Samuel J Thomas

Principal Data Scientist
Address: 70 W Clear Lake Ln, Westfield IN 46074
samuel.thomas522@gmail.com +1 317 696 9214 linkedin.com/in/samueljthomas1/ | *Updated:* April 3, 2024

Education

Indiana University, Ph.D. Biostatistics May 2021

Purdue University, M.S. Mathematics

University of Notre Dame, B.S. Electrical Engineering

Society of Actuaries, Associate of the Society of Actuaries (ASA)

Professional and Academic Experience

Budget and Finance

Guidewire: Lead modeler and prototype developer for Merger and Acquisition cyber-risk model intended for a leading financial services client. This model is used to quantify the cyber risk of merging two organizations. Adjustments in tail risk and average annual loss are based on the combined risk profile of the two companies. Provided direction to contractors on the report and chart design.

Capital Group: Developed machine learning and statistical models to redesign the North American Distribution sales territories. These models optimized the geographic size and location of territories for over 200 sales professionals across the United States. One direct impact of this analysis was the creation of a new sales territory region in the San Francisco area. Modeling techniques used included K-means clustering, 2-dimensional optimization, and a customized adaptation of genetic algorithms.

Capital Group: Developed an attribution model to inform the business regarding which types of events are most profitable to the organization. The model automatically estimates lift in sales and visits for many sponsored events in Azure Databricks. This model incorporates a deep learning model called an autoencoder for advisor matching, and a Bayesian generalized linear model for attribution. Long-term plans including evolving this model for recommendations.

Capital Group: Estimated impact of mutual fund placements on recommended lists using a generalized linear mixed effects model. This model is used to quantify $50B in sales opportunity, delineated at the mutual fund level.

Capital Group: Developed a sequence analysis using the TraMineR package to quantify the sales preferences of 40k financial advisors. This analysis demonstrated that the firm's sales in one product did not cannibalize sales in another product. An initiative is planned to identify leads for the newer product, based on the results of this analysis.

Capital Group: Led a team of 3 consultants to build analytics for three new departments of the North American Client Group. These departments work with the home offices of financial services firms for product placement and sales initiatives.

Capital Group: Developed a simulation model in R to estimate the sales lift for a major coverage redesign initiative. This model is based on historical and projected market share, based largely on activity data reported through salesforce.com.

Capital Group: Extensive experience with using internal and external data to drive business decisions related to sales analytics. Worked with senior business leadership to build the foundations for reporting and analytics.

Capital Group: Fit a Generalized Linear Mixed Effects model (GLMM) to sales and activity data to identify the most significant drivers of work for Internal Wholesalers (IW). This model was used to assign IW coverage by balancing capacity between regions in the United States.

Forecasting and Outbreak Analytics

Capital Group: Developed the forecasting methodology and implementation for the American Funds Service Company budget process. This process has been adapted and reproduced for multiple departments in the organization. I developed regression models with ARMA errors using Rob Hyndman's forecast package in R. These models related macroeconomic indicators to fund account volumes and were used in various scenario analysis. This forecast model was consistently accurate within 3% of account volumes. I served as lead data scientist in developing statistical models for the AFSG budget process. Models are used to develop hiring plans to support a >1,500-person service center and by Global Finance for revenue projections. Accurate forecasting is critical to the financial health of the organization.

Capital Group: Identified contact rate influencers using Generalized Least Squares (GLS) models. Contact rates are the number of phone calls per period of time (usually months). Daily seasonal factors were fit using Fourier series and macroeconomic variables such as the U3 unemployment rate, the United States GDP, and changes in the Dow Jones Industrial Average index for the stock market. Discussion with Data Science Interest group contributed to increased leveraging of AFS data for analysis for broader Capital Group applications, including the forecasting process mentioned above. For example, results from this steady helped the short term workforce planning team to staff the call center more appropriately after a sharp shift in the stock market.

Capital Group: Developed a custom forecast package in R called mlts (multi-level time series) to facilitate the annual forecast process. This package automatically combines hierarchies of low level forecast. This approach facilitated business adjustments on any of hundreds of forecasts for use in the budget process.

WellPoint/Anthem: I pulled data from the CDC to produce an influenza report for the Corporate Actuarial team. This report compared the levels of influenza like illness for the current year to previous years by state. This data was used by the actuarial teams to help estimate the reserves needed based on the incoming health claims patterns.

