Linxiao Bai

E-mail: linxiaobai3@gmail.com

Website: https://pagliaccl.github.io/

Tel: 217-305-1947

Education

University of Rochester Sept.2016-May.2017

- MS of Data Science, CS concentration.
- GPA: 3.9
- Enrolled with 30% tuition waiver

University of Illinois at Urbana Champaign June.2013-Dec.2015

- BS of Statistics, CS focus.
- GPA: 3.7
- Distinction Honor.
- *John E. Gieseking* Scholarship

Working Experience

Monsanto, Genomic Prediction Nov.2017-Now

Data Scientist, Pipeline Architect

Focus 1, Genome Wide Recommendation System

- In charge of designing and implementing a recommendation system that utilizes genomic information to guide breeding advancement decisions. The system also incorporated genomic variability and "*breeding value*" to the final prediction score.
- Conducted massive hyper-parameter tuning for GBDT+LR and DNN based prediction models.
- Constructed multi-variate regression model using MCMC (pymc3, theano-GPU).
- Responsible for designing and implementing a data QC algorithm to production. The algorithm finds the best subset that maximize the normality of the training.
- Serve as a subject matter expert to implementing multi-processing parallelism for massive model training in cloud. Reducing the training time/ AWS budget by 5 times.

Focus 2, Pipeline Architecture Design

- In charge of the development and maintenance of data driven modeling pipelines on AWS using Apache Airflow.
- Design and implemented a dynamic pipeline for production deployment. It is dynamically updated and allows data scientist to quickly deploy new models and pipelines by simply interacting with a flask UI.
- Implemented a flask UI for pipeline automation. It uses AWS s3 hook at the back end to interact with a series of production configuration stored in the cloud.

University of Rochester Medical Center July.2017-Now

Research Associate

- Publication: *Ann Marie White, Linxiao Bai, Christopher Homan, "Does Reciprocal Gratefulness in Twitter Predict Neighborhood Safety? Comparing 911 Calls Where Users Reside or Use Social Media," AAAI International Conference on Web and Social Media (ICWSM), Stanford, CA, June 2018.*

- In charge of data driven researches focusing on public safety and community welfare.
- Serve as a subject matter expert to experiments design and machine learning modeling in the projects. Such experiments and models are used to discover important factors to predicting community safety index.
- Responsible for large-scale (1 million) geocoding using Spark parallelism and ArcGIS on a Linux cluster.
- Implemented an interactive visualization of geo-tagged tweets and 911 emergency calls using google map API and Echarts (a JavaScript charting library).

---

iFlytek, CloudPlatform/ Computational Advertising Jan.2016-Aug.2016

Data Scientist/ Spark Software Engineer

Focus 1, Anti-Fraud System with Spark-MLlib:

- Serve as a subject matter expert in the designing and developing of the Anti-Click-Fraud System of *iFlytek Ad Exchange Platform, Demand Supply Platform*.
- Designed a click-pattern likelihood prediction model using Bayesian network and LSTM. The model uses likelihood to evaluate the probability of a sequential actions.
- Trained and deployed an auto-encoder-based fraud detection system.
- Designed and implemented an IP address blacklist based on click-pattern regression, and Poisson hypothesis testing.
- In charge of constructing and maintaining MAC address blacklists based on illegal hardware information rules.
- Designed a series of online A-B testes to validate performances of strategies.
- Proposed two influential indexes that quantify the performance of the Anti-Fraud system. The indexes are still being used in the company.

Focus 2, Data Mining, Machine Learning, CTR prediction:

- Contributed as a subject-matter expert to data mining tasks of *iFlytek Data Management Platform* using Spark.
- In charge of parameter tuning and feature designing of the CTR prediction model, GBDT+LR and Deep Learning.
- In charge of GPS trajectory mining with clustering algorithm using Spark-MLlib to discover user living space.
- In charge of constructing supervised learning algorithms for gender and “heavy-gamers” predictions.
- Constructed user-label matching strategies using keyword/NLP approaches on Spark cluster. Audience include: influenza population, middle class, new car-owners, and so on.

Technical Skills

- Experienced with Python and ML packages (pandas, sklearn, keras, pymc3), familiar with R.
- Expert in Apache Airflow DAG design and AWS s3 hook.
- Experienced with Spark and its job optimizing and resource saving.
- Familiar with Docker, MySQL, AWS-CLI, YARN-hdfs.
- Dynamic visualization with html and JavaScript charting libraries.
- GitHub version control.