Amir Behbehani, Ph.D.

236 West 10th, New York, NY, 10014
United States

617-433-9772 / nycAlgos@gmail.com
amirbehbehani (LinkedIn) ♦ amir.path2 (Skype)

Summary

Data Scientist with 10+ years experience developing machine learning algorithms to drive revenue, increase efficiency, and enhance the portfolio of intellectual property within the enterprise. Specialize in building and managing data pipelines, and analyzing data using Python (NumPy, SciPy, pandas, scikit-learn), Apache Spark (SparkSQL, MLlib), TensorFlow, Postgres, MATLAB, Mathematica, and developing algorithms and developing software. Extensive experience with advanced mathematics, statistics, and applied machine learning, as well as presenting and visualizing complex concepts to diverse audiences.

Education

Massachusetts Institute of Technology – Ph.D., Computational Economics
Thesis: Price Optimality in Bipartite Markets Exhibiting Network Externalities
Stanford University – B.S., Mathematics

Experience

Lead Data Scientist, Founder – Serial Metrics*, San Francisco, CA* 01/2010 – Present

- Developed a first-to-market platform that translates English questions into **SQL** queries, enabling non-technical users to generate ad-hoc reports and insights from relational data without having to write SQL code. Built using **Python** and leverages a **Bayesian engine** to fine-tune query recommendations and leveraged 15 years' background in **natural language processing (NLP)**.
- Designed and patented an automated **ETL** (extract, transform, and load) system, able to ingest data from disparate databases, infer foreign key relationships, and join data sets to form a canonical training set for subsequent algorithm development. I wrote the core engine in **C++** and developed a series of proprietary, graph-theoretic algorithms to traverse a relational schema to create a "master join table," serving as a training set for subsequent predictive modeling.
- Developed and integrated an auto-parametric modeling framework for a point-and-click **machine learning**-based back-end system.
- Developed a semantic engine based on the Cilibrasi Normalized Google Distance metric, which determines the cognitive similarity of terms from the cooccurrence rate in Google search queries, solving the long-tail of search terms, increasing the horizontal scalability of the product by ensuring search queries could process all English synonyms.
- Interviewed, hired, and mentored the engineering team.

Data Scientist – Medibio*, San Francisco, CA* 08/2016 – 01/2017

- Developed a method to track circadian data from an Apple Watch in order to determine if the device wearer is at risk for a variety of mental health illnesses, including clinical depression, bi-polar disorder, anxiety, and dementia.
- Used **Python** libraries to perform time-domain to frequency-domain transforms before subjecting data to **random forest models** to generate differential diagnoses.
- Wrote a proprietary app to transform EDF (European Data Format) to CSV files.

Data Scientist, Founder – Zuli.io (Acquired by Google)*, San Francisco, CA* 03/2014 – 06/2016

- Designed and patented a system to use wireless **RSSI** values collected from smart electrical plugs within a house to detect the location of each person and automatically turn on and off appliances in a smart home.
- Implemented a **BLE** communication system between hardware plugs, a plug control and historical energy usage UI.
- Prototyped the machine learning code in Python and converted it to **Objective C** to load into a proprietary iPhone app.
- Final predictive models use **random forest** and **multinomial logistic regression** to produce predictions hovering around 99.99% accuracy.
- Patent information: https://patents.google.com/patent/US20160323393A1
- Model summary: https://drive.google.com/file/d/1v7ujn2Vv2yiy4uliO

Consulting Data Scientist – Jobr (Acquired by Monster)*, San Francisco, CA* 06/2013 – 01/2014

- Developed a **random forest** model to match job seekers and job posters by presenting candidates and recruiters, respectively, members of the other population that (1) they are most likely to be interested in, and (2) are most likely to reciprocate that interest.
- Model summary: https://drive.google.com/drive/u/0/folders/1l0QP4Z

Data Scientist – Identified.com*, San Francisco, CA* 01/2012 – 06/2013

- Developed a geo-inferencing system, able to infer the location of users that had not explicitly stated their location, using **Facebook user data** ingested from **Hadoop** and leveraging a set of **random forest models** trained on those data to predict user locations within a metropolitan statistical area.
- Generated complex SQL queries to convert relational data sets into a **dyadic data format (graph data set)** to ensure the statistical model *incorporated interaction effects, as opposed to the more traditional statistical methods of controlling for interaction effects.*
- Leveraged Census Bureau data sets to cluster **Lat/Long** data in accordance to a defined set of **Metropolitan Statistical Areas** to create a predictive set of user locations.
- Once deployed, the production system was able to take 700,000 Facebook accounts with known location data, and infer the location of 50 million Facebook accounts with no corresponding location data.
- The deployed models performed at approximately 95% accuracy.
- Model summary: https://drive.google.com/file/d/1Kw9W7t2K-BjYK_B1B

Data Scientist – Zipongo*, San Francisco, CA* 01/2011 – 01/2012

- Developed a **topic modeling framework,** written in **R, to structure the unstructured nutrition data,** which I extracted from FDA data sets, and **API**'s.
- Created a set of algorithms leveraging a variety of techniques including **multinomial logistic regression, support vector machine** and **random forest** to classify food ingredients as "healthy," "neutral," or "unhealthy.”

Economist – Eventbrite*, San Francisco, CA* 01/2010 – 01/2011

- Conceived, architected, and developed the dynamic pricing system for Eventbrite.
- Once implemented, Eventbrite realized an increase of up to 700X ticket revenues.
- Patent information: http://www.patentsencyclopedia.com/app/20120316924

Skills

Programming Languages: Python, C/C++, Java, Perl, JavaScript, DCPL
Databases: MS SQL Server, Oracle, HBase, Amazon Redshift, MS SQL
Statistical Methods: Hypothetical Testing, Exploratory Data Analysis (EDA), Confidence Intervals, Bayesian Analysis, Principal Component Analysis (PCA), Dimensionality Reduction, Cross-Validation, Auto-correlation
Machine Learning: Regression analysis, Naïve Bayes, Decision Tree, Random Forests, Support Vector Machine, Neural Network, Sentiment Analysis, Collaborative Filtering, K-Means Clustering, KNN, CNN, RNN and Ada Boosting
Data Visualization: Tableau, MatPlotLib, Seaborn, ggplot2, d3.js
Packages: ggplot2, caret, dplyr, Rweka, gmodels, RCurl, tm, C50, twitter, NLP, Reshape2, rjson, plyr, pandas, numPy, seaborn, sciPy, matplot lib, scikit-learn, Beautiful Soup, Rpy2, sqlalchemy
Hadoop Ecosystem: Hadoop 2.x, Spark 2.x, MapReduce, Hive, HDFS, Pig
Cloud Services: Amazon Web Services (AWS) EC2/S3/Redshift
Reporting Tools: Tableau Suite of Tools 10.x, Server and Online, Server Reporting Services(SSRS), MS Office (Word/Excel/Power Point/ Visio)
Version Control: Tools SVM, GitHub