Shimin Zhang

zhang.shimi@northeastern.edu | LinkedIn | Github | +1 617-314-5190

Education

Old Dominion University — Norfolk, VA, USA
Master in Modeling and Simulation Engineering — *Sep 2020 - May 2022*
Coursework: Model Engineering, Parallel Computing, Advanced Analysis for Modeling& Simulation, Machine Learning, Web Programming

Northeastern University — Boston, MA, USA
Master in Data Analytics — *Jan 2019 - Sep 2020*
Coursework: Data Warehousing&SQL, Data Management&Big Data, Data Visualization, Statistics, Data Mining, Analytics Systems Technology

University of Electronic Science and Technology of China — Chengdu, China
Bachelor in Finance — *Sep 2013 - June 2017*

Skill Summary

- **Languages**: Python(Numpy, Pandas, sklearn, Matplotlib, Dash, PySpark), SQL, R, JavaScript, C++
- **Tools**: AWS(S3, EMR, Athena, QuickSight), BigQuery, PostgreSQL, MySQL, Tableau, Git, Google Analytics
- **Professionals**: Google Data Analytics Professional Certificate, Data Analyst with SQL Server Track

Work Experience

Research Assistant at Old Dominion University | Norfolk, VA — Sep 2020 - Sep 2021

- Trained and deployed end-to-end image&CNN-based self-driving algorithms on cars in 3D simulators and physical robots.
- Designed simulated transportation system to collect various driving behavior data of autonomous cars to database.
- Transformed the deep JSON structure data of transportation systems to non-relational data format for advanced modeling.
- Defined and analyzed indicators to quantify the safety hazards of attacked autonomous cars(**best presentation** reward).

Data Analyst Intern at HAI Analytics Inc. | Boston, MA — June 2020 - Aug 2020

- Drilled down college admission, financial aid, and student enrollment details with interactive dashboards in **Tableau**;
- Applied statistics and machine learning algorithms to predict student enrollment probability for university to decide.

Data Scientist Intern at GE Aviation | Boston, MA — Jan 2020 – Apr 2020

- Pre-processed millions of industry data with encoding, joining, feature engineering for visualization and modeling;
- Derived insights about malicious logins from **6 million** system logs through exploratory data analysis and visualization;
- Compared different **machine learning** algorithms (like decision tree, XGBoost) in terms of the detection performance;
- Evaluated the final model with confusion matrix along with F1 score and optimized it on **PySpark** and AWS **SageMaker**.

Game Operation Analyst at Tap4Fun | Chengdu, China — May 2017 – Dec 2017

- Collected users' feedback using customized **ETL** pipelines from questionnaires and APIs of game communities;
- Worked with **MySQL database** to query large amount of data and analyze patterns to stimulate user's payments;
- Designed **A/B tests** to evaluate the performance of different operation activities, interface layouts.
- Designed and optimized **models** to do player segmentation, life-time value forecasting, and bug detection.

Selected Projects

- **Predict User Churn in Music Streaming Service**: Python, Machine Learning, Classification
  - Cleaned the dataset with imputation, one-hot encoding,scaling and compare indicators of different groups of listeners by charts;
  - Used SMOTE method to create new instances of the churn users (cancel and downgrade subscription) to balance the dataset;
  - Predicted potential churn users with decision tree after comparing accuracy rate of various machine learning models.
- **Real Estate Pricing Prediction**: Python, Machine Learning, Regression
  - Analyzed a dataset of housing sales in California to find the correlations between sale price and the location, number of rooms, etc;
  - Applied regression models and techniques such as stepwise, Lasso, Ridge regression, Random Forest, XGBoost to predict house price;
- **Real-time Data Visualization Dashboard**: Python, Dash, Plot.js, HTML, Visualization
  - Automated the pipeline of collecting and pre-processing(cleaning, merging, calculation) real-time Covid-19 data from open resources;
  - Developed dynamic dashboard to visualize indicators such as new cases, death, filtering by country, along with bar, line, map charts.
  - Developed interactive dashboard to demonstrate sale performance changing over time, comparing among regions and departments.
- **ETL Pipeline Development**: Python, SQL, Functional Programming, Database
  - Customized the pipeline to fetch real-time business data from Yelp API and load them to relational database;
  - Extracted song and log data from JSON format and transform to 5 different tables in star schema, then load them to Postgres database.