**Kai Cui**

**Summary:**

* Highly efficient data scientist and quantitative researcher with 7+ years of experience in Machine Learning, Data mining with large data sets of structured and unstructured data, data acquisition, data validation, predictive modeling and data visualization, with ability to build/create new opportunities for organizations.
* Strong ability to analyze sets of data for signals, patterns, ways to group data to answer questions and solve complex data puzzles
* Proficient in advising on the use of data for compiling personnel and statistical reports and preparing personnel action documents.
* Work experience in analytics, working with data to convert large volumes of structured and unstructured data into actionable insights and business values.
* Ability to analyze raw data, drawing conclusions and developing recommendations.
* Proven experience in hypotheses development, identification of patterns within data, analyzing data and interpreting results.
* Skilled in Advanced Regression Modeling, Time Series Analysis, Statistical Testing, Correlation, Multivariate Analysis, Forecasting, Model Building, Business Intelligence tools and application of Statistical Concepts.
* Proficient in: Data Acquisition, Storage, Analysis, Integration, Predictive Modeling, Logistic Regression, Decision Trees, Data Mining Methods, Forecasting, Factor Analysis, Cluster Analysis, ANOVA, Neural Networks and other advanced statistical and econometric techniques.
* Adept in writing code in Python and R to manipulate data for data loads and extracts.
* Proficient in data entry, data auditing, creating data reports & monitoring data for accuracy.
* Ability to extract Web search and data collection, Web data mining, Extract database from website, Extract Data entry and Data processing.
* Extensively worked on using major statistical analysis tools such as Python, R, SQL and MATLAB.
* Strong knowledge in all phases of the SDLC (Software Development Life Cycle) from analysis, design, development, testing, implementation and maintenance with timely delivery against deadlines.
* Good knowledge and understanding of data mining techniques like classification, clustering, regression techniques and random forests.
* Extensive experience with creating MapReduce jobs, SQL on Hadoop using Hive and ETL using PIG scripts, and Flume for transferring unstructured data to HDFS.
* Strong ability to work with Mahout, for applying machine learning techniques in Hadoop Ecosystem.
* Strong Oracle/SQL Server programming skills, with experience in working with functions, packages and triggers.

**Professional Skills:**

* **Operating systems**-Windows, Ubuntu, Mac.
* **Languages**-Python, R, Hive/Impala/Pig/MapReduce
* **Databases**-SQL-Server, My SQL, MS Access, MongoDB
* **Packages**-Python: Pandas, numpy, scipy, scikit-learn, H2O, Beautiful Soup, matplotlib, MDP, Rpy2, neurolab, NLTK, etc. R ggplot2, caret, dplyr, Rweka, gmodels, RCurl, tm, twitteR, NLP, Reshape2, rjson, plyr Etc.
* **Version controls**-Git, SVN

|  |  |
| --- | --- |
| **Education:**  **PhD in Statistical Science**  Duke University  **MBA Candidate**  New York University – Leonard N. Stern School of Business  **BS in Mathematics and Biological Science** Tsinghua University Beijing china | Durham, NC    New York, NY |

**Professional Experience**

#### AMC Networks, New York NY Apr. 2017 – Present

#### Data Engineer Consultant

In this role, multiple data sources were ingested to data warehouse and database to generate regular reports on revenue and subscriber growth, customer cohort analysis and user behaviors on AMC Networks’ new digital business. Served as a general consultant on (big) data-related initiatives.

**Responsibilities**

* Served as a general consultant on data evaluation, data interpretation, metrics construction and data quality assurance.
* Ingested structured and unstructured data and streaming data from multiple data vendors to data warehouse (Amazon Redshift) and in-house database.
* Built and delivered reports on revenue and subscribers growth metrics, customer cohort analysis and user behaviors on AMC Networks’ new digital business, in both Python and R.
* Built and tested the workflow scheduling framework for data ingesting and daily reporting tasks based on Apache Airflow and python, with multi-layer dependency structures.
* Build automated web scraping and data collecting processes via Selenium in Python.

**Environment**: Python, R, Git, SQL Server, Apache Airflow, Amazon Redshift

#### Point72 Asset Management L.P., New York NY Jan. 2015 – Mar. 2017

#### Data Scientist

In this role, large structured and unstructured data were employed to conducts quantitative and qualitative analyses to deliver fundamental insights. Company and industry key performance indicators (KPIs), trends, and patterns were analyzed and predicted using machine learning algorithms. Such algorithms are either proprietarily developed or commonly used in practice. The prediction methods are compared based on their efficiency and scalability in terms of estimation complexity and in terms of memory requirements for real-time predications.

