JOHN M. NIEHAUS

Webpage: jmniehaus.github.io
GitHub: github.com/jmniehaus

EDUCATION

M.S., M.A.; Statistics, Political Science; *Texas A&M University*; 2021

B.A., Political Science, Psychology; *University of Missouri*; 2014

SOFTWARE, PROGRAMMING & LANGUAGES

R (6 years); **Python** (2 years); **tensorflow/keras** (6 months); **SQL** (2 years); **git** (3 years); **Tableau** (1 year); **Bash** (1 year); **jupyter notebooks** (2 years); **Stata** (2 years)

EMPLOYMENT

- **Senior Data Scientist**, *Mercy Health*; May 2023 – present
- **Data Scientist**, *Veterans United Home Loans*; Jun. 2021 – May 2023
- **Graduate Research Assistant**, *Texas A&M University*; Aug. 2016 – Jun. 2021
- **Teaching Assistant, Advanced Time Series**, *University of Michigan*; Summer 2018, '19
- **Replication Analyst**, *Cambridge Press*; Aug. 2018 – Aug. 2020

PROJECTS

Cervical Cancer Diagnosis (*personal project*) – GitHub Repo

I predict the probability that a patient has cervical cancer to minimize the necessity of invasive testing procedures. After walking users through extensive exploratory data analysis, I write a wrapper class around the miceforest package's MICE imputer in order to gain robust support for MICE imputation through the sklearn API, and therefore to permit MICE imputation in randomized cross validation pipelines. After addressing this API inconsistency, I train gradient boosted trees to maximize the F2 score using repeated, nested randomized search cross-validation. Although the average precision is bad–reaching only 0.31–this is nearly double the average precision reached by the original authors of this data, and does not involve advanced techniques beyond simple gradient boosted trees.

Competition Risk Modeling (*Veterans United Home Loans*)

I modeled the probability that a potential home-buyer will exit our home-buying process and close a mortgage with competing lenders. This first involved writing several SQL stored procedures to import and clean disparate datasets, primarily via window functions, joins, and aggregate functions. Next, I used this training data to impute missing values a MICE imputer followed by training a logistic elastic net model with parameters tuned by randomized cross validation. Finally, I analyzed the factors most associated with competition risk, suggesting to the business that we incentivize these borrowers with greater credits on their mortgage.

R Package – bizicount (*personal project/master's thesis*) – Package Link

As part of my master's degree and as a continuing side project, I wrote an R package that estimates bivariate, zero-inflated count models using copulas. In addition, it interfaces with other packages to print publication-quality regression tables, and carry out non-parametric post-estimation model diagnostics. In writing this software, I gained heavy experience with object oriented programming principles, debugging, version control in git, and meta-programming. Moreover, I learned how to handle numerical issues in optimization, and how to recast constrained optimization problems as unconstrained problems while still accurately computing uncertainty estimates.