Tengfei Cui

Email: ctengfei2021@163.com · Tel: (+86) 13064027783 · GitHub: https://github.com/ITCUI-XJTLU

EDUCATION BACKGROUND

University of Washington, Seattle (UW), USA 09/2023 – 06/2025
MS Biostatistics (Thesis Track)

Xi'an Jiaotong-Liverpool University (XJTLU), China 09/2019 – 06/2023
BSc Applied Mathematics | Cum GPA: 3.9/4.0 | 2020-21, 2021-22 University Academic Excellence Award (Top 5%)

Coursework (XJTLU): Advanced Linear Algebra | Multivariable Calculus | Analysis | Partial Differential Equations | Applied Probability | Probability and Statistics | Abstract Algebra | Population Dynamics | Optimization Theory | Operation Theory | Mathematical Analysis | Numerical Analysis | Java Programming |

Coursework (Coursera): Deep Learning | Supervised Machine Learning: Regression | Applied Data Science

SKILLS

Programming Languages: R | Java | Python | MATLAB
Libraries: NumPy | SciPy | Pandas | TensorFlow | Scikit-Learn | Pytorch | Matplotlib
Software: Microsoft Office | Visual Studio | PyCharm | Git | Jupyter Notebook

RESEARCH EXPERIENCES

Survival Analysis on MIMIC III Database
***Research Assistant* | Xi'an Jiaotong-Liverpool University, Advisor: Prof. X. J. Zhu** 5/2022 – 10/2022

- Building survival analysis models to predict survival time of patients with malignant neoplasm of liver
  - Extracted and processed lab test data of patients with malignant neoplasm of liver and get an insight of dataset by drawing K-M curves
  - Built Cox PH models with a single predictor, chose potential risk factors, and built Cox PH models with multiple predictors
  - Improved performance of Cox PH model by best subset selection, AIC value, and linear regression models
  - Used Cross-Validation and several models evaluating indexes including ROC curves and time-dependent AUC curves to perform model comparison and selection; Built nomogram of the constructed models for prediction

Automated clinical coding by Deep Learning
***Research Assistant* | University of Hong Kong, Advisor: Prof. L. Q. Yu** 5/2022 – 10/2022

- Building a new automated medical code prediction system which could transform discharge summaries of the MIMIC III dataset into ICD-9 codes
  - Tried the current dominant CNN-based models for the medical code prediction task, such as CAML, MultiResCNN, JointLAAT and MSMN
  - Finetuned several outstanding pre-trained language models, including BERT, BioBERT, ClinicalBERT and BlueBERT, in multilabel text classification tasks and compared the performance with the CNN-based models
  - Tried new long sequence transformer models, such as Clinical-Longformer and Clinical-BigBird, to improve the performance of transformer-based models
  - Designed new hierarchical fine-tuning architectures to extend the maximum input sequence of transformer-based models
  - Integrate knowledge or semantic information in coding classification systems and ontologies, such as CCS and UMLS, to improve the performance of deep learning models

Impact of COVID-19 on Academic Research – A Data-Driven Study
***Research Assistant* | Nanjing University & University of Virginia, Advisor: Prof. Q. Z. Du** 12/2021 – 11/2022

- Collectively leveraged an array of machine learning techniques to study the impact of the ongoing COVID-19 pandemic on the number and quality of the recently published academic papers
  - Developed a streamlined workflow for key words extraction from academic publications and the associated creation of a knowledge graph
  - Leveraged TFIDF numerical statistics for keywords extraction and latent semantic analysis (LSA); clustered semantic vectors using cosine similarity, allowing summarization of semantic themes
  - Utilized regular expression, n-gram model, and self-attention mechanism to drive pattern extraction, with a focus on converting semi-structured resume data into fully structured data
  - Facilitated knowledge graph construction using Neo4j graph database, with a focus on subject-predicate-object three-element groups

