Linxiao Bai

E-mail: lbai2@ur.rochester.edu

Website: https://pagliaccl.github.io/ Tel: 217-305-1947

Education

- **University of Rochester**, MS of Data Science, CS concentration. Sept.2016-May.2017
  GPA: 3.90 (Enrolled with 30% tuition waiver)
- **University of Illinois, Champaign**, BS of Statistics, CS focus, Distinction Honor. June.2013-Dec.2015
  GPA: 3.69 (*John E. Gieseking* Scholarship)

Work Experience

University of Rochester Medical Center, Saunders Research Lab July.2017-Now

Research Assistant

- In charge of data driven researches focusing on public safety, and development of a community crime alert system using geo-tagged tweets and python NLP package.
- Visualizing large-scale geo-tagged tweets (18 millions) and emergency calls (1 million) data by overlapping their heat-map layers using google earth/ Echarts (html based plotting library) and creating dynamic word-cloud pictures.
- In charge of a large-scale (1 million) geocoding and real-time GPS mapping project using Spark 2.0 parallelism and ARCMAP on large scale Linux cluster.
- Negotiating pipelining/ API with the mobile-app developing team to come up with solutions to transferring data, updating models, and code version control.

iFlytek, CP BigData Team/ Computational Advertising Team Jan.2016-Aug.2016

Data Scientist/ Spark Software Engineer

Focus 1, Anti-Fraud System, Online Advertising. Using Apache Spark on AWS to:

- Design and developed the Anti-Click-Fraud system for *iFlytek AD Exchange Platform, Demand Supply Platform*.
- Design and implemented the first IP blacklist rules based on click-time patterns and Poisson hypothesis testing.
- In charge of constructing and maintaining hardware blacklists based on illegal hardware information detection.
- Design a series of online A-B testes to validate performances of strategies.
- Compose the first Anti-Fraud system performance evaluation report.
- In the report, proposed two influential indexes that quantify the performance of the Anti-Fraud system. The indexes are still being used in the industry.

Focus 2, Data Mining, Machine Learning, Online Advertising:

- Contributed as a subject-matter expert to data mining projects of *iFlytek Data Management Platform using Spark*.
- In charge of GPS trajectory mining strategies with clustering algorithm using Spark-MLlib. The goal is to predict user live spaces.
- In charge of constructing supervised learning algorithms used for gender and “heavy-gamers” predictions.
- Constructed user-label matching strategies using keyword/NLP approaches on Spark cluster. Audience include: traveling people, Influenza population, working class, breeding mothers, new car owners, and so on.
- Contributed to the composing of *iFlytek BigData White Book,* a thorough documentation about company’s overall data condition and possible data mining directions that are supported by the data.

Huishang Bank, Data Analytics Lab June.2015-Aug.2015

Intern Analyst

- Designed a linear predictive model for recommending new financial products to targeted customer using R.
- Responsible for analyzing predictors significance in the linear model. The result of important predictors is being used in the industry to help designing financial products and explaining customer interests.
- Responsible for day-to-day data cleaning, summary statistical analysis using R and other exception handling skills.
- Frequently analyze significant data index and compose weekly data report.

Research Projects

- VisualDX Anti-Crawler System Development, Using Deep Learning to Detect Traffic Suspiciousness.
- Electricity Abnormal Usage Detection, Feature Engineering of Time Series Data and Classification.
- Predict NYC Ground Traffic Using Cab Trajectory, A Deep Learning Application.
- American Domestic Airline Network Analysis, An Improvement of the Configuration Model.
- Change Point Detection of Probabilistic Model in Time Series Data, An Application of EM Algorithm.

Technical Skills

- Python, R, MySQL, Java, Matlab.
- YARN-hdfs based Spark with Scala, Python.
- Spark job effectiveness optimizing. YARN resource saving.
- Spark SQL fast query and Zeppelin BD visualization.
- Maven project management, GitHub version control.