Amir Reza Aghamousa Farashi, Ph.D. | Curriculum Vitae

Principal Data Scientist, **Coupang (www.coupang.com) LLC in USA (Feb. 2018~present)**
Tower 730, Songpadae-ro 570, Songpa-gu, Seoul, South Korea
+82-10-7480-0824 | amir.aghamousa@coupang.com

Job experience:

- Postdoctoral researcher, statistical modeling & data mining in cosmology (May 2016~Feb. 2018)
  Affiliation: Korea Astronomy and Space Science Institute (KASI), South Korea
- Postdoctoral researcher, statistical modeling & data mining in cosmology (April 2014~April 2016)
  Affiliation: Asia Pacific Center for Theoretical Physics (APCTP), South Korea

Academic Qualifications:

- Doctor of Philosophy in Physics (Feb/2008 ~ Sept/2013), University of Pune, India
- Master of Science in Physics (First Class) (July/2001 ~ Nov/2004), University of Pune, India
- Bachelor of Science in Physics (Sept/1993 ~ July/1999) Ferdowsi University of Mashhad, Iran

Skills:

- Machine learning (classification, regression, clustering), Deep Learning, statistical modeling
- Deep understanding in statistics (Bayesian and Frequentist), mathematics and numerical analysis
- Professional programmer in Python and R and programmer in Scala, Big data and distributed computing (Spark), SQL, Fortran, Shiny: web application framework
- Data analysis, visualization, exploratory data analysis, social network analysis, natural language processing, image classification, web scraping
- Cloud computing and AWS services (S3, EC2)
- Strong problem-solving skills
- Excellent communication skills. Proficient in communicating complex concepts with others
- Ability in team management to archive the main goal efficiently
- Linux, Windows, LaTeX and hardware computer

Projects:

- **Improving search advertisements** *(www.coupang.com)*
  I have built the User Feedback (UF) model, a simple model that utilizes the users' clicked data (in www.coupang.com) to assign the advertisement keywords to the products to propose the most relevant advertised products to customer's queries in the search result page. After training and testing carefully the model it has shown a high performance with more than 90% precision. 89% item coverage along with a query coverage from Top to Tail queries range up to rank 1M. The experiment successfully deployed to the Search production with +9% contribution in the increment of the main metric i.e. First Placement Ad GMV per customer and other important metrics like GMV per customer +3.9% and Ad GMV per Customer +15.03%.
- **Search tags cleansing** *(www.coupang.com)*
  In an online shop like www.coupang.com, Search tags are the keywords added by sellers to the products. This extra information can improve search results by customers however wrong search tags have a destructive effect on search results. Using fastText (NLP library), I have trained a model that used different features of products like title, brand, category, etc to classify the associated search tag as relevant or irrelevant to the product. The model shows 80% precision and 50% recall which is a valuable result for cleansing tags and improving search results.
- **Unsafe image detection** *(www.coupang.com)*
  Some products have images that are not suitable for some customers. These include nudity or sexual images or those related to adult products. Using some labeled product images in Coupang database, I have made a fine-tuned model based on EfficientNet to classify images into Safe and Unsafe. In addition, I have leveraged the efficiency by making a combination of fine-tuned EfficientNet and NudeNet model. The combination shows a precision more than 80% for the fashion category. This project is in progress now to be employed for other categories of products.
- **Modeling the risk of bank loans** *(http://www.aghamousa.com/credit-risk-modeling/)*
  Using logistic regression, decision tree and random forest I have identified the risk of bank loans based on German credit data set with 24 variables. Assuming a 15% acceptable bad rate I have made a logistic model

with 14 variables that shows 76% accuracy with 79%, 68% and 0.44 for specificity, sensitivity and kappa values respectively which is comparable with the results of alternative random forest model.

- **Scientific network analysis** *(http://www.aghamousa.com/scientific-network-analysis/)*
  I have investigated the strong scientific connection between a sample of 31 astronomers working in South Korea based on their collaborative publications. Using graph analysis I have shown/visualized many characteristics of society such as components, path length distribution and ranking astronomers and universities based on the number of collaborators, power of connections and international connections.
- **Supernova classification**
  Using the Gaussian process and the other statistical tools we have designed an algorithm to classify the supernova type Ia based on the associated light curves (collecting data via web scraping). The proposed algorithm shows around 80% accuracy and is applicable for poor light curves even with 5 data points.
- **Gaussian process for model consistency test** *(https://arxiv.org/abs/1705.05234)*
  Using the marginalized likelihood in the Gaussian process to check the consistency of cosmological models with data. In this approach, we consider the hyper-parameters values of covariance kernel function and associated likelihood to assess the level of consistency.
- **Time delay analysis in time series** *(https://arxiv.org/abs/1603.06331)*
  Designing a novel statistical algorithm based on cross-correlation and kernel regression using R, for time delay estimation of sparse and non-equispaced time series of strong lens systems. Participating in Time Delay Challenge including more than 5000 pairs of time series, our algorithm results in just 5% bias on average with fast processing time ~ 2 minutes per pair light curves (using normal PC). An updated version of the algorithm has improved bias with 87%!.
- **Nonparametric regression** *(https://arxiv.org/abs/1412.3552)*
  We have deeply developed REACT a nonparametric regression technique and used for CMB power spectrum estimation and cosmological model selection using data from WMAP and Planck surveys. Our development contains risk minimization related to the weighted loss function, calculating weight matrix for any given basis and using in non-equispaced data.

Publications:

Available here