Wes Shi

781-718-8038 | wesshi818@gmail.com | **Portfolio** | LinkedIn

EDUCATION

Columbia University New York, NY — Aug 2022-Dec 2023
MSc Operations Research (**STEM**) - **Analytics** track — GPA 3.88/4.0

Brandeis University Waltham, MA — Jan 2019-May 2022
BS Applied Mathematics & Quantitative Economics — GPA 4.0/4.0

SKILLS

Programming: **Python** (Sklearn, TensorFlow, Beautiful Soup, NLTK, NumPy, Pandas, Matplotlib, Plotly), **SQL**, R, Java
Methods: A/B Testing, Time Series Forecasting, Natural Language Processing, Statistical Modeling, Web Scraping
Tools: AWS S3, Excel (VBA), Hadoop, Word, PowerPoint, GitHub, Git, Redshift, Stream lit, Snowflake, SPSS, Tableau

PROFESSIONAL EXPERIENCE

System2 — Jun 2023-Aug 2023
Data Scientist — New York, NY

- **ETL Process:** Automated and optimized ETL process of actively trading stocks' **alternative data (1 TB+)** using **AWS S3** and redshift database, reduced extraction time by **60%+**
- **Machine Learning:** Performed quantitative portfolio analytics based on market risk data, adapted **random forest** for selecting 6+ influential risk metrics, supported making investment decisions
- **Data Visualization:** Developed **3+** customizable app (Stream lit) with scalable **Python & SQL** pipeline for visualizing portfolio return prediction based on **LSTM** and **XGBoost**, achieved MSE of 1.7 on short-term prediction
- **A/B Test:** Designed and analyzed A/B test for clients, evaluated the effectiveness of new credit card offer tag by **T-test**, identified ~**3% growth** on active users
- **Market Growth Estimation:** Constructed pipeline for scraping **8k+** gyms' info periodically and automatically fulfill/update AWS RDS database; innovated **market growth estimation algorithm** and provided growth strategy insights

DG Venture Law Firm — Jan 2023-May 2023
NLP Data Scientist — New York, NY

- **Product Design:** Confirmed business needs of NDA review product by communicating with legal department, identified **5+ essential end-user needs** (e.g., cloud-based pipeline), developed detailed documentation
- **Product Development:** Led a team to design and develop an **end-to-end solution** for detecting, highlighting clauses containing 10+ types of deal-breaking information in NDAs, providing modification suggestions
- **Text Classification:** Implemented text feature extraction with bag of words model and trained classifiers including LinearSVC (f1 ~82.3%), Multinomial Naïve Bayes for clause classification
- **NLP Modelling:** Fine-tuned **BERT** model and **neural network** to perform clauses' classification, achieved 95%+ f1 score; utilized corresponding deal-breaking info searching, captured **99%+ improper clauses**, saved **20+** minutes per NDA review

JOANN — May 2022-Aug 2022
Data Scientist — Hudson, OH

- **Exploratory Data Analysis:** Performed EDA to raw data from HANNA using Python; identified 14+ types of inaccurate records and automated data cleaning pipeline, increased **data integrity by 10%+**
- **Feature Engineering:** Constructed **SQL** table including 280+ ship routes' distance, origin, and destination ports for revamping the original geo-map straight distance, improved feature accuracy by **52%**
- **Machine Learning:** Cooperated with analytics team, applied advanced analytical methods including XGBoost prediction model, achieved model error reduction from 14 to 8.4 days, **improved warehouse efficiency by 30%+**
- **Data Visualization:** Analyzed the effect of replacing HAP with snowflake, including reducing the query time consuming by **60%+**, presented the insights to stakeholders through **Tableau dashboards**

PROJECT

LLM Research Assistant (Columbia University) — Sep 2023-now

- **Literature Review:** Collaborated with fellow RA on literature review for LLMs in social dilemmas, examining their dynamics in 5+ traditional games, including prisoners' dilemma, ultimatum game, etc.
- **LLM application:** Pioneered testing techniques for LLMs in sequential multi-agent social dilemmas like fruit gathering; developed **Streamlit dashboard** to display real-time actions of 2 different LLM based agents' actions

Credit Card Default Project (Capital One) — Jun 2021-Aug 2021

- **Data Preprocessing:** Deployed Synthetic Minority Oversampling Technique (SMOTE) based on 790K+ imbalanced data to create synthetic data points, promoted the recall and **addressed class imbalance problem**
- **Fine-tuning:** Tuned hyperparameter with Grid Search by k-fold cross-validation, achieved **75.68% auc-roc** in default prediction