David Zhao

zdwzhaoyikai@gmail.com | 979.264.1086
ykzhao.com

Research methodologist and business-minded **data scientist** with passion in applying academic rigor to facilitating decision-making in business via engaging communication of data-driven insights. 4+ years of experience in research design and data analytics.

- **Statistics**: Generalized Linear Models, ANOVA, (non)parametric hypothesis testing, power calculation, A/B testing, etc.
- **Predictive modeling**: Xgboost, Random Forest, Elastic net, Neural Network, SVM, KNN, Kmeans, ResNet, BERT, etc.
- **Survey methodologies**: nonresponse adjustment, post-stratum weighting, raking, missing imputation, survey sampling, etc.
- **Interpretable machine learning**: making sense from the 'black box' models; SHAP, LIME, partial dependence plot, etc.
- **Causal inferences**: propensity score, covariate matching, instrumental variables, regression discontinuity, etc.
- **Structural equation modeling**: mixture models (Latent class and Latent Profile modeling), growth curve modeling, etc.
- **Natural language processing**: multiclass and multilabel classification, word2vec, BERT, etc.
- **Visualization**: Tableau, ggplot2, matplotlib; author of the R visualization package: LCAplotter

EDUCATION

TEXAS A & M UNIVERSITY | Ph.D. in Communication | M.S. in Statistics

May 2020 | December 2019 | College Station, TX

*Dissertation title: Experimenting with different NLP deep learning architectures in frame (textual) analysis

NORTH CHINA ELECTRIC POWER UNIVERSITY | B.A. in Advertising

May 2015 | Beijing, China

EXPERIENCE

PUBLIC POLICY RESEARCH INSTITUTE | Data scientist

May 2018 – Sept. 2018 | Bryan, TX

- Conducted end-to-end data analyses independently for three large projects: Coastal Resilience survey, World Value Survey and Nativism world trend (Ipsos); presented insights to academic conferences and survey sponsors
- Developed the R package LCAplotter for visualizing the Latent Class Models more dynamically
- Advised the design of questionnaire for Regional Health Survey in College Station, TX
- Developed an Shiny app for visualizing diverse clusters from a Latent Class Model for a political science conference
- Parsed news content from Lexis-Nexis database using Beautiful Soup and Regex for a political communication research project
- Conducted Measurement Invariance test for the Nativism variable for the Ipsos global nativism survey

TEXAS A & M UNIVERSITY | Course Instructor

Sept. 2015 – present | College Station, TX

- Designed and instructed the course COMM-308 Research Methods in Communication (survey, interview & content analysis)
- Instructed the course COMM-203 Public Speaking

MACHINE LEARNING PROJECTS

May 2017 – present

- Presented a method to incorporate Pytorch-Transformers BERT models into Fastai framework for NLP projects; widely used by NLP community
- Achieved Top 15% in an Kaggle Histopathologic Cancer Detection computer-vision contest with an ensemble of DenseNet and ResNet
- Introduced machine learning interpretation method, Tree-based SHAP values, for house pricing prediction in a STAT seminar
- Optimized and implemented ML models from scratch: regularised Multinomial Logistic regression, Kmeans and ResNet

PROGRAMMING & TOOLS

PROGRAMMING

Proficient:
Python • R • Pytorch • MSSQL • Tableau
Familiar:
HTML • CSS • STATA

COMPUTATIONAL TOOLS

Cloud computing • Google Colab • Git
Shiny (R) • Markdown • LaTeX
ML essentials:
scikit-learn • fastai • Pytorch-transformers • SHAP