

# Mustafa Murat Arat

STATISTICIAN · DATA SCIENTIST

Ankara, Turkey

✉ arat.murat@gmail.com | 🌐 mmuratarat.github.io | 📧 mmuratarat | 📄 mustafa-murat-arat | 🌐 mmuratarat | 📧 murat.arat

## SUMMARY

PhD-level educated, bilingual statistician and data scientist, who design, develop and implement advanced predictive methods, powered by machine learning and deep learning algorithms, in order to provide actionable insights from large volumes of real-world, structured and unstructured data in order to satisfy the business needs for decision making. My research interests include but are not limited to the following areas: machine learning and deep learning algorithms, supervised learning, and unsupervised learning methods, time series forecasting and modeling, big data, distribution theory, statistical analysis, optimization and operational research.

During my Ph.D. process, I completely switched from MATLAB/R to Python and obtained working knowledge of, and hands-on experience in building Recurrent Neural Networks (RNN) (Long Short-term Memory (LSTM) and Gated Recurrent Units (GRU)), Convolutional Neural Networks (CNN), Autoencoders and other types of deep neural network architectures in Deep Learning frameworks, i.e., low-level TensorFlow and its high-level Python API Keras. I also have a deep understanding of best practice data analytics procedures to solve complex business problems using various Python packages, such as numpy, pandas, SciPy on JupyterLab, and data visualization tools such as Matplotlib and Seaborn in addition to Scikit Learn, and StatsModels for other sophisticated, statistical modeling projects (including some classification/regression algorithms such as Logistic Regression, Lasso Regression, Ridge Regression, Support Vector Machines (SVM), Decision Trees (tree-based algorithms) and Ensemble Learning (Bagging, Boosting, Stacking, Random Forest), some dimension reduction methods, e.g., Principal Component Analysis (PCA), and some clustering algorithms, i.e., K-Means and Mixture Models with EM Algorithm and others). As a team, we have extensively used structured query language (PostgreSQL) to extract data from and version control systems, such as Github to collaborate. At the same time, I constantly try to improve myself and learn new emerging technologies, tools, and platforms to add new skills to my skill set. Recently, I am acquiring skills in Apache Spark with Python (PySpark / Spark MLlib) and Hadoop's MapReduce, two distributed computing platforms for building ETL processes; Amazon Web Services (AWS) for cloud computing; Docker / Swarm / Kubernetes which are used for containerized microservices and orchestration systems; Tableau, an interactive business intelligence tool for data visualization and Elastic Search (plus Kibana - data visualization dashboard for Elasticsearch), a distributed full-text search and analytics engine for schema-free JSON documents, which is based on Apache Lucene.

I believe that I have a team-oriented, can-do attitude, strong enthusiasm and a deep curiosity about the intersection of business and technology. I am passionate, highly motivated, innovative, energetic, reliable, inclusive and pro-diversity, and have excellent written and verbal communication skills.

## EDUCATION

### Doctor of Philosophy in Analytics

*Knoxville, TN, U.S.A.*

DEPARTMENT OF BUSINESS ANALYTICS & STATISTICS, HASLAM COLLEGE OF BUSINESS, UNIVERSITY OF TENNESSEE

*August 2015 – May 2020*

- GPA: 3.92 / 4.00
- Coursework includes: Prescriptive Analytics, Statistical Inference, Probability/Stochastic Process, Bayesian Modeling and Computations, Data Mining Methods and Business Applications, Applied Multivariate Methods, Image Reconstruction, Reinforcement Learning, Applied Time Series, Probability and Mathematical Statistics, Multivariate Data Mining Techniques.
- Completed my dissertation "Advances and Applications in Deep Learning" under the supervision of Dr. Michel Ballings (mb@utk.edu).

### Master of Science in Statistics

*Ankara, Turkey*

DEPARTMENT OF STATISTICS, INSTITUTE OF NATURAL SCIENCES, HACETTEPE UNIVERSITY

*August 2011 – June 2014*

- GPA: 3.93 / 4.00 (ranked 1st of 15)
- Coursework includes: Linear Models, Time Series, Stochastic Processes, The Analysis of Contingency Tables, Simulation Techniques, Multivariate Statistical Methods, Survival Analysis.
- Completed my master thesis titled "A Study on Support Vector Machines" under supervision of Prof. Dr. Turhan MENTES (mentes@hacettepe.edu.tr).

