DOI: https://doi.org/10.31410/ITEMA.2023.73

Deep Machine Learning for Time Series Inbound Tourism Forecasting

Ivanka Vasenska[1]

Received: October 18, 2023
Accepted: October 19, 2023
Published: May 28, 2024

Keywords:
Time series;
Deep machine learning;
Artificial intelligence;
Bulgaria inbound tourism forecast

Creative Commons Non Commercial CC BY-NC: This article is distributed under the terms of the Creative Commons Attribution-Non-Commercial 4.0 License (https://creativecommons.org/licenses/by-nc/4.0/) which permits non-commercial use, reproduction and distribution of the work without further permission.

Abstract: *Accurate inbound tourism flow forecasting has been a challenge for all stakeholders related to the sector. The multidisciplinary character of the tourism product which has been directly and indirectly influenced by all types of risks, cataclysms and crises further exposed its intangible nature to shocks and flows disruption. Thus, forecasting inbound tourism flows with advanced data science and AI (artificial intelligence) methods has been gaining momentum, which the COVID-19 pandemic boosted. Therefore, this paper aims to examine the relevant AI forecasting methods by applying a deep machine learning technique comparing different Python time series forecasting libraries via a Jupyter Notebook computer environment. Bulgaria's inbound tourism data has been used to develop an advanced deep neural network with the DARTS Python library and compare its accuracy with other Python library models.*

1. INTRODUCTION

Deep machine learning (DML) is a branch of AI and a sub-branch of machine learning (ML) using artificial neural network (ANN) architectures that are significantly advanced and can be applied through computer iterations to speech recognition, natural language processing and other domains (Géron, 2019; Goodfellow et al., 2016). Artificial neural networks use network architectures (similar to the biological neural networks (NNs) that the human brain uses) with a large number of interconnected processing layers. The birth of idea and description of an artificial neural network was first published in the 1940s, presenting a simplified model of how the human neuron works. McCulloch and Pitts (1943), describe in a paper the mathematical structure of a simplified neural model viewed as a Threshold Logic Element (THL) which ever since has been considered as the first mathematical model of a neural network. Building on the ideas of Turing (1936) the paper by McCulloch and Pitts provides a way to describe brain functions in abstract terms and demonstrates that simple elements connected in a neural network can have enormous computational power.

NNs are models that resemble human brain activity and, in particular, the connections between neurons (nodes) in the human brain. A node or "neuron" is a computational unit that has one or more weighted values and acts as a miniature mathematical function. Depending on the task at hand, the nodes are organized in a layer to make up a network - AAN, with a possible size from one layer to 100 layers of nodes. The first layer is the input layer, and the output layer is the resulting function transformation, once propagated through the middle "hidden" layers, which is where the term "deep" in DML comes from - the depth of the network's hidden layers.

DML algorithms are applied in tourism and hospitality sectors to recognize faces in an image or footage, in check-in/ check-out procedures, at airports through automatic face recognition,

[1] South-West University "Neofit Rilski", 60 Ivan Mihaylov str., Bulgaria

OPEN ACCESS

and even to detect emotions in people passing a certain point (eg. the happiness of those leaving the facilities).

The most common methodology for building an NN architecture structure is by applying the Python computer language and its libraries on potentially thousands of multi-GPU servers through computational graphics and a Python library. The main Python library ecosystem applied for time series forecasting is SciPy and it consists of the depicted-on Figure 1. libraries.

