AVA (TONG) YANG

avayang0926@gmail.com | (929)386-8805 | LinkedIn | Portfolio Website

Full-Stack Data Scientist with experience of building end-to-end **Machine Learning** solutions on distributed systems to identify key insights and drive business values. Well-versed in collaborating with cross-functional teams and communicating results to stakeholders from various background.

TECHNICAL SKILLS

- **Language:** Python, PySpark, SQL, Git, Linux | scikit-learn, pandas, ml-lib
- **Platform:** Databricks, VS Code, AWS (S3, EMR), Airflow, Snowflake, Github
- **Data Analysis & Visualization:** Databricks, Snowflake, Plotly | PowerPoint, Excel
- **Hands-on Experience:** Model Development, Deployment & Productionization, Data Preprocessing, ETL, Feature Engineering, Model Evaluation, Data Visualization, Big Data, Version Control

WORK EXPERIENCE

Capital One (US Card, Transaction Intelligence) **New York, NY**

Senior Machine Learning Scientist (PySpark Pipeline, Databricks, AWS, Git, SQL) *Jul. 2021 – Now*

Developed and delivered Tree-based models and PySpark pipeline enhancements for the identification and prediction of Recurring Transactions, enabling Block Charges and Continue Charges features in mobile app.

- Achieved a 6% lift in prediction accuracy by initiating a model release with new features. Conducted correlation analysis to support the decision, retrained the model on AWS and productionized the code change within 2 weeks.
- Implemented and managed model monitoring practices for the team, automated codes to generate quarterly monitoring reports within 10 hours. Evaluated model performance on metrics such as detection rate, PSI, and PR-AUC.
- Led the source data migration project, rewriting the ETL process on Databricks and resolving data quality issues to ensure the source data consistency, which reduced data downloading time in the pipeline by 30%.
- Initiated to revisit confidence score thresholds after model refit. Proposed a monthly subscription threshold increase from 0.5 to 0.8, achieved a 3% reduction (18M in a quarter) of False Positive cases and significant cost savings.
- Identified and resolved a critical production issue in Q1 2022. Collaborated with engineers to identify the root cause, validate results, and backfill data into Snowflake database, designed data quality checks into daily pipeline.
- Partnered with product team to derive key insights and optimal recommendations, worked with engineers to deploy model changes in production, presented to stakeholders multiple times and obtained positive feedback.

Articence (Startup, Intelligent Hiring Platform) **New York, NY**

Data Science Intern (Natural Language Processing, Python, AWS) *Jun. 2020 – Aug. 2020*

- Built a Job & Resume Analyzer Web App product using BERT-based model in Python, to predict key skills match with model accuracy achieved of 90%, delivered personalized career guidance to customers.
- Scraped software websites and clustered ~500 text reviews utilizing Topic Modeling algorithm, identified marketing campaign opportunities and improved 7% of Click-Through Rate.
- Collaborated with engineers to cut product's runtime by 85%; partnered with product team and communicated user behavior insights with data visualization.

CreditX (Fintech Company) **New York, NY**

Data Science Student Intern (Credit Risk Monitor, Python, SQL, GCP) *Feb. 2020 – May 2020*

- Reduced loan default risk by 10% from user behaviors by deploying predictive models with Random Forest, XGBoost, and LSTM in Python, monitored Credit Risk in Online-Lending business.
- Mined 200K operations data and integrated Natural Language Processing (NLP) techniques to preprocess text data; conducted exploratory data analysis (EDA) in Tableau and data manipulation in SQL for modeling.
- Customized Transformer Model to extract 2 new features from alternative data on cloud service (GCP), used by financial analysts to facilitate credit risk analysis process.

EDUCATION

Columbia University **New York, NY**

Master of Science in Operations Research, Concentration in Data Science, GPA 3.6/4.0 *Aug. 2019 – Dec. 2020*

Coursework: Machine Learning, Deep Learning, Business Analytics, Data Visualization, Stochastic Models

Zhongnan University of Economics and Law **Wuhan, China**

Bachelor of Science in Information and Computer Science, GPA 3.8/4.0 *Sept. 2015 – Jun. 2019*

Coursework: Statistical Inference, Operations Research, Econometrics, Financial Mathematics