Independent consulting: Created an automated forecasting model to predict fuel demand in various locations in the UK. This model was used to anticipate geographical fueling needs for a trucking company.

Independent consulting: Designed a statistical sampling plan for Medscape to estimate treatment plan adoption among physicians.

Bayesian Statistics

Guidewire: Designed and developed an original algorithm to quantify supply risk in a Bayesian network. Algorithm is implemented in python with a polynomial time complexity. This algorithm replaced the initial proposal with an exponential time complexity, which was too slow for a production application. Developed the original set of inputs or priors for this model based on cyber risk data, industry studies, original research, and associations of sources of revenue between companies (correlation matrix).

Guidewire: Lead catastrophe modeler responsible for Mass Ransomware and Mass Data Breach model development. Use PyMC to develop a Bayesian Generalized Linear Mixed Effects model with random intercepts for industry categorization. PyMC uses Hamiltonian Monte Carlo, an advanced Bayesian MCMC technique, to fit regression models.

Capital Group: Used PyMC to fit a generalized linear model using MCMC techniques (Hamiltonian Monte Carlo). Model identified economic factors associated with Plan Premier Participant counts.

Indiana University April 2024: Giving two guest lectures on Stan software and statistical computation research, including software package development.

Indiana University 2019: Co-instructor for PBHL-B 646 Advanced Generalized Linear Models with Wanzhu Tu, Spring 2019, Indiana University.

Machine learning

Capital Group: Developed a deep learning autoencoder in H2O to reduce the dimensionality of the characteristics of a data set of 270k financial advisors in the United States. The resulting dataset was used to run KNN on advisors attending Capital Group sponored events. The KNN model matched the test set of advisors to a pseudo-control set of advisors with similar characteristics. A comparison of these sets of advisors was estimated using a GLMM to control for a set of covariates and estimate sales lift as a result of these events.

Capital Group: Developed a machine learning algorithm (based on C5) to identify retirement plans at risk for attrition. At-risk plans, as identified by the model, are twice as likely to leave as non-risk plans.

Revelant: Used text mining to analyze open-responses to survey questions from a call center. This analysis was used to identify drivers of caller satisfaction.

Guidewire: Using PyTorch to develop an autoencoder to impute firmographic and technographic features for companies we lack such data (experimental).

Guidewire: Led second-place team in Guidewire hackathon. Developed NLP model to mine cyber attack data for new ransomware gangs and categorizing types of attacks. This project mined from open text data using the named entity recognition (NER) features in the Spacy package.

Independent consulting: Developed machine learning model (random forests) for a research project on developing "soft" sensors in an environmental application. The idea of this project was to estimate the feasibility of using data to replace hard sensors in an environmental application, saving money on installing and maintaining these parts. Over 400 features were evaluated in this model. PCA was used to help with dimensionality reduction.

Academic coursework: Developed a convolutional neural network model in Tensorflow to classify minerals based on spectral imaging from the planet Mars.

Independent consulting: Developed a regression model to match the automated pricing results from a website. The model was developed using statistical software and translated to Excel for the client's use.

Cloud computing

Guidewire: Completed redesign and implementation of a risk model for cyber catastrophe events for the Cyence product. This model aligns risk of attack to a combination of historical experience, cyber expert judgment, and customer feedback. Improved the computational efficiency of this model from a runtime of 12 hours for 10k simulations to 40 minutes using coarse-grained parallelism and vectorized numpy calculations in AWS EMR.

Capital Group: Used Databricks to perform data engineering using PySpark on multiple data sources on events for financial advisors. This process was used to estimate the sales lift of financial advisors attending these events.

Optimization

Capital Group: Developed a custom optimization model in Python to estimate the number of Shareholder Services associates needed if processing work was outsourced. Model influenced decision to retain processing work in-house. This model was based on the discrete event simulation of calls into a call center by time of day.

Capital Group: Identified the best sales territory locations by optimizing sales market against a minimum driving distance.

Capital Group: Developed a mathematical model to estimate the steady-state account volumes based on queuing theory (M/M/∞ queue).

Data Mining

Guidewire: Developed a data mining process to identify the number of mass catatrophic events from a dataset of approximately 5000 historical incidents. This process combined record counts and basic NLP to filter the data down to 35 candidates. Manual research and consultation with cyber experts helped to narrow down this set to the final set of incidents used in the catastrophe model.