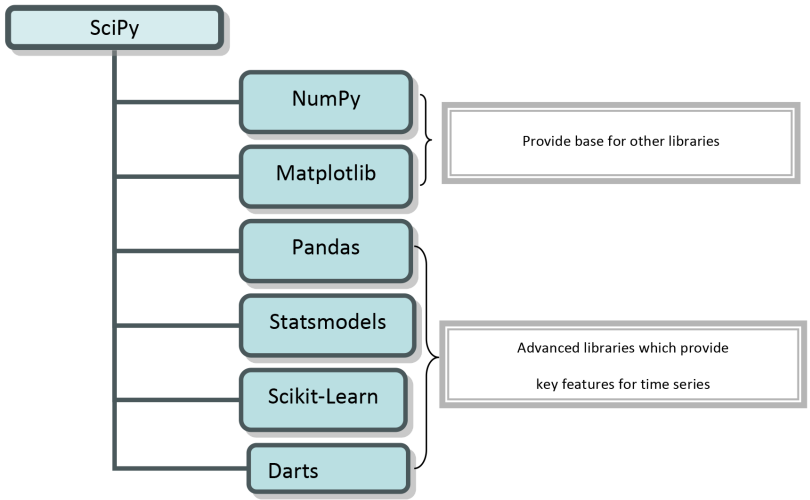

Figure 1. SciPy ecosystem libraries for time series forecast
Source: Own processing based on Lazzeri, 2021

- SciPy is developed in the open on GitHub, through the consensus of the SciPy and wider scientific Python community and is an open-source software for scientific computing package, developed openly and hosted on public GitHub repositories under the Scipy GitHub organization.
- NumPy - multidimensional data and mathematical functions working with the data;
- Matplotlib is a Python plotting library that produces publication-quality figures in a variety of hardcopy formats and interactive environments across platforms.
- Pandas - data manipulation and analysis;
- Statsmodels is a Python module that provides classes and functions for the estimation of many different statistical models as well as for conducting statistical tests and statistical data exploration. The results are tested against existing statistical packages to ensure that they are correct (Lazzeri, 2021).
- Scikit-Learn - data modeling;
- Darts is a Python library for user-friendly forecasting and anomaly detection on time series.

Among the above-mentioned ecosystem, there are many more essential Python libraries, such as:

- TensorFlow - distributed numerical computation using data flow graphs;
- Keras - python wrapper library, can be built independently on top of TensorFlow;
- Natural Language Toolkit (NLTK) with its lexical resources FrameNet, WordNet, Word2Vec;
- Spark MLlib - computing scaling;
- Thano - scientific computing on a large scale;
- MXNet - fast model training.

For the purpose of the current paper, the DARTS open-source data forecasting library has been used to predict the inbound tourist flows to Bulgaria for 2023 based on time series data from 2005 to 2022 with DML architectures by Python, in a computer-generated environment – Jupyter Notebook.

2. LITERATURE REVIEW

Artificial Intelligence - AI, machine learning - ML, data mining, big databases and smart data are just some of the many buzzing trends that have taken a dominant position in science, business and media in recent years (Egger, 2022). Although these increasingly popular phrases may seem catchy, they still have an attractive force, mainly because they have undoubtedly infiltrated our daily lives, whether in the personal or professional sphere. The rapidly advancing digitalization of our society has laid the foundations for this (Neuburger et al., 2018) increasing computing power, greater storage capacity, faster internet connections, the rapid development of powerful algorithms, and the availability of vast amounts of data for the purposes of analysis are only some of the driving forces that have and continue to allow us to apply new analytical methods and generate useful knowledge for science, business and ultimately society as a whole (Egger, 2022; Skiena, 2017). Furthermore, the application of emerging technologies in the post-COVID-19 era must adapt to changes in consumer behavior (perceptions of risk, last-minute bookings, desires for advance bookings in new contexts in museum exhibitions, need for highly personalized tourist packets) and likely changes in interaction mode (from physical touch to voice or from input to automated discovery) (Gretzel et al., 2020).

Today, we all witness how sensor devices monitor everything that can be monitored: video streams, social media interactions, and generally the position of anything that moves. Cloud computing allows us to use the power of a huge number of machines to systematize, process and analyze this data. It's hard to imagine, but hundreds of computers kick in every time we search Google or any metasearch engine on the Internet, scrutinizing all of our previous activity just to decide which is the best ad to generate based on our search history. The result of all this is the birth of data science, a new field dedicated to maximizing the value of vast databases of information. As a discipline, data science sits somewhere at the intersection of statistics, computer science, and machine learning, but it is building its own strength and character (Skiena, 2017). Furthermore, data science is focused on quantitative data collection and interpretation by the use not only of statistics but also by the application of scientific methods, processes for data systemization, visualization and analysis to extract meaningful insights for business.

