YOU WANG

(201) 736-4733 | yw6127@nyu.edu | linkedin.com/in/you-wang-7094231a3

EDUCATION

New York University, New York, NY — Sep 2021 – May 2023 (Expected)
Master of Science in Data Science | **GPA**: 3.9/4.0
Relevant Courses*: Introduction to Data Science, Probability & Statistics, Optimization & Linear Algebra, Machine Learning, Big Data, Deep Learning, Natural Language Processing.*

Sun Yat-sen University, Guangzhou, China — Sep 2016 – Jun 2020
Bachelor of Science in Biotechnology | **GPA**: 3.8/4.0
Relevant Courses: *Database System, Computer Programing Language, Biostatistics, Machine Learning Fundamental*

TECHNICAL SKILLS

Tools: Python (NumPy, Pandas, Scikit-learn, Maplotlib) | R | SQL | LaTex | Tableau | Excel
Proficiencies: A/B Testing | Data Mining | Machine Learning | Big Data Analysis | EDA | Data Visualization| Statistics

PROFESSIONAL EXPERIENCE

Data Analyst Intern - **China Guangfa Bank Credit Card Center** — Aug 2020 – Nov 2020
- Extracted, cleansed & transformed 1.3M credit card transactions & customer demographics data using SQL queries, stored the processed data in a MySQL database daily, computed & visualized daily trends in 8 key KPIs with a Tableau dashboard.
- Applied feature engineering like One-Hot encoding, data binning & feature combination, augmented 25% new metrics & monitored daily key metrics of credit card operations data.
- Created 70+ interactive visuals, discovered data-driven insights with intuitive storylines & actionable recommendations daily.

Data Analyst Intern - **Datastory Information Technology** — Oct 2019 – Jan 2020
- Conducted exploratory data analysis for PepsiCo, Procter & Gamble and Bear Electric, identified & treated data anomalies using SQL & Python, segmented customers based on social media data, provided targeted audience profiles to brands.
- Tracked product upgrades & marketing campaigns performance with intuitive graphs & plots in PowerBI dashboards.
- Streamlined & automated 2 ETL processes, defined data cleaning & transformation rules, augmented data annotation with user portraits, and engineered 10+ new features from unstructured data, saved repetitive manual efforts by 60%.

Data Analyst Intern - **Kangmei Pharmaceuticals** — July 2019 – Sept 2019
- Implemented daily data ETL & processing pipeline for handling~400k sales records, produced daily & weekly reports with SQL & Excel for 10+ business users, saved ~20 hours per week of manual efforts.
- Analyzed sales variations with 9 other key product marketing metrics, performed correlation analysis & visualized trends with Python Plotly graphs & charts, empowered evidence-based decision making for product promotions.

PROJECTS

Diabetic Hospital Readmission Prediction | *Python, Machine Learning, Random Forest, Xgboost, Lasso Regression*
- Cleansed & processed 100+ features related to diabetes patient health & demographic data using Python Pandas & NumPy, performed correlation analysis for features selection & reduced to 44 significant features.
- Built & evaluated Random Forest, XGBoost and Logistic Regression classifiers with Sci-kit learn in Python to predict short-term & long-term readmission rates for diabetics, achieved an accuracy of 91% & AUC of 0.80.
- Constructed & optimized Lasso regression models to predict length patient stay in the hospital with MAPE of 12.6%.

CNN-Based Nucleic Acid Sequence Classifier | *Python, TensorFlow, Deep Learning, CNN*
- Processed HLA DNA sequence and 16s rRNA sequences into a 4 X 1 dimensional array using NumPy in Python.
- Designed a CNN with 6 convolutional, 6 pooling & 2 fully connected layers using TensorFlow in Python, achieved an accuracy of 93+% on all types of sequences.

Healthcare Provider Fraud Analysis | *Python, Machine Learning, Stacking, Random Forest, Xgboost, MLP*
- Cleansed & processed 160k+ rows of medical and insurance data from 5410 healthcare providers, studied and selected the features using Exploratory Data Analysis, created 25 new features
- Built a stacked model based on Random Forest, Xgboost and Multilayer Perceptron to predict the fraudulent behavior of healthcare providers, achieved an F1 score of 0.58 and an AUC score of 0.92

Citi Bike Network Flow Analysis | *Python, Machine Learning, RNN, Clustering*
- Preprocessed ~3M Citi Bike usage big data with 14 variables, incorporated multi-processing & multi-threading techniques to optimize date cleaning & preprocessing, decreased preprocessing time by 75%.
- Clustered 1000+ stations to 6 groups with K-Means algorithm in Python to reduce computational cost, transformed the data into 6X6 matrix to record trip counts within 2-hour intervals for each of the clusters.
- Built Random Forest, Graph Convolution Network & RNN models to predict the imbalancedness in bike distribution for each cluster within 2 hours, evaluated model performance & achieved MSE of 70.4 with RNN model.

Movie Recommendation System | *PySpark, Recommender System, Latent Factor Representation, ALS*
- Preprocessed 1.2M movie ratings from MovieLens, performed data cleaning & outlier treatment using Pandas in Python,
- Implemented collaborative filtering recommendation using Spark ALS method with latent factor representation of viewers & movies in PySpark, applied hyperparameter tuning with GridSearchCV, achieved MAP@100 of 0.001 & nDCG@100 of 0.011

Haizhu Wetland Ecosystem Service Value Analysis | *R, Python, Logistic Regression, Linear Regression*
- Applied Coarsened Exact Matching algorithm to deal with data imbalance, implemented Logistic Regression model applied backward elimination, quantified the impact of features on people's willingness to pay for ecosystem protection.