu partenerii. În prezent este șeful departamentului RoTLD și Cloud Computing.

Ionuț SANDU graduated university with a BS în Computer and Systems' Science (2006) and obtained a Master's Degree in Electronic Public Administration in 2007. In 2010, he became Scientific Researcher within the .ro Domain Administration Department (RoTLD) of the National Institute for Research and Development in Informatics, ICI Bucharest, and since 2015 is Scientific Researcher Grade III and Head of the Technical Division of RoTLD, with responsibilities in systems' administration, development of new services, development and maintenance of communication infrastructures. He is also in charge with maintaining a close relationship with RoTLD's Partners. Currently, he is Head of the "RoTLD and Cloud Computing" Department.

Mihail DUMITRACHE este absolvent al Facultății de Electrotehnică, Universitatea Politehnica din București, specializarea Inginerie Asistată de Calculator, inginer și doctor în Inginerie Electrică. Deține studii masterale în specializarea Inginerie Electrică, Universitatea Politehnică din București și în specializarea Administrație Publică Electronică, Universitatea din București. Și-a început activitatea profesională în cadrul Institutului Național de Cercetare-Dezvoltare în Informatică - ICI București în anul 2002 ca programator. În prezent este Cercetător Științific III, Șef Serviciu Administrare domenii RoTLD și Lector Universitar la Universitatea din București. Este autor și coautor al unor studii și articole de specialitate.

Mihail DUMITRACHE PhD graduated from “Politehnica” University of Bucharest, Faculty of Electrical Engineering, with an Engineer’s Degree and, later on, a PhD in Computer Assisted Engineering. In between, he obtained two Master’s Degrees, one in Electrical Engineering, at “Politehnica” University of Bucharest and one in Electronic Public Administration, at Bucharest University. His professional career started at the National Institute for Research and Development in Informatics, ICI Bucharest in 2002 as a computer programmer. Currently, he is Scientific Researcher Grade III and Head of the .ro Domain Administration Department (RoTLD), and also Lecturer at the University of Bucharest. He is author and co-author of several scientific studies and articles.