According to Encyclopedia Britannica AI, is the ability of a digital computer or computer-controlled robot to perform tasks commonly associated with intelligent beings (Copeland, 2023). AI is a broader concept of machines' capability to perform tasks that normally require human intelligence, such as understanding natural language, recognizing objects and sounds, and solving empirical problems. And by machines, we don't just mean the physical robots, but also the software running on, for example, your computers, phones and connected home devices. In addition, AI research areas include rule-based reasoning, ML and DML where, additional layers, complex neural architectures subject to ML techniques are added, processing of natural language - Natural Language Processing (NLP), computer vision, speech analytics and robotics (Egger, 2022). AI can be seen as a key driver of innovative solutions for businesses of all sizes and industries (Mich, 2020), including tourism.

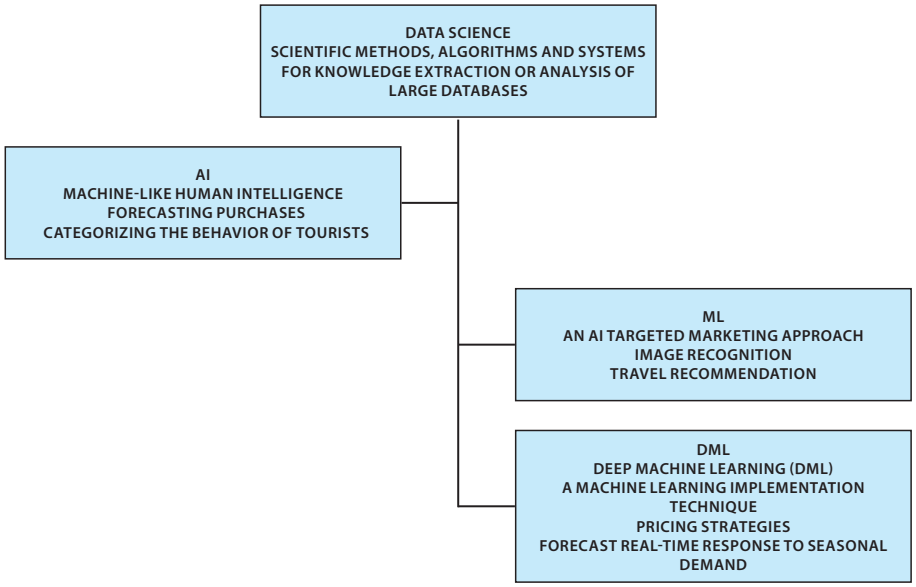

Figure 2. Data science and AI impact on tourism
Source: Own processing

Data science lies at the intersection of computer science, statistics, and their applied domains. From computer science comes machine learning and high-performance computing technologies to handle large-scale computation. From statistics comes a long tradition of exploratory data analysis, significance testing, and visualization. From the fields of application in business and science come challenges and evaluation standards to assess when they have been adequately overcome (Skiena, 2017). These principles and techniques are widely applied across functional areas in business as well as tourism. Probably the most extensive business applications are in marketing, for target marketing tasks, online advertising and cross-selling referrals. Data science is also applied in overall customer relationship management to analyze customer behavior to manage human capital and maximize expected customer value (Provost & Fawcett, 2013).

According to some researchers, data science, as an important aspect of AI, includes a comprehensive set of methods, algorithms and systems that are applied in various sectors of an interdisciplinary field (Egger, 2022), such as tourism. By observing Figure 2, the infernal deduction can lead us to the conclusion that data science combines computer science, mathematics and statistics, as well as tourism domain-specific knowledge, in order to obtain valuable information from large sets of structured, semi-structured and unstructured data (Egger, 2022). This therefore helps to explain and understand tourism phenomena and processes in the present and, with its predictive power, to some extent, future forecasts as well. Thus, computer languages such as Python are pivotal, especially for the current paper's aim achievement when the object is a data analysis with a forecasting task.