### Bachelor of Science in Statistics

*Ankara, Turkey*

DEPARTMENT OF STATISTICS, FACULTY OF SCIENCE, HACETTEPE UNIVERSITY

*September 2007 – June 2011*

- GPA: 3.24 / 4.00 (top 10%, ranked 9th of 91)
- Coursework includes: Categorical Data Analysis, Biostatistics, Econometrics, Time Series, Experimental Design, Multivariate Statistical Methods, Stochastic Processes, Sampling, Survey Design, Operation Research, Actuary, Regression Analysis, Probability, Mathematical Statistics, Nonparametric Statistical Methods, Statistical Decision Making, Linear Algebra, Calculus, (Advanced) Mathematics.

## HONORS & AWARDS

2009–2010 **Supreme Success Award**, Faculty of Science, Hacettepe University

*Ankara-Turkey*

2015–2020 **Scholarship/Assistantship**, Department of Business Analytics & Statistics, Haslam College of Business, University of Tennessee

*Knoxville, TN, U.S.A*

Fall 2015 **Statistics Excellence Fund**, Department of Business Analytics & Statistics, Haslam College of Business, University of Tennessee

*Knoxville, TN, U.S.A*

## SKILLS

<b>Programming Languages/Libraries</b>	Python, Numpy, Pandas, scikit-learn, statsmodels, psycopg2, spaCy, NLTK, OpenCV and Matplotlib.
<b>Deep Learning Libraries</b>	Tensorflow, and Keras.
<b>Container and Orchestration Systems</b>	Docker, Swarm, Kubernetes.
<b>Database Systems</b>	PostgreSQL
<b>No-SQL/Full-text Search Systems</b>	Elasticsearch/Kibana
<b>Big Data Frameworks</b>	Hadoop's MapReduce and Apache Spark
<b>Scripting Languages</b>	R, and MATLAB.
<b>Visualization Software</b>	Tableau.
<b>Statistical Softwares</b>	SPSS, and Minitab
<b>Version Control</b>	Git/Github
<b>Markup Languages</b>	TeX, Beamer, Jekyll, and Markdown/R-Markdown.
<b>Miscellaneous</b>	Extensive usage of Jupyter Lab and comfortable with Linux/Unix command line.
<b>Languages</b>	Turkish (Native Language), English (Full professional proficiency), French (Elementary proficiency)

## PROJECTS

- **The Short-term Prediction of Electrical Load Demand of a Power System**

Electricity load forecasting has always been the essential part of efficient power system planning and operation. For this study, we applied Support Vector Regression (SVR) to short-term forecast electrical load of the city Brabant of the Netherlands using hourly data, comparing it with two well-known forecasting models, Nonlinear Auto-regressive Network with External Input and Double Seasonal Exponential Smoothing. SVR yields the lowest Mean Squared Error and the highest coefficient of determination.

- **Building and Training Long Short-term Memory Networks One Layer at a Time**

Long short-term memory (LSTM) is an artificial recurrent neural network architecture capable of learning order dependence in sequential data. A common LSTM unit is composed of 4 different gates with each having its own weight matrices and biases. Even though initial weights are simply initialized usually according to some random distribution, e.g., Xavier Initialization, sometimes, a custom weight initialization might be needed. However, Tensorflow's built-in LSTM function (tf.keras.layers.LSTM) does not allow for custom weight initialization. Therefore, we created our own LSTM cell implementation using the source code of Tensorflow. Thus, we can freeze an LSTM layer and initialize next layer with the previously trained weights, or we can use the pre-trained weights in a subsequent model. By doing so, we had three main advantages: (i) we could build the network from scratch as deep as needed in real time, (ii) speed: adding each new layer amounts to training a shallow network with only one hidden layer; and (iii) resilience to overfitting. The process of freezing old layers and inserting new ones is repeated until additional layers cease to improve performance. This indicates that it's time to stop adding new layers and consider the network complete. So we avoid the pitfalls of backpropagation, including its high computational cost and its struggle to effectively adjust deep parameters. This also substantially decreases training time.