**Responsibilities**

* Worked with data vendors that cover different sectors (consumer/retail, TMT, industrial, finance and healthcare), and different sources (transactional data, email text data, clickstream and web scraped data, etc.)
* Participated in all phases of research including data collection, data cleaning, data mining, feature engineering, developing models, validation/backtesting, visualization and performance monitoring.
* Implemented customer segmentation using Self-Organizing Maps, Decision Trees and Distance-Based Clustering, prototyped dynamic visualization of clustering results in Python, R and Tableau.
* Conducted data analyses for company-level predictive models on key performance indicators (e.g. sales, subscriber models, market shares, user churn/acquisitions, etc.), cross-sectional analysis, industry/macro indicators, customer segmentations and customer cohort analysis by using Python.
* Designed predictive and statistical modeling, data mining approaches and scalable analyses with >10TB daily data from multiple data sources on HDFS. Used pandas, numpy,scipy, H2O, matplotlib, scikit-learn, NLTK in Python for developing various machine learning algorithms on extracted and cleaned data.
* Prepared large volumes of structured and unstructured data, performed data quality checks, data cleaning and preprocessing, and performed ETL with Hadoop, Impala, Hive and Pig on HDFS and SQL Server.
* Performed Natural Language Processing in Python to mine unstructured data using document clustering, topic analysis, named entity recognition, document classification and sentiment analysis.

**Environment**: Python, R, Git, SQL Server, HDFS/Hadoop, MapReduce, Impala, Hive, Pig, Excel.

#### Validus America Inc, Validus Holdings Ltd. New York NY Dec. 2012 – Dec. 2014

#### Sr. Quantitative Analyst/ Data Scientist

#### This role was based on analyzing massive claims data, natural disaster risk data, reinsurance/insurance portfolio characteristic data to develop and apply efficient and scalable algorithms and models in areas including feature selection, recommendation model learning, correlation modeling, Monte-Carlo simulations, tail risk quantification and quantitative asset pricing.

#### Responsibilities

* Led efforts in developing the theoretical framework and scalable analytical models for reinsurance statistical pricing with big internal data on Amazon Web Services (AWS) by using MapReduce (EMR), Hive, Pig, Sqoop, Flume and Python, with >1TB input data and >10TB outputs, to support Validus’ insurance-linked securities (ILS) portfolio construction and reinsurance portfolio pricings.
* Generated statistical models in production on various business needs by using Python, R and Matlab. Selected models implemented include: multivariate correlation and copula models, factor analysis, ensemble models, Monte-Carlo simulation methodologies, decision trees, random forest, (generalized) linear regressions, supervised and unsupervised classification/clustering, dimension reduction and feature selection, and missing value inferences.
* Developed quantitative models and strategies for portfolio optimizations by using Python and Matlab. Implemented and developed algorithms on convex and non-convex optimizations in investors’ portfolio constructions with different risk constraints (VaR, market share, risk preferences, etc.)

**Environment:** Python, R, Mahout, Matlab, MySQL, SQL Server, Hadoop, MapReduce, Pig, Hive, Impala

#### Novartis Pharmaceuticals. East Hanover, NJ May 2012 – Aug. 2012

#### Statistical Methodologist

This role developed statistical models for conflict diagnostics for evidence synthesis in a multiple testing framework.

#### Responsibilities

* Developed a Bayesian graphical model to perform a network meta-analysis to estimate treatment effects in smoking cessation programs and an evidence synthesis to estimate HIV prevalence in Poland.
* Implemented the model developed in evidence synthesis and multiple testing problems in R and WinBugs.

**Environment:** R, WinBugs, SQL

#### CITIC TRUST Ltd, Beijing China May 2009 – Sep. 2009

#### Statistical Methodologist

#### This project was based on Chinese REIT’s historical performance data, historical default data and company characteristics to generate REIT statistical default models and portfolio optimization.

#### Responsibilities

#### Developed a copular based default model and performed feature engineering/selection using dimension reduction techniques in Python and Matlab.

**Environment:** Python, Matlab, SQL.