Xuan Wang

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

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

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

PROFESSIONAL EXPERIENCE

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

• Programmed and executed a Python-based reporting automation tool on Cloud run for 10+ clients, resulting in significant improvements such as reduced manual input requirements, decreased human errors, and a more than fivefold acceleration.
• 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 to engineering teams.
• Compiled, and optimized effective SQL queries on BigQuery to meet business requirements and demonstrated significant findings to support the modeling team through interactive Tableau dashboards.
• Performed exploratory data analysis on terabytes of email campaign data using PySpark to process inbound data from multiple channels, making data ready for downstream purposes.
• 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.

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

• Utilized NLP to perform multi-label classification of textual data, resulting in the prediction of 64 tags for 28k+ beauty products.
• Created the first content-based ML solution in Spark to generate makeup recommendations by wrangling 500+ surveys.
• Modeled structured and unstructured data, deployed ML models with Flask on AWS EC2 to visualize the matching results.
• **Impact**: Implemented methodologies utilizing neural networks and exhaustive search to create personalized makeup recommendations based on individual preferences. Established relevant metrics to support decision-making process.

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

• Tasked with implementing an automated web scraping solution by applying Python libraries such as requests, BeautifulSoup, and selenium for data extraction.
• Fetched 12k+ vendor records through Salesforce API & data ETL, fed as inputs to 15 websites for targeted file collections.
• **Impact**: Developed a multi-threaded Python application with optimized procedures and functions to fulfill extensive data requirements, leading to a time efficiency >10X, and significantly reduced oversight by contract managers by 90%.

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

• Designed Git-based automated Python grading scripts with seamless integration, reduced human supervision by 70%.
• Developed S3 and DynamoDB handlers to facilitate distributed data manipulation with AWS Python SDK in virtual environments.
• Led a group of 60+ students in container exploration and successfully deployed of Helm Charts on a GKE single-node cluster.

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

• Guided students in leveraging EC2 GPUs to experience quicker execution capabilities of TensorFlow/Keras deep learning models.
• Demonstrated key operations of RDD in Apache Spark, transformed and merged 1TB+ sized distributed datasets to build a collaborative-filtering-based movie recommend system.

ACADEMIC PROJECTS

Clinical Narrative Analysis with Apache cTAKES Jan 2021 - May 2021

• Researched on an open-source NLP system and leveraged it to extract and transform 30k+ EMR records into SNOMED CT.
• Designed an LSTM-based model for performing Named Entity Recognition and executed unstructured text analytics to identify significant insights.

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

• Conducted analysis on anomaly detection methodologies and predicted readmission probabilities for 175k+ clinical records.
• Employed under-sampling and dimension reduction methods to remediate severe imbalanced data & obtained 7 key features.
• Optimized machine learning models using Grid Search and Stratified K-fold cross-validation resulted in a 92.5% accuracy rate with XGBoost.