- **Sales Elasticity of Emotional Displays**

Partnered with a Direct-response TV Network, we study over 46 million frames of video (over 12,792 hours of video footage) in which a salesperson makes her/his sales pitch to customers. We use state-of-the-art artificial intelligence technology (transfer learning of a Haar-feature based cascade classifier and mini-Xception model) to detect the presence of a human face and then extract salesperson's facial expressions from the unstructured data, and then match each salesperson's moment-by-moment display of emotions to customers' purchase data to understand the true nature of the relationship. This research contributes to marketing theory and practice in five ways: (1) It applies state-of-the-art AI technologies to detect human faces and extract emotional displays, and formulate new models to estimate the sales impact of emotional displays in the presence of marketing mix variables; (2) It establishes that the elasticities of a salesperson's emotional displays are uniformly negative, including that of happiness, which is a provocative finding because it partially contradicts the external validity of the contagion theory. (3) It introduces the sales impact of marketing mix activities (i.e., product, price, display duration, and free shipping) in the emotional displays literature; (4) It marks the first empirical study to show that the presence of a salesperson's face in the video frames is not only positively associated with sales, but also increases sales by 0.61%; and (5) It offers guidance to firms on re-training of sales personnel to support the implementation of "selling with a straight face" as a maxim for their sales professionals.

- **Learning High-Cardinality Categorical Features with Sparse Updates**

We define high dimensional categorical variables as data with an intractable number of categories, lacking intrinsic ordering to them, with prime examples being ZIP-codes, product identifiers, and customer identifiers. These data pose a challenging problem in neural networks considering that two dominant methods are one-hot encoding and continuousification of a categorical variable, with prime example being mean target encoding. For this study, we seek a general-purpose approach for statistical analysis on categorical entries that is suited to a very large number of categories without dependency to the distribution of the target variable and without any human intervention and unrealistic expansion of feature matrix, that might lead to amount of computational and memory requirements. Using the key-value store idea, we propose an algorithm which is low dimensional, thus it is very efficient in processing time and memory and it can be computed with online learning setting in any Neural Network architecture. Besides, it learns a weight per categorical level, which can differentiate between every unique categories.

## PROFESSIONAL EXPERIENCES

### Research Assistant

DEPARTMENT OF STATISTICS, FACULTY OF SCIENCE, HACETTEPE UNIVERSITY

Ankara, Turkey

October 2011 – present

- Teaching a wide range of undergraduate-level courses from Bayesian Statistics to Probability Theory.
- Consultant for undergraduate students.
- ERASMUS program coordinator.
- Managing seminars in the department.

### Graduate Teaching Assistant

DEPARTMENT OF BUSINESS ANALYTICS & STATISTICS, HASLAM COLLEGE OF BUSINESS, UNIVERSITY OF TENNESSEE

Knoxville, TN, U.S.A.

August 2015 – July 2020

- Assisting students during the lab sessions and marking written reports of the assignments for courses such as Categorical Data Analysis, Customer Analytics, Applied Multivariate Methods, Applied Time series and many more.
- Co-creating a Deep Learning class for the first time in the faculty with my advisor Dr. Michel Ballings, covering a wide range of topics.

## ONLINE CERTIFICATES

1. **The Complete SQL Bootcamp by Jose Portilla**, Certificate can be seen here. August 20, 2019
2. **Spark and Python for Big Data with PySpark by Jose Portilla**, Udemy.com - Certificate can be seen here. December 9, 2019
3. **Tableau 10 A-Z: Hands-On Tableau Training For Data Science! by Kirill Eremenko**, Udemy.com - Certificate can be seen here. December 20, 2019
4. **Bash Shell Scripting: Crash Course For Beginners by Francesco Santi**, Udemy.com - Certificate can be seen here. April 01, 2020
5. **Docker Mastery: with Kubernetes +Swarm from a Docker Captain by Bret Fisher**, Udemy.com - Certificate can be seen here. May 07, 2020
6. **Python for Time Series Data Analysis by Jose Portilla**, Udemy.com - Certificate can be seen here. June 23, 2020
7. **Taming Big Data with Apache Spark and Python - Hands On! by Frank Kane**, Udemy.com - Certificate can be seen here. June 23, 2020
8. **Complete Guide to Elasticsearch by Bo Andersen**, Udemy.com - Certificate can be seen here. July 8, 2020

## ONLINE PUBLICATIONS

I also have a passion for data science advocacy and education. Being able to share what I know brings me a great deal of joy, and therefore I started a blog on my home page <https://mmuratarat.github.io>, which accelerated my learning in data science, bolstered my ability to communicate technical content to a general audience, and mentor other people.

