JACQUELINE ARAYA

DATA SCIENTIST

London, UK | 077-6131-4054 | ja3076@columbia.edu | jacquelinearaya.github.io | arayajacqueline

A researcher-turned data scientist eager to transform my analytical thinking and experience in applied research into actionable methods to tackle business problems using data as the vehicle to get insights, predictions, and recommendations that speak to every business stakeholder

EDUCATION

M.Sc. in Data Science | Columbia University **Sept 2018 – May 2020 | New York, USA**

- Capstone Project: "Asset tracking: identifying patterns in spatio-temporal trajectory datasets" in collaboration with GE Research Centre

Professional Engineer - Industrial Engineering | University of Chile **Jul 2013 – Mar 2015 | Santiago, Chile**

- Thesis: "Quantitative Analysis of scientific virtual communities of practice"
- Dean's list in 2012 and 2013 for academic excellence

BSc. in Engineering Science | University of Chile **Mar 2008 – Jun 2013 | Santiago, Chile**

SKILLS

Excellent: Python (Numpy, Pandas, TensorFlow, Keras, Scikit-Learn, Matplotlib, NLTK, Web Scraping), R (ggplot, tidyverse, RMarkdown), SQL – Google Cloud Platform: BigQuery Proficient: PySpark, Tableau, Git, Linux Power Shell Familiar: Hadoop File System (HDFS), JavaScript, Stata, MATLAB, Java

PROFESSIONAL EXPERIENCE

Research Associate | Columbia Business School **Aug 2016 – Oct 2018| New York, USA**

- Worked with the Management Division faculty over two years on the analytical and technical side of applied research projects related to organizational behaviour, entrepreneurship, firm strategy, and labour market dynamics.
- Conducted the experimental design and implementation of an A/B test on 15,000 job seekers from Indeed tracking the impact on engagement using automated Python and bash scripts.
- Collected 6.5m player/match records of the e-sports game Dota2 from Steam API, using distributed computing, bash, and Python scripts.
- Scraped data from 10 static and non-static websites using Python (Selenium - BeautifulSoup) creating structured data sets.
- Translated Java code into Python to calculate similarity measures between artists for a music dataset of 10,000 artists and 115,000 songs.
- Analysed survey data set of 606 MBA students using a multinomial logistic regression model to study the impact of different choices when facing job prospects upon graduation.
- Mentored research assistants in the use of Linux-based high-performance servers and helped interns and Ph.D. students to debug code and translate technical requirements into coding implementation.
- Worked closely with the IT Director in the promotion of a data driven culture within the research community by introducing new Linux-based tools.

Project Analyst | Nic Chile Research Labs (Chile's DNS research centre) **Jun 2012 – Mar 2015 | Santiago, Chile**

- Effectively secured $40,000 in public funding for the development of high-quality measurement software for mobile Internet.
- Forged a partnership with one of Chile's largest Internet service providers to develop quality measurement software for mobile Internet.
- Collaborated with industry stakeholders in the design of creative business models for software prototypes.

Digital Marketing Intern | Bank of Credit and Investment (BCI) **Summer 2012 | Santiago, Chile**

- Performed market segmentation of customers based on its overdraft credit behaviour to improve decision making on the marketing mix.
- Designed a new marketing email campaign targeted to different segments that achieved a 3% lift in credit usage.
- Built a data visualization dashboard in Excel of 5 credit products with forecasted versus actual sales and automated for weekly update with direct reporting to area manager.

DATA SCIENCE PROJECTS

Data processing using Spark - Wikipedia Database | Columbia University **Mar 2020 | New York, USA**

- Used Google Cloud Platform cluster to launch a Spark script over HDFS (Hadoop File System) using Python to clean and parse the Wikipedia database to generate a webgraph of its internal links and use it as input to a Spark PageRank algorithm script to generate the ranking of each Wikipedia page.

Movie recommendation system model using Spark ML | Columbia University **Dec 2019 | New York, USA**

- Used Spark machine learning to build a recommendation system in Python designing and testing different experiments with collaborative filtering models to better predict the last rating of every user for a database of 20M ratings (27,000 movies and 138,000 users).

Neural network classification model | Columbia University **Dec 2019 | New York, USA**

- Built a Python end-to-end pipeline that take photos as input and classify them by predicting its correct label using a deep learning model (CNN architecture) that train and test on a dataset of photos taken by me.

RELEVANT COURSEWORK

M.SC. IN DATA SCIENCE	Algorithms for Data Science, Statistical Inference, Machine Learning, Deep Learning, Data visualization, Computational systems and Databases, Product Recommendation, Causal Inference
PROFESSIONAL ENGINEER - INDUSTRIAL ENGINEERING	Statistics, Applied Econometrics, Operations Management, Marketing Strategy, Marketing Analytics, Consumer Behaviour, Financial Analysis, Financial Engineering, Data Mining, Data Warehousing, Industrial Organization, Information and Communication Technologies
BSC. IN ENGINEERING SCIENCE	Single and Multivariate Calculus, Linear Algebra, Computer Science, Probability and Statistics, Modelling and Optimization, Operations Research, Economics, Microeconomics, Macroeconomics