Linda (Guanqi) Zeng

2616 Erwin Road, Durham, NC 27705 | +1 336-671-6045 | guanqi.zeng@duke.edu

SUMMARY

Highly self-motivated data scientist with one year's hands-on industry experience in Machine Learning, Data Visualization, Data Transformation for structured and unstructured dataset

- Programming Language: Python (NumPy, scipy, sklearn, pytorch) and R (CART, randomForest, brms)
- Database management & Data Transformation: SQL (Azure SQL), pyspark, dplyr, pandas
- Data visualization: Tableau, Power BI, ShinyApp, matplotlib, seaborn, Bokeh

EDUCATION

Duke University Durham, NC
Master of Science in Statistical Science August 2020-May 2022
GPA: 3.7/4.0

- Related Courses: Predictive Modeling, Bayesian Modeling, Statistical Inference, Probabilistic Machine Learning, Introduction to Deep Learning, Theory & Algorithms in Machine Learning, Time Series and Dynamic Models

Wake Forest University Winston-Salem, NC
Bachelor of Science in Mathematical Statistics August 2016-May 2020

- Major GPA: 3.9/4.0
- Related Courses: Computational and Nonparametric Statistics, Data Structures & Algorithms, Multivariate Statistics

WORK EXPERIENCE

PowerSecure Durham, NC
Data Analyst Intern June 2021- Now

- Developed data pipeline to extract, clean and transform 7 years' raw minutely utility data by SQL, Pandas and Matlab
- Conducted data exploratory analysis & investigated business potentials for targeted customers through various charts
- Identified sites with abnormal utility loads through abnormal detection analysis by isolation forest algorithm and hypothesis testing with summary statistics
- Standardized existing customer segmentation strategy by classifying customers into multiple benefit levels with rule-based methods
- Summarized model analysis results by creating an interactive report to reflect data health and business insights via Power BI

PROJECT EXPERIENCE

SMART news Project Duke University
Team Member May 2021

- Cleaned news text data by lower casing, removing punctuation and stops words; tokenized & transformed the data to TF-IDF matrix through NLTK
- Implemented passive-aggressive classifier to identify fake news; obtained 94% accuracy for test data
- Built a recommendation system for recommending similar content by clustering news content through matrix reduction with SVD and linkage hierarchical clustering

Case Study on Morphine Street Price Duke University
Team Member March 2021

- Improved data quality & expanded 30% of the analyzable records by refining manual records with census data
- Developed linear fixed-effect model to discover geographical differences in drug prices for various regional clusters
- Estimated base morphine price of each state and detected outliers through hierarchical modeling; discovered the heterogeneity between states and the negative relationship between dosage strength and purchase size

DataFest: Non-medical use of prescriptions in Germany Duke University
Team Leader April 2021

- Developed logistic regression model with 70.2% accuracy for identifying drug abuse cases with 156 survey questions data on demographic background, physical health, and substance use history
- Implemented Random Forest model with 80.3% accuracy by socio-economic status data on income levels, education, age, and psychiatric medication usage

Understanding Happiness Project Duke University
Team Member November 2020

- Assessed worldwide happiness scores (0-10) under extreme social settings by fitting a linear regression model
- Predicted happiness scores for each country using Random Forest, XGBoost, and SVM