Find Members of Praesepe Star Cluster

***Research Assistant* | XJTLU, Advisor: Prof. X. Y. Pang** 02/2021 – 04/2021

- Designed, implemented, and validated a data-driven workflow in Python to search for the members of the Praesepe star cluster based on an array of features, including space location, velocity, and age:
  - ✧ Collectively employed NumPy, Pandas, Matplotlib, and Seaborn to drive data preprocessing, visualization, exploratory data analysis, and pattern extraction
  - ✧ Performed member selection clustering and member properties clustering using Hertzsprung–Russell diagrams
  - ✧ Bolstered hands-on skills for selecting appropriate descriptive data analytics method following real-world situations, improving model performance using real data, and communicating outcomes of projects

Sequence to Sequence Learning with Neural Networks

***Research Assistant* | XJTLU, Advisor: Prof. H. K. Li** 12/2020 – 02/2021

- Implemented in Python a newly published general end-to-end approach to sequence learning that makes minimal assumptions on the sequence structure for machine translation and text summarization purposes:
  - ✧ Used a multilayered Long Short-Term Memory (LSTM) to map the input sequence to a vector of a fixed dimensionality, followed by using another deep LSTM to decode the target sequence from the vector.
  - ✧ Systematically characterized the new model's performance in
    - ✓ English to French translation based on WMT'14 dataset, focusing on testing its accuracy and robustness in handling out-of-vocabulary words, long sentences, and sensible phrase and sentence representations that are sensitive to word order and are relatively invariant to the active and the passive voice
    - ✓ Text summarization based on Stanford Sentiment TreeBank (SST2) dataset, focusing on benchmarking against other classical NLP models, including Word2Vec, Bert with self-attention mechanism, etc.

INTERNSHIP EXPERIENCES

***Data Analyst Intern* | Ruguo Technology Co., LTD, Baoding, China** 06/2021 – 09/2021

- Investigate the technical architecture and data architecture of the Internet of Vehicles system
- Preprocessed data to ensure the quality of the data, generated data analysis reports from the perspective of operation, mined the data value from various dimensions, and provided references for determining the business value and operation work based on the data analysis results
- Assisted the technical team in realizing the architecture design and functions of data access, data processing, data mining, and data visualization on the big data algorithm platform and supported the project implementation and implementation

***Data Analyst Intern* | Wuhan Tianheng Information Technology Co. LTD, Wuhan, China** 01/2021 – 03/2021

- Predicted passenger flow of an urban rail transit system for the government with the aim to support the rational allocation of transportation resources, optimization of the operational plan, and evaluation of economic benefits:
  - ✧ Applied support vector regression (SVR) with various kernels, ridge regression, Lasso regression, and elastic-net regression models to enable prediction; evaluated and compared the performance of each method
  - ✧ Built a workflow to drive data collection and preprocessing, with a focus on effectively handling large-scale, multi-source, heterogeneous data

DATA SCIENCE PROJECTS EXPERIENCES

Kaggle: Marketing Data Analysis on Bike Sharing Business 10/2020 – 12/2020

- Performed descriptive and predictive data analyses on bike sharing business in China:
  - ✧ Analyzed the e-bike return rates using various visualization libraries in Python, with a focus on performing time series and geospatial analyses so as to identify insightful patterns and trends; optimized operational strategy based on analyses outcome to promote e-bike returns
  - ✧ Predicted the demand for bike sharing service in an array of metropolitan cities in China based on various supervised learning models; trained k-nearest neighbors, logistic regression, support vector machine, gradient boosting decision tree, random forest models to drive prediction; performed bagging, voting classification and stacking to drive ensemble learning so as to improve predictive accuracy

SOFTWARE DEVELOPMENT EXPERIENCES

Development of a Star Wars Video Game in Python 07/2020 – 08/2020

- Designed and implemented a Star Wars video game in Python, covering power-ups, an array of level layouts, and layered visuals:
  - ✧ Optimized various components in the game design to enhance the player experience, including mechanics, feedback, pacing and interface
  - ✧ Systematically leveraged 1) various object-oriented features and design patterns in Python to make the app modular, extensible, robust, and maintainable, 2) the unified modeling language to guide software design, 3) Github to promote team collaboration, and 4) unittest library to streamline and automate testing