Selected online publications can be seen below:

- **Dimensions of matrices in an LSTM Cell** - <https://mmuratarat.github.io/2019-01-19/dimensions-of-lstm>
- **Cross Entropy for Tensorflow** - <https://mmuratarat.github.io/2018-12-21/cross-entropy>
- **Tensorflow Metrics - Accuracy/AUC** - <https://mmuratarat.github.io/2018-12-20/tf-metrics-acc-auc>
- **Backpropagation Through Time for Recurrent Neural Network** - <https://mmuratarat.github.io/2019-02-07/bppt-of-rnn>
- **Batch Normalization** - <https://mmuratarat.github.io/2019-03-09/batch-normalization>
- **How to use embedding layer and other feature columns together in a network using Keras?** - <https://mmuratarat.github.io/2019-06-12/embeddings-with-numeric-variables-Keras>
- **Ensemble Learning in extensive details with examples in Scikit-Learn** - <https://mmuratarat.github.io/2019-08-06/ensemble-learning-details-examples>
- **Regularized Linear Models** - <https://mmuratarat.github.io/2019-09-01/regularized-linear-models>
- **Support Vector Machines in theoretical details** - [https://mmuratarat.github.io/2019-10-17/svm\\_in\\_details](https://mmuratarat.github.io/2019-10-17/svm_in_details)
- **Inverse Transform Method for Continuous Distributions and Sampling from Normal Distribution** - <https://mmuratarat.github.io/2019-10-02/inverse-transform-and-box-muller>
- **The Expectation-Maximization (EM) in theoretical details** - [https://mmuratarat.github.io/2019-11-13/em\\_in\\_details](https://mmuratarat.github.io/2019-11-13/em_in_details)
- **Using PySpark to connect to PostgreSQL locally** - <https://mmuratarat.github.io/2020-06-18/pyspark-postgresql-locally>

## TRANSLATIONS

1. Lipschutz, S. Lipson, M. (2011). Schaum's Outline of Probability (2nd edition), New York, NY, U.S.A.: The McGraw-Hill Company; Danacıoğlu, N., **Arat, M.M.**, Gürünlü Alma, Ö., Çankaya, E., Olmuş H., Kardiye, F. & Khaniyev, T. Schaum Serisi-Olasılık (İkinci Basımdan Çeviri) September, 2013, Nobel Yayın Dağıtım, <https://www.nobelyayin.com/detay.asp?u=3893>.
2. Roussas, G. (2006). Introduction to Probability. California, U.S.A.: Elsevier Academic Press; Danacıoğlu, N., **Arat, M.M.**, Gürünlü Alma, Ö., Yüksel, G., Çankaya, E., Yurt Öncel, S., Kardiye, F., Olmuş H. & Dağalp, R. Olasılığa Giriş September, 2014, Nobel Yayın Dağıtım, <https://www.nobelyayin.com/detay.asp?u=4207>.

## POSTER PRESENTATIONS

### 6th International Congress on Psychopharmacology

Antalya, Turkey

GÖGCEGÖZ GÜL, I., ERYILMAZ, G., HIZLI SAYAR, G., ÖZTEN, E., TARHAN, N., **ARAT, M.M.**, "EVALUATION OF THE EFFICACY OF THE CONTINUATION ELECTROCONVULSIVE THERAPY IN TREATMENT-RESISTANT SCHIZOPHRENIA"

September 19 – 21, 2013

## COLLOQUIUMS/MEETINGS/CONFERENCES ATTENDEE

1. 8th International Statistics Students Colloquium, Izmir, Turkey, May 14 – 15, 2011.
2. The Joint Meeting of Young Business and Industrial Statisticians and Young Portuguese Statisticians, Lisbon, Portugal, July 23 – 26, 2012.
3. 9th International Statistics Days Symposium, Antalya, Turkey, May 10 – 14, 2014.