Capital Group: Extensive data engineering to combine disparate data sources for the Gatekeeper project. Some of these data sources included Morningstar mutual fund data, internal salesforce CRM data, third party marketshare data, and recommended lists typically in PDF form. The purpose of this project was to produce a reporting and analytics dashboard for use by the Gatekeeper sales team. So-called Gatekeepers work at the home offices of major financial services firms. Used SQL (SSMS) primarily, some R and Excel for this project.

Revelant: Leveraged data to identify opportunities for quality and efficiency improvement opportunities for the 1-800-MEDICARE service center. One project involved mining large log files from customer service representatives with data on how representatives migrated through a software application to find certain scripts. I used the TraMineR package in R to identify the most common click and search patterns and make recommendations for improvements based on these patterns.

Independent Consulting: Performed multiple statistical analyses and maintained an Access data base for Healthcare Performance Consulting. The database was used to produce a standard set of reports on a monthly basis.

Skills

- Statistical Computation
- Bayesian Analysis: Markov Chain Monte Carlo, Hamiltonian Monte Carlo
- R, Python, Stan, PyMC, C++, SQL, Azure Databricks, Tensorflow, Tableau, AWS Cloud, EMR, PySpark

PUBLICATIONS

Thomas, S. and Tu, W. (2022). *Riemannian Monte Carlo*. In Wiley StatsRef: Statistics Reference Online (eds N. Balakrishnan, T. Colton, B. Everitt, W. Piegorsch, F. Ruggeri and J.L. Teugels). doi:10.1002/9781118445112.stat08392

Thomas, S. and Tu, W. (2020). *Hamiltonian Monte Carlo*. In Wiley StatsRef: Statistics Reference Online (eds N. Balakrishnan, T. Colton, B. Everitt, W. Piegorsch, F. Ruggeri and J.L. Teugels). doi:10.1002/9781118445112.stat08243

Thomas, S. & Tu, W. (2021) *Learning Hamiltonian Monte Carlo in R*, The American Statistician, DOI: 10.1080/00031305.2020.1865198

Green, Brice and Thomas, Samuel, *Inference and Prediction of Stock Returns using Multilevel Models* (August 31, 2019). Available at SSRN: https://ssrn.com/abstract=3411358 or http://dx.doi.org/10.2139/ssrn.3411358

In Progress

Thomas, S. and Tu, W. 2020. *An R package for Bayesian Multivariate Generalized Additive Models*.

Talks

A Bayesian Analytical Software Based on Hamiltonian Monte Carlo. Regenstrief Institute, 12/4/2019. https://www.youtube.com/watch?v=sBA3lAoNhto

Using Fourier Series to Model Daily Seasonal Patterns of Redemptions. Capital Group Companies, Data Science Interest Group, 2018

Improving Capacity and Financial Planning, a Guide to Business Forecasting with Alteryx. Inspire 2016 Alteryx Conference, San Diego, CA.

Predicting At-Risk Plans Using the C5 Algorithm. Capital Group Companies, Data Science Interest Group, 2015

UseR 2012 at Vanderbilt University. UseR 2012 Vanderbilt University.

Software developed

hmclearn: CRAN. An R package to fit statistical models with Hamiltonian Monte Carlo. https://cran.r-project.org/web/packages/hmclearn/index.html

bayesGAM: CRAN. An R package to fit semiparametric regression models using Hamiltonian Monte Carlo. https://cran.r-project.org/web/packages/bayesGAM/index.html

mlts: Proprietary. An R package to automatically develop forecasts and perform cross-validation for bottoms-up forecast models. Internal package for Capital Group Companies.

Employment

Principal Data Scientist

Guidewire Software 2022–present 40+ hours/week

Senior Data Scientist

Capital Group Companies 2008, 2010–2022 40+ hours/week

Principal, Data Scientist

Revelant Technologies 2008–2010 40+ hours/week

Associate Actuary

Milliman 2007–2008 40+ hours/week

Assistant Actuary

WellPoint/Anthem 2005–2007 40+ hours/week

Independent Consulting

2006-2022 1-10 hours/week Clients include: Medscape/WebMD, Healthcare Performance Consulting, and Arks32

Teaching

Co-instructor, Biostatistics

Indiana University Two guest lectures for Advanced GLM class Apr-2024

Indiana University 2019

PBHL-B 646 Advanced Generalized Linear Models

Adjunct Professor, Mathematics

Ivy Tech University 2006–2014

Courses taught:

- Math 211 Calculus
- Math 136 College Algebra
- Math 135 Finite Math