Yunseok Jang

ryanjang123@gmail.com https://github.com/yunseokj1372

EDUCATION

New York University, New York, NY Expected Graduation: May 2023
- Master of Science, Data Science: GPA 3.8

University of Southern California, Los Angeles, California May 2021
- Bachelor of Science, Applied and Computational Mathematics: Cumulative GPA 3.9 (Magna Cum Laude)

TECHNICAL SKILLS

Programming Language: Python, MySQL, MATLAB
Tools: Numpy, Pandas, Matplotlib, sklearn, PySpark, Dask, HDFS
Relevant Coursework: Intro to Data Science, Probability and Statistics for Data Science, Computational Linear Algebra and Optimization, Machine Learning, Natural Language Understanding and Computational Semantics, Big Data

RESEARCH & DEVELOPMENT EXPERIENCE

University of Southern California, Los Angeles, California January 2020 – May 2020
Independent Research
- Chaired a team of independent research of Computational Probability for building probabilistic models.
- Simulated randomness through MATLAB in a variety of distribution by method of Pseudo Random Number Generator which is the foundation of modern random models.

PROFESSIONAL EXPERIENCE

NYU Tandon Department of Chemical and Biomolecular Engineering, Brooklyn, NY June 2022 – August 2022
Research Assistant / Data Scientist
- Build machine learning models to predict the key parameters or physiochemical properties of enzyme.
- Encode amino acid residue sequence with one-hot encoding, Protein Analysis Bag-of-Words, Bigram, BLOSUM62, and random frozen embedding to reformat sequence into vectors with numerical entries.
- Conduct resampling methods such as cross validation and jackknife to evaluate the performance of the machine learning models used.
- Model random forest regression and gradient boosting to predict three outputs: kcat and km, half-life, and ki.

NYU Langone Health, Remote May 2022 - Present
Research Data Associate
- Preprocess Cardiac Arrest Survivorship experience text data collected from NYU Langone Health with CountVectorizer, TfidfVectorizer, Sentence Tokenizing, and Stemming and Lemmatization.
- Produced a stratified split of dataset to make a balanced dataset.
- Build supervised models such as Logistic Regression, Random Forest Classifier, and BERT to predict classification of experiences as relevant versus irrelevant experience of Cardiac Arrest Survivorship.

DATA SCIENCE PROJECTS

Recommender System April 2022 – May 2022
- Partitioned MovieLens datasets into train, validation, and test set based on user ID and timestamp with PySpark SQL. (small: 9000 movies and 600 users, large: 58000 movies and 280000 users)
- Predicted customers' movie preference by building Latent Factor Model through both ALS model of PySpark Collaborative Filtering and Lenskit.
- Achieved 0.0509 for NDCG (normalized discounted cumulative gain) metric on PySpark ALS model and 0.0997 for Lenskit ALS model.

Toxicity Detection Dataset with r/WallStreetBet Comments March 2022 – May 2022
- Scraped comments from reddit WallStreetBet community which our team defined as a toxic community.
- Labelled comments based on features provided by Perspective API to create a dataset that showcases the limitations of several state-of-the-art language models.
- Tested on language models such as BERT, RoBERTa, DeBERTa, and GPT-3 few-shot and zero-shot to check performance on the dataset.
- Evaluated the dataset on three metrics which are F1 