## PEER-REVIEWED PUBLICATIONS

1. Tan, O., Hizli Sayar, G.H., Ünsalver, B.Ö., **Arat, M. M.** & Karamustafalıoğlu, O. (2014). "The correlations of nicotine addiction with the levels of impulsiveness, depression and anxiety in obsessive-compulsive patients". *Journal of Dependence*, 15(3), 124 – 133.
2. Gogcegoz Gül, I., Eryilmaz, G., Hizli Sayar, G., Özten, E., **Arat, M.M.**, & Tarhan, N. (2014). "Evaluation of the Efficacy of The Continuation Electroconvulsive Therapy in Treatment-Resistant Schizophrenia". *Revista de Psiquiatria Clínica*, 41(4), 90 – 94
3. Eryilmaz, G. Gogcegoz Gül, I., Yorbik, O., & **Arat, M.M.** (2014). "Evaluation of Clinical Response According to Plasma Paroxetine Level in Paroxetine-Responsive Major Depression". *International Journal of Internal Medicine*, 3(3), 39 – 42.
4. Tan, O., Hizli Sayar, G., Ünsalver, B.Ö., **Arat, M. M.** & Karamustafalıoğlu, O. (2015). "Combining transcranial magnetic stimulation and cognitive-behavioral therapy in treatment resistant obsessive-compulsive disorder". *Anatolian Journal of Psychiatry*, 16(3), 180 – 188.
5. **Arat, M.M.**, Aktas, S. (2016). "Generalized Maximum Entropy Approach to Unreplicated Factorial Experiments". *Statistics and Its Interface*, 9(3), 295 – 302.
6. Bharadwaj, N., Ballings, M., Naik, P., Moore, G.M., **Arat, M.M.** (2019). "Sales Elasticity of Emotional Displays: Large Scale Evidence for Selling with a Straight Face". *Journal of Marketing*, Under Revision.
7. **Arat, M.M.**, Ballings, M. (2019). "Learning High-Cardinality Categorical Features with Sparse Updates". Ready to be submitted.

## ORAL PRESENTATIONS

### Applied Statistics 2012

Ribno (Bled), Slovenia

**MUSTAFA MURAT ARAT**, "TESTING EXPORT-LED GROWTH HYPOTHESIS: THE CASE OF TURKEY, 1961-2010"

September 23 – 26, 2012

### EURO-INFORMS Joint International Meeting: 26th European Conference on Operational Research

Rome, Italy

ELCİN ERGİN, **MUSTAFA MURAT ARAT**, CEM İYİGUN, İNCİ BATMAZ, "SHORT-TERM ELECTRICITY LOAD FORECASTING VIA NONPARAMETRIC PREDICTION METHODS"

July 1 – 4, 2013

## European Conference on Data Analysis by The German Classification Society (GfKl) and the French speaking Classification Society (SFC)

MUSTAFA MURAT ARAT, ELCIN ERGIN, "SHORT TERM LOAD FORECASTING USING  $\epsilon$ -SUPPORT VECTOR REGRESSION"

Luxembourg, Luxembourg

July 1 – 4, 2013

## The 13th Annual Conference of the European Network for Business and Industrial Statistics (ENBIS-13)

MUSTAFA MURAT ARAT, SERPIL AKTAS ALTUNAY, "GENERALIZED MAXIMUM ENTROPY APPROACH TO UNREPLICATED FACTORIAL EXPERIMENTS"

Ankara, Turkey

September 15 – 19, 2013

## y-BIS 2013: Joint Meeting of Young Business and Industrial Statisticians, sponsored by International Society for Business and Industrial Statistics (ISBIS) and European Network for Business and Industrial Statistics (ENBIS)

MUSTAFA MURAT ARAT, "COMPARISON OF SVM AND LS-SVM FOR REGRESSION"

Istanbul, Turkey

September 19 – 21, 2013

## Columbia University

MICHEL BALLINGS, NEERAJ BHARADWAJ, PRASAD NAIK, GEORGE MILLER MOORE, MUSTAFA MURAT ARAT, "BUT WAIT, THERE'S MORE! DEEP LEARNING OF SALES ELASTICITY OF SALES PITCHES. THEORY + PRACTICE IN MARKETING"

