JAIME ARBOLEDA CASTILLA

Data Scientist | Mathematician | Software Engineer

Born in 1985 in Algeciras (Province of Cádiz, Spain).

I love gaining and sharing insights from data, as well as using those insights to make better decisions. I am always passionate about learning new things and I am able to do it fast. I am interesting in research and open problems, keeping up with new approaches in Machine Learning. Lately I have been very interested in Bayesian modeling and causal inference, although I have mostly worked on predictive modeling. I like programming and contributing to the open source community. I am easy to work with and I have always maintained good relationships with all my coworkers.

PROFESSIONAL EXPERIENCE

2023 | 2021

Data Scientist

European Commision — Remote work

On Safety and Security Analytics (SSA) team in ICS2 Project, TAXUD

Description

- Researching, designing and continuous developing of SSA models.
- Supporting deployment and orchestration of SSA models and analytic solutions for real time use.
- Supporting designing, organizing and monitoring of the SSA entire workflow.
- Developing Similarity Search and Anomaly Detection capabilities.
- Providing training for Public Officers across all Member States.

Projects

- Similarity search engine. It's built on top of several custom components, including a new neural network architecture based on Transformers.
- Model comparison tool for performance analysis, built with Dash and Docker.
- Synthetic XML generation.
- XML Splitter tool, which is a custom pipeline for transforming a complex XML into a Pandas Dataframe, with many options.

Technologies

- Dataiku, RStudio, JupyterLab.
- Denodo, Oracle, PostgreSQL, Neo4j.
- SQL, Python, R, Bash, XML, JSON, HTML.
- Numpy, Pandas, Keras, Tensorflow, Dash.
- GitLab, Jenkins, Docker, Kubernetes, Apache Kafka.

CONTACT INFO

blog
github
linkedin
email

SKILLS

- Machine learning and Deep Learning
- Data Analysis and Statistics.
- Bayesian Inference.
- Optimization.
- Structured approach to problem solving.
- Critical thinking.
- Easy to get along with.
- Python, R, Spark, Java, Rust.
- Tidyverse and tidymodels frameworks.
- Numpy, Pandas, Scikit-Learn, Tensorflow and Keras frameworks.
- Git and Docker.

2021
|
2020

Head of IT Unit and Data Scientist

Spanish Tax Agency — Madrid, Spain

On Subdivision of Information Analysis Technologies and Fraud Investigation

Description

- Spanish Public Officer belonging to the Senior Corps of Systems, Information and Communication Technologies of the State Administration.
- Data Scientist and manager of a team of 8 people working on analytic models for tax fraud detection.

Projects

- Custom (based on KNN) clustering algorithm that, using the information of purchases and sales of each corporation, predicts whether its declared sector of economic activity is accurate or not. It was programmed in Scala.
- Identification of statistical position and undervaluation of goods fraud in Customs declarations by modifying an existing algorithm provided by European Commission.
- Classifier (using XGBoost) for predicting whenever a taxpayer is most likely to make a mistake when modifying some parts of its draft of Personal Income Tax Declaration. The goal was to provide a nudge message to those taxpayers in the event of modification, aiming at reducing filling errors.
- Classifier for predicting the risk of non-payment of debts with the Tax Agency, in order to anticipate precautionary measures.
- Classifier for predicting the risk of not paying its tax liabilities in due time for a given taxpayer. This model makes use of near real time information regarding all invoices collected in the previous months of the prediction.
- Regression model for predicting the total (declared or undeclared) incomes of a given family using all available information.

Technologies

- Python, Scala, Spark.
- Pandas, Numpy, Scikit-Learn, Xgboost, Luigi.
- Linux, Cloudera.
- SyBase IQ, DataStage.

2019 | 2017

Head of IT Unit

Spanish Tax Agency — Madrid, Spain

On Subdivision of Software Development and Applications

Description

- Spanish Public Officer belonging to the Senior Corps of Systems, Information and Communication Technologies of the State Administration.
- Manager of a team of 15 people working on SW development for Personal Income Tax and applications.

Projects

- Web service for personal data ingestion, for the web app of Personal Income Tax Declarations.
- Cryptographic service for granting access credentials for the presentation of the Income Tax Declarations.
- Datawarehouse ingestion of Personal Income Tax Declarations.
- Risk analysis (combining rule-based risks, statistical risks and simple predictive models) for Personal Income Tax Declarations.
- Software development for the management and lifecycle of Personal Income Tax Declarations.

Technologies

- COBOL, Java, HTML, JavaScript.
- Web Services.
- DB2, Oracle.
- Z/OS, Linux.
- SyBase IQ, DataStage.

2017 | 2013

Head of IT Unit

Spanish Tax Agency — Madrid, Spain

On Subdivision of Software Development and Applications

Description

- Spanish Public Officer belonging to the Senior Corps of Systems, Information and Communication Technologies of the State Administration.
- Manager of a team of 8 people working on SW development for Corporate Tax and applications.

Projects

- Datawarehouse ingestion of Corporate Tax Declarations.
- Risk analysis (combining rule-based risks, statistical risks and simple predictive models) for Corporate Tax Declarations.
- Software development for the management and lifecycle of Corporate Tax Declarations.

Technologies

- COBOL, Java.
- DB2, Oracle.
- Z/OS, Linux.
- SyBase IQ, DataStage.

TEACHING EXPERIENCE

2023

Big Data

BBVA — Remote work

Professor of Data Scientist Fundamentals Course, delivered to 23 workers of BBVA from South America. The course had a duration of 42 days (168 hours), and covered, among other topics:

- Big Data tools in BBVA (Datio, Stratio, Crossdata).
- Python.
- Data Wrangling with Numpy and Pandas.
- Data Visualization with Matplotlib and Seaborn.
- Machine Learning with scikit-learn.
- Big Data with Spark and SparkSQL.
- Machine Learning with SparkML.
- Deep Learning with PyTorch.

