Xuan Wang

Austin, Texas | 512-200-6763 | xxxuan0213@gmail.com | linkedin.com/in/xxuan-wang/ | github.com/xuanwangg

SKILLS

Programming Languages: Python, SQL, Apache Spark, Git, R; Scikit-learn, Pandas, NumPy, Tensor-flow, Keras
Tools: Tableau, AWS (S3, DynamoDB, EC2), GCP BigQuery, Salesforce, Jupyter, Kubernetes, Azure, Docker, Excel

PROFESSIONAL EXPERIENCE

Data Analyst (AI/ML) **- Movable Ink.** Mar 2022 - Jan 2023

• Worked closely with engineers on customizing GitHub Actions to accomplish CI/CD for machine learning models; Assisted in troubleshooting, resolving, escalating data-related issues, and validating data to improve data quality.
• Developed ETL data pipelines to pre-process terabytes of email campaign data with PySpark, dealt with multichannel inbound data, making data ready for downstream purposes.
• Designed, optimized, and compiled effective SQL queries on BigQuery based on business needs, created interactive Tableau dashboards, and demonstrated key insights to support the internal modeling team.
• Created and implemented a reporting automation tool in Python with in Cloud run for 10+ clients; Minimized manual input needs, reduced human errors, and sped up report generation > 5X.
• Built comprehensive analysis and defined file health metrics in Python to identify high-converting campaigns; Communicated and suggested future initiatives to key stakeholders ranging from CX teams to engineering teams.

Data Science & Analysis Intern **- Renzoe Box, Inc.** Sep 2021 - Feb 2022

• Built data pipelines for structured & unstructured data on AWS & Setup web connectors for pre- trained machine learning models using Flask in Python. Processed multi-label classifications on text data with NLP to predict 64 tags for 28k+ products;
• Created the first content-based ML solution using Spark for makeup recommendation systems by wrangling 500+ responses.
• **Impact**: Implemented methodology & define metrics to support makeup preference decisions and customize recommendations with the help of neural networks and exhaustive search.

Cloud Computing TA **- UT Austin, Department of Computer Science** Jan 2021 - May 2021

• Utilized Git to implement and integrate automatic grading Python scripts; Reduced human supervision by 70%.
• Programmed S3 and DynamoDB handlers to manipulate distributed data with AWS SDK, Python (Boto3) in virtual environments.
• Explored containers with 60+ students & deployed Helm Charts on Google Kubernetes using a single node GKE cluster.

Data Research Assistant **- Texas Department of Information Resources** Dec 2020 - June 2021

• Responsible for developing an automatic web scraper by leveraging requests, BeautifulSoup, and selenium Python packages.
• Extracted 12k+ vendor info via Salesforce database REST API & data ETL; Fed to 15 websites as inputs to scrape required files.
• **Impact**: Automated manual efforts through multi-threaded Python application with targeted procedures & functions for meeting extensive data requirements, leading to a time efficiency >10X; Eliminated 90+% oversight by contract managers.

Big Data & Distributed Programming TA **- UT Austin, McCombs School of Business** Aug 2020 - Jan 2021

• Led students to utilize EC2 GPUs to accelerate TensorFlow and Keras deep learning models and enabled fast experimentation.
• Demonstrated key operations of RDD in Apache Spark; Transformed and merged 1TB+ sized distributed datasets to build a collaborative-filtering-based movie recommend system.

EDUCATION

MS In Information Studies - The University of Texas at Austin | GPA: 3.45/4.00 Aug 2019 - May 2021
BS In Electronic Commerce - Dalian University of Technology | GPA: 3.40/4.00 Sep 2015 - Jun 2019

ACADEMIC PROJECTS

Clinical Narrative Analysis with Apache cTAKES Jan 2021 - May 2021

• Researched an open-source NLP system and leveraged it to extract & transform 30k+ EMR into SNOMED CT using Python.
• Developed an LSTM model for Named Entity Recognition and conducted unstructured text analytics to highlight critical findings.

Readmission Prediction with MIMIC-III Database Sep 2020 - Dec 2020

• Investigated anomaly detection techniques and forecasted readmission probabilities for over 175k clinical records; Applied Under-sampling and dimension reduction methods to remediate severe imbalanced data & obtained 7 key features.
• Performed ML models, tuned with Grid Search and Stratified K-fold cross validation, achieved 92.5% accuracy with XGboost.