New York, NY, USA

May 18, 2019

## Interactive Marketing Research Conference

MUSTAFA MURAT ARAT, GEORGE MILLER MOORE, MICHEL BALLINGS, "MAXIMIZING INSIGHTS FROM CUSTOMER DATA STREAMS"

Houston, TX, USA

March 27 – 29, 2019

## INFORMS Annual Meeting 2019

MUSTAFA MURAT ARAT, MICHEL BALLINGS, GEORGE MILLER MOORE, "BREAKING THROUGH BARRIERS TO DEEP LEARNING ADOPTION IN CUSTOMER BEHAVIOR MODELING"

Seattle, WA, USA

October 20 – 23, 2019

## 2020 Winter AMA Academic Conference

NEERAJ BHARADWAJ, MICHEL BALLINGS, PRASAD NAIK, MILLER MOORE AND MUSTAFA MURAT ARAT, "PURCHASE IMPACT OF A SALESPERSON'S FACIAL EXPRESSIONS: LARGE-SCALE VIDEO ANALYSIS USING DEEP LEARNING"

San Diego, CA, USA

February 14 – 16, 2020

\*\*\*Underline means who presents the paper.\*\*\*

## SCIENTIFIC ACTIVITIES

### GIS 2013: Genç İstatistikçiler Sempozyumu (Young Statisticians Symposium)

MEMBER OF THE LOCAL ORGANIZING COMMITTEE

- <http://www.gis2013.hacettepe.edu.tr>

Ankara, Turkey

September 10 – 11, 2013

### AIK 2013: Araştırmacılar ve İstatistikçiler Konferansı (Researchers and Statisticians Conference)

MEMBER OF THE LOCAL ORGANIZING COMMITTEE

- <http://www.aik2013.hacettepe.edu.tr>

Ankara, Turkey

September 12 – 13, 2013

### y-BIS 2013: Joint Meeting of Young Business and Industrial Statisticians, sponsored by ENBIS (European Network for Business and Industrial Statistics) and ISBIS (International Society for Business and Industrial Statistics)

MEMBER OF THE INTERNATIONAL SCIENTIFIC COMMITTEE AND ORGANIZING COMMITTEE

- <http://ybis13.msgsu.edu.tr>

Istanbul, Turkey

September 19 – 21, 2013

### y-BIS 2019: Joint Meeting of Young Business and Industrial Statisticians, sponsored by ENBIS (European Network for Business and Industrial Statistics) and ISBIS (International Society for Business and Industrial Statistics)

MEMBER OF THE INTERNATIONAL SCIENTIFIC PROGRAM COMMITTEE

- <http://ybis2019.msgsu.edu.tr>

Istanbul, Turkey

September 25 – 28, 2019

## INTERSHIPS

---

**Turkish Statistical Institute**

WWW.TURKSTAT.GOV.TR

Ankara, Turkey

August 9 – August 27, 2010

## REVIEWS FOR

---

**Applied Soft Computing**

Impact Factor: 4.873

REVIEWER

- Imprint: ELSEVIER
- ISSN: 1568-4946
- Abstracting and Indexing: Science Citation Index Expanded, Scopus, CompuMath Citation Index, Current Contents/Engineering, Computing & Technology
- <http://www.journals.elsevier.com/applied-soft-computing/>

## PROJECTS

---

**Turkish Nourishment and Health Project (Türkiye Beslenme ve Sağlık Arastirmasi)**

Ankara, Turkey

RESEARCHER

August – September, 2013

- Project No: 013A601001
- by Republic of Turkey - Ministry of Health
- <http://www.saglik.gov.tr/>

## PROFESSIONAL SERVICES

---

**Turkish Statistical Association**

Ankara, Turkey

ASSOCIATE MEMBER OF "THE BOARD OF DIRECTORS"

2012 – 2014

- <http://www.turkistatistik.org>

## PROFESSIONAL AFFILIATIONS

---

**European Network for Business and Industrial Statistics (ENBIS)** member

**Bernoulli Society for Mathematical Statistics and Probability** member

**Institute of Mathematical Statistics (IMS)** member

**International Society of Business and Industrial Statistics (ISBIS)** member

**International Society for Bayesian Analysis and its Industrial Statistics Section (ISBA)** member

**American Statistical Association (ASA)** member

**The Institute for Operations Research and the Management Science (INFORMS)** member