Python is an object-oriented programming-interpreted language. Python uses code modules that are interchangeable instead of a single long list of instructions that was standard for functional programming languages. Python libraries are called "modules". These modules provide commonly used functionality in the form of different objects or functions. For example, there is a module that has functions you can use in your code to test if files exist on your hard drive; there are modules that have functions for implementing web server, or web-browser functionality; there are modules to work with images; there are modules to create charts and graphs; there are modules to parse XML or HTML files; etc. To achieve the present research aims the Darts Python module/ library was applied.

Scientist working with time series already knows that time series are special creatures. If you possess regular tabular data, you can often just apply Scikit-learn to do most ML operations from preprocessing to prediction and model selection. Nevertheless, with time series, they should be different. Furthermore, the time series task can require one library for pre-processing (e.g. Pandas to interpolate missing values and re-sample), a different one to detect seasonality (e.g. Statsmodels), a third one to fit a forecasting model (e.g. Facebook Prophet), and what is more a backtesting and model selection routines must be performed as well. Such a process can be quite tedious, as most libraries need different APIs and data types. Also, in cases involving more complex models based on NN, or issues involving external data and additional dimensions the task can be more time-consuming. If that is the case a self-made model or so-called "use-case" should be implemented r, for instance using libraries such as Tensorflow or PyTorch. On the other hand, the Darts Python library can be seen as an attempt to smooth the end-to-end time series machine learning experience in Python.

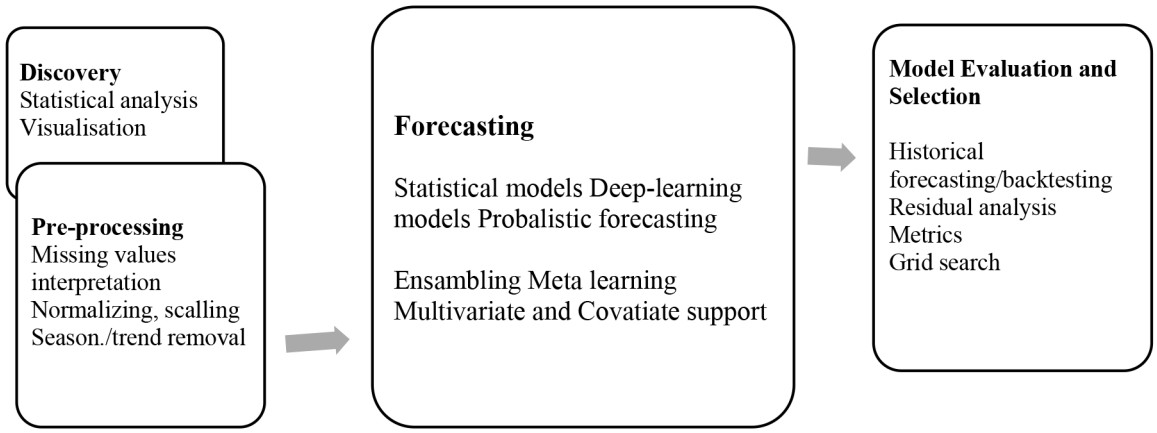

Figure 3. DARTS learning experience

Source: Own processing based on Herzen et al. (2022)

The library allows forecasting model applications in the same way similar to scikit-learn, using fit() and predict() functions (Herzen et al., 2022). The library developers also consider that by applying it the following processes are pretty straightforward - to backtest models, combine the predictions of several models, and take external data into account. What is more, the library supports both univariate and multivariate time series and models (Herzen et al., 2022). All ML or DML-based models can be trained on potentially large datasets containing multiple time series, by applying the library and what is more, some of the models give good support for probabilistic forecasting. Darts also offers extensive anomaly detection capabilities. For example, it is trivial to apply Python Outlier Detection models on time series to obtain anomaly scores, or to wrap any of Darts forecasting or filtering models to obtain fully fledged anomaly detection models (Herzen et al., 2022). Therefore, applying the Darts library is considered less time-consuming, more accurate data processing, better model pre- and fully-processing and last but not least simultaneous backtests performed together.