I enjoyed the experience a lot, and I received very positive feedback from all the students (every student ranked the instructor with the maximum score).

2022 | 2020

Mathematics Teacher

Academia Castiñeira — Madrid, Spain

I regularly give a course on Mathematical Methods (Differential Equations, Harmonic Analysis and Complex Analysis) for students of third course of Aeronautic Engineering.

2021

Lecturer

Webinar — Universidad Complutense, Madrid, Spain

I gave a lecture in the Fiscality and Artificial Intelligence Webinar to present a project held at the Tax Agency, related to using the Nudge philosophy to boost voluntary tax compliance.

2020 | 2017

Mathematics Teacher

Academia Montero Espinosa — Madrid, Spain

I gave multiple courses to students of the Degree of Mathematics, including:

- Complex Analysis.
- Topology.
- Advanced Algebra.
- Statistics.
- Probability.
- Galois Theory.
- Integration and Measurement Theory.
- Physics.

PUBLICATIONS

2021

Nudge Project

Paper — Aranzadi Thomson Reuters

Paper on Nudge project, carried out at the Tax Agency, consisting of the application of Artificial Intelligence to help the assistance to the taxpayer and voluntary compliance with tax obligations. Published in Aranzadi Thomson Reuters, along with other works presented in the Webinar “Fiscality and Artificial Intelligence” organized by the Complutense University of Madrid.

OPEN SOURCE COLLABORATIONS

2023

Collaborator of category_encoders

category_encoders — Remote work

I fixed an issue related to the compatibility between this library and `sklearn` when using composed objects (`Pipeline` and `ColumnTransformer`). My pull request was merged in the main branch.

2022

Collaborator of Keras

Keras — Remote work

I raised an issue after finding a bug, and weeks later I was able to fix it with a pull request that was merged on the main branch.

2021

Developer of open source library

Nested Cross Validation — Remote work

Python package that performs hyperparameter optimization, train, probability calibration and error validation of classification models using a Nested Cross-Validation approach.

EDUCATION AND TRAINING

2023

DataTalksClub

Zoomcamp — Remote

- Data Engineering Zoomcamp

2022 | 2021

European Commission Training

Internal training — Remote

- Cibersecurity
- Software Development and Agile Methodologies.

2021 | 2013

Spanish Tax Agency Training

Internal training — Madrid, Spain

- Advance Data Analysis.
- Machine Learning and Big Data.
- Geospatial Data Processing in R.
- Software Development and Agile Methodologies.
- Automatic Reporting tools.
- Data Warehousing.
- Blockchain and Cryptocurrencies.
- OSGI for Java development.
- SCRUM and Agile methodologies.
- Advanced Java programming.

2022 | 2017

Coursera

Courses and Specializations — Remote

- Probabilistic Graphical Models: Representation
- Probabilistic Graphical Models: Inference
- Probabilistic Graphical Models: Learning
- Bayesian Statistics: From Concept to Data Analysis.
- Bayesian Statistics: Techniques and Models.
- Bayesian Statistics: Mixture Models.
- Neural Networks and Deep Learning.
- Improving Deep Neural Networks: Hyperparameter Tuning, Regularization and Optimization.
- Structuring Machine Learning Projects.
- Convolutional Neural Networks and Computer Vision.
- Sequence Models and Natural Language Processing.
- Machine Learning.

2013 | 2012

National Institute of Public Administrations (INAP)

Selective course — Madrid, Spain

Course for accessing the Senior Corps of Systems, Information and Communication Technologies of the State Administration.

- Passed the competitive examination to the Corps in the first attempt and with the second best place of 600 candidates.
- In the internship phase, I worked on an analysis of the security systems of the Ministry of Defense.

2016 | 2014

Universidad Española de Educación a Distancia (UNED)

Master's Degree in High School Teacher Training — Madrid, Spain

- Average mark of 8.1.
- 215 hours of teaching internships at high school.
- Finished with a work on teaching Mathematics using programming tools.

2010 | 2009

Universidad Complutense de Madrid (UCM)

Master's Degree in Mathematical Research — Madrid, Spain

- Average mark of 8.8.
- Carried out under the direction of Mr. Francisco Presas with a research work in geometric quantization.

2009 | 2005

Universidad Autónoma de Madrid (UAM)

Double Degree in Mathematics and Computer Science — Madrid, Spain

- Average mark of 9.5 and 25 Honor Distinctions.
- I studied additional courses out of the curriculum.

HONORS AND AWARDS

2011 | 2007

Consejo Superior de Investigaciones Científicas (CSIC)

Scolarships — Madrid, Spain

- I was granted a scholarship oriented to the realization of doctoral thesis with CSIC, stopped for personal reasons after obtaining the Master's Degree in Mathematical Research.
- I was awarded to be part of a program to expand studies in Mathematics during summers. This allowed me to attend intensive courses during the summers of 2007 and 2008 at the Complutense University of Madrid.

2009

Universidad Autónoma de Madrid

Honorable mention — Madrid, Spain

Winner of honorable mention to the student with the overall best marks during the degree.

2009 | 2004

Comunidad de Madrid

Scolarships for outstanding academic performance — Madrid, Spain

- Course 2008/2009: Teaching supprt on Differential Geometry.
- Course 2007/2008: Work on numerical methods for solving equations.
- Course 2006/2007: Work on Neural Networks.
- Course 2005/2006: Work on Commutative Algebra.
- Course 2004/2005: Work on cryptography and number theory.