Siyi Wei

Master Student of Statistical Sciences at University of Toronto

(647) 870-8349 | weisiyi2@gmail.com | www.wei-siyi.com | superp0tat0 | siyi-wei

Education

University of Toronto — *Toronto, ON*

Master of Science in Department of Statistical Science — *Aug 2020 - Aug 2021 (Expected)*

- Relevant Courses: Applied Statistics, Monte Carlo, Data Science Methods, Spectral Features, Applied Deep Learning
- Relevant Activities: Teaching Assistant for Undergraduate Probability Theory, Introduction to Machine Learning, Advanced Statistical Learning, Linear Algebra, Statistics and other courses.

Honours Bachelor of Science — *August 2015 - April 2020*

- Specialist in Statistics, Statistical Machine Learning and Data Science Stream, Minor in Computer Science
- Relevant Courses: Big Data Analytics (Rank: 1/40), Intro & Advanced to Machine Learning (A), Deep Learning (Audit), Software Engineering (A), Data Collection(Rank: 1/300), Database(A+), Data Structures(A-), NLP & Algorithms Specialization (Online)
- Relevant Activities: Teaching Assistant for Introduction to Machine Learning, Linear Algebra, Probability, Statistics and other courses.

Skills

Languages	Python(Scikit-Learn, Pandas, Numpy, PyTorch, PySpark), R(Tidyverse, RShiny, mlr3, qqplot2), SQL, Git, C#
Technologies	Spark, Hadoop, Amazon Web Services, TensorFlow, Linux (Ubuntu), Jenkins, Selenium
Technical Skills	Statistical Modeling, Machine Learning, Data Mining, Data Visualization, Time Series Analysis, A/B Testing, CI/CD

Experience

Google Summer of Code — *Toronto, ON*

Data Scientist Student Intern — *May 2021 - Present*

- Develop mlr3 fairness package to help mlr3 users to detect and correct the fairness problems in multiple approaches. Implement debiasing strategies as pre and post processing PipeOperators in the style of the mlr3 pipelines, which supports some bias mitigation algorithms like reweighing, equalized Odds Postprocessing and Adversarial Debiasing.
- Create popular fairness metrics like confusion matrix, AUC and ROC curves then mitigate into other mlr3 packages. Implement visualizations and a clear API for auditing using either ggplot2 or inherit from other fairness packages.
- Create an introduction vignette, demos and a well documented wiki page for debiasing algorithms to showcase the new package.

University of Toronto Scarborough — *Toronto, ON*

Research Assistant — *May 2019 - April 2020*

- Built interactive Data Visualization using R Shiny to help professors detect suspicious activities from students tests. Using MOSS (anti-plagiarism system) data to locate different groups of students and report their grade distribution across semesters and assignments.
- Using logistic regression to analyze the relations between multiple factors and student's pass/fail status. Applied PCA regression to ease the collinearity founded in dataset. Design hypothesis tests to verify potential hypotheses made from exploratory data analysis.

Rakuten Kobo — *Toronto, ON*

QA and Developer Intern — *January 2019 - April 2019*

- Worked on development and QA for new Kobo Audio subscription services using Jenkins, Ruby and selenium. Wrote SQL and Ruby scripts for daily regression testing on production stage. Collaborate with Manager to design A/B Testing to improve current features.
- Responsible for the QA automation and bug tracking system for 4 months, prevent service failure for multiple times and received praise from team lead and promoted to work on web development and ETL pipeline for new subscription.
- Developed web page and ETL pipeline using C# and MongoDB via CI/CD pipeline, automated the data collection process and developed new metrics for subscription data. Received praise from Big Data Team for increasing their efficiency of the data collection process.

Projects

Correct Algorithmic bias using Bayesian Hierarchical Model

- Research on the famous COMPAS dataset and identify the algorithmic bias caused by COMPAS algorithm and classification trees on the ethnicity. Found the disadvantageous position of African American by reporting the subgroup false positive and false negative rates.
- Compare the accuracy and subgroup performance of logistic regression and classification trees. Working on the Exploratory Data Analysis and identifying the bias could come from different subgroup patterns like age and gender distribution.
- Implement different hierarchical logistic regression models with prior information in stan to correct the algorithmic bias. Decide using a three stage bayesian hierarchical model to adjust the subgroup fairness. Successfully balanced the subgroup fairness by reducing 67%FP-37%FN to 48%FP-52%FN for African American with only 1.7% loss in total accuracy. Further extend this project to GSoC 2021.

Unsupervised pre-training with spectral PCA (Graduate Research Project)

- Research on recent published paper on CNN's working principle with random labelled data. Worked on transfer learning using PCA combined with neural networks. PCA preprocessed data could converge 20% faster than original data and with more interpretability.
- Conduct further research with professor and raise hypotheses on spectral features that could be learned by convolutional layers. Working on empirical study with Google Colab, pytorch, numpy and other machine learning libraries. Extract spectral features using different approaches like Kernel PCA or Convolutional layers. Those spectral features could be useful in few shot learning or transfer learning and contribute to the protection of user privacy.