3. RESULTS AND DISCUSSION

Challenging time series forecasting problems have been an issue for scientists and practitioners for several decades. Thus, we have implied some novel automated methodologies that combine classical methods for time series forecasting in tourism with up-to-date models performed in a Python web-based platform - Jupiter Notebook (https://jupyter.org/) using a taxonomy and

framework of questions. We have applied naive, classical, machine learning and deep learning forecasting methods in order to estimate which of them is the most suitable and has better parameters for the task of forecasting in tourism, namely the overnight stays in the Bulgarian accommodation facilities by month for the period 2005 - 2023. For accurate task performance, we have chosen a specific user-friendly library of Python – Darts, which has been developed particularly for forecasting and anomaly detection on time series (Herzen et al., 2022).

The input data is the number of overnight stays registered in Bulgaria for the period 2005 to 2022 and our output data is the prediction made by the models which have been compared with the real-time data - overall of 216 observations. The data was obtained via the websites of The National Statistical Institute of the Republic of Bulgaria (2023) and Eurostat - the statistical office of the European Union (Eurostat, 2023).

From the python environment, a TensorFlow environment was activated and then Jupyter Notebook environment was created. All the bellow mentioned modules were generated with the functions from () and import ():

```
import math
import numpy as np
import pandas as pd
from pandas import read_csv

import matplotlib.pyplot as plt

import scipy.stats
import scipy.optimize
import scipy.spatial
from statsmodels.tsa.seasonal import seasonal_decompose
import tensorflow as tf
tf.autograph.experimental.do_not_convert
from darts import TimeSeries
from keras.preprocessing.sequence import TimeseriesGenerator
from sklearn.model_selection import train_test_split
from tensorflow.keras.layers import Input, Dense, LSTM
from tensorflow.keras.models import Sequential
from sklearn.metrics import mean_squared_error
from math import sqrt
from tensorflow.keras.optimizers import Adam
from tensorflow.keras.callbacks import TensorBoard, ModelCheckpoint
```

Then the data was imported into the Jupyter Notebook environment:

```
df = pd.read_csv('Bulgaria.csv', sep = ";", parse_dates = True)
```

and then the data itself was observed via the pandas library:

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 216 entries, 0 to 215
```

```
Data columns (total 2 columns):
 #   Column    Non-Null Count  Dtype
---  ------    --------------  -----
 0   Month     216 non-null    object
 1   arrivals  216 non-null    int64
dtypes: int64(1), object(1)
memory usage: 3.5+ KB
```

Following this, for the forecasting task, the time series were split (Figure 4) as the last 36 months of data were used for the prediction:

```
series1, series2 = series[:-36], series[-36:]
series1.plot()
series2.plot()
```

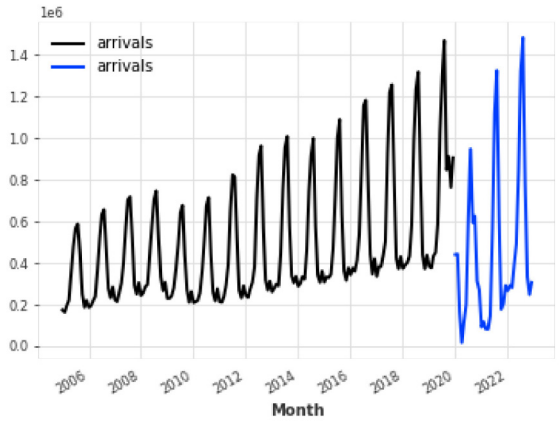

Figure 4. Time series overnights stay per month train test split

Source: Own processing

Following the data observation, the models for the forecast were created and performed, to evaluate the best-performing model, namely

- ExponentialSmoothing - Holt-Winters Exponential Smoothing, is used for time series forecasting when the data has linear trends and seasonal patterns.
- TBATS - Trigonometric seasonality, Box-Cox transformation, ARMA errors, Trend and Seasonal components.
- AutoARIMA - Automatically discover the optimal order for an ARIMA model.
- Theta - is a simple forecasting method that involves fitting two θ -lines, forecasting the lines using a Simple Exponential Smoother, and then combining the forecasts from the two lines to produce the final forecast.

For model estimation, the Mean Absolute Percentage Error (MAPE) was used as it is quite convenient and scale-independent for empirical experiment purposes since measures the average magnitude of error produced by a model, or how far off predictions are on average. In Darts it is a simple function call:

```
from darts.metrics import mape
```

The results were generated as follows:

```
model ExponentialSmoothing() obtains MAPE: 49.96%
model TBATS() obtains MAPE: 56.45%
model AutoARIMA() obtains MAPE: 77.88%
model Theta() obtains MAPE: 51.13%
```

Since the best performing model so far is the Exponential Smoothing with MAPE value of 49,96% which means that the average absolute percentage difference between the predictions and the actuals is nearly 50% a better result may be achieved with the probabilistic forecast with Monte Carlo samples describing the distribution of the time series values with simple Exponential Smoothing model. Here the MAPE was estimated at 49.34%.

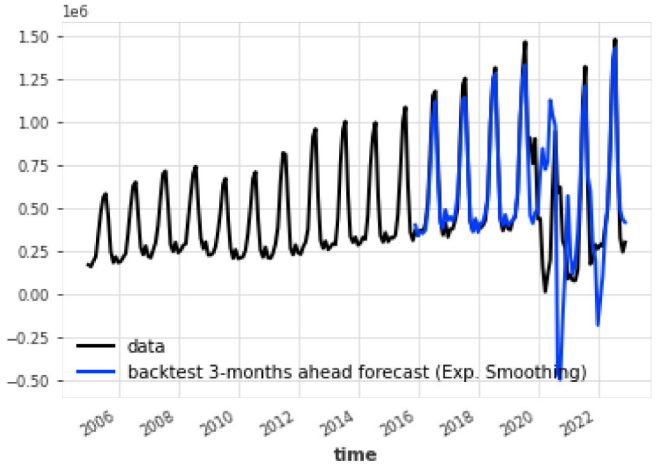

Figure 5. Monte Carlo simple Exponential Smoothing model

Source: Own processing

The unsatisfactory results from the above-described models had initiated further empirical tests this time with more sophisticated models based on ML and DML. The first one tested was the Long Short-Term Memory Neural Network (LSTM). LSTSM is a variety of recurrent neural networks (RNNs) that are capable of learning long-term dependencies, especially in sequence prediction problems. LSTM has feedback connections, i.e., it is capable of processing the entire sequence of data.

```
RNNModel(model=LSTM, hidden_dim=20, n_rnn_layers=1, dropout=0, training_length=20, batch_size=16, n_epochs=100, optimizer_kwargs={'lr': 0.001}, model_name=Nights_RNN, log_tensorboard=True, random_state=42, input_chunk_length=14, force_reset=True, save_checkpoints=True)
```

Here the error metrics are different, namely the loss error as with ML and DML the main objective is to minimize the loss to evaluate the model performance. The estimation came as: train_loss=0.00142, val_loss=0.0335, and thus our model based on the LSTM had generalized the results good which means it can be applied to another forecasting task of the type described above. For model evaluation, the training horizon was expanded and a large portion of the data set – from 2016, was used as again the MAPE was the error metric. The model was tested to predict from 12 to 96 months in advance from that point and the best results were described in Table 1 below:

Table 1. The model

Prediction horizon	MAPE in %
12 months	8.63
24 months	7.35
36 months	6.77
48 months	15.56
60 months	26.81
72 months	34.74
84 months	35.49
96 months	35.49

Source: Own calculations

An up-worth progression of the MAPE can be observed especially when the COVID-19 pandemic started which can explain the model's bigger evaluation error. On the other hand, due to the small volume of the time series, respectively the small validation set the overfitting of the data was inevitable.

4. FUTURE RESEARCH DIRECTIONS

Since the results from the DML model performance are satisfying to a large extent a further more sophisticated state-of-the-art test can be performed bearing in mind the need for a larger data set for a better model performance and more accurate evaluation. Such can be the application of another library, e.g. Facebook Prophet, or another model, such as the NBEATS Model.

5. CONCLUSION

The article demonstrated the usage of ML and DML Python models for forecasting tourism data tasks. As demonstrated by the results an assumption that DML models outperform the basic and probabilistic forecast can be made which means that all the benefits from forecasting with the Darts Python library can be applied to other datasets with the same success. Furthermore, forecasting with AI development should be observed and more experiments with innovative models and libraries must be performed for scientific clarification and precise estimation. The ML and DML for tourism forecast purposes are on the verge of transformation and AI for science development together with the COVID-19 pandemic are one of the main catalyzers of this process. Another push in this direction can be considered all industry stakeholders involvement in AI, ML and DML in additional tourism operations, development and practical application.

References

Copeland, B. (2023, October 11). *Artificial intelligence*. *Encyclopedia Britannica*. https://www.britannica.com/technology/artificial-intelligence

Egger, R. (2022). *Applied Data Science in Tourism*. (R. Egger, Edd.) Springer Nature. https://doi.org/10.1007/978-3-030-88389-8

Eurostat. (2023). *Arrivals at tourist accommodation establishments - monthly data.* Retrieved from Tourism Industries - monthly data: https://ec.europa.eu/eurostat/databrowser/view/TOUR_OCC_ARM__custom_6400112/default/table

Géron, A. (2019). Hands-On Machine Learning with Scikit-Learn and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems (Vol. 1). O'Reilly Media. ISBN: 9781492032649

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep Learning.* MIT Press. ISBN-13: 978-0262035613

Gretzel, U., Fuchs, M., Baggio, R., Hoepken, W., Law, R., Neidhardt, J., Pesonen, J., Zanker, M., & Xiang, Z. (2020). e-Tourism beyond COVID-19: a call for transformative research. *Information Technology & Tourism, 22*(2), 187-203. https://doi.org/10.1007/s40558-020-00181-3

Herzen, J., Lässig, F., Piazzetta, S. G., Neuer, T., Tafti, L., Raille, G., & Grosh, .. &. (2022). Darts: User-Friendly Modern Machine Learning for Time Series. *The Journal of Machine Learning Research, 23*(1), 5442-5447. Retrieved from https://unit8co.github.io/darts/

Lazzeri, F. (2021). Python Open Source Libraries for Scaling Time Series Forecasting Solutions. Data Science at Microsoft. https://medium.com/data-science-at-microsoft/python-open-source-libraries-for-scaling-time-series-forecasting-solutions-3485c3bd8156

McCulloch, W. S., & Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. *The bulletin of mathematical biophysics, 5*. https://doi.org/10.1007/BF02478259

Mich, L. (2020). Artificial Intelligence and Machine Learning. *Handbook of e-Tourism*, 1-21. https://doi.org/10.1007/978-3-030-05324-6_25-1

The National Statistical Institute of the Republic of Bulgaria. (2023). *Tourism.* Retrieved from Business statistics: https://www.nsi.bg/en/content/1847/tourism

Neuburger, L., Beck, J., & Egger, R. (2018). Chapter 9 The 'Phygital' Tourist Experience: The Use of Augmented and Virtual Reality in Destination Marketing. *Tourism Planning and Destination Marketing*, 183-202. https://doi.org/10.1108/978-1-78756-291-220181009

Provost, F., & Fawcett, T. (2013). Data Science and its Relationship to Big Data and Data-Driven Decision Making. Big Data, 1(1), 51-59. https://doi.org/10.1089/big.2013.1508

Skiena, S. S. (2017). The Data Science Design Manual. *Texts in Computer Science*. https://doi.org/10.1007/978-3-319-55444-0

Turing, A. M. (1936). On computable numbers, with an application to the Entscheidungsproblem. *Journal of Math*, *58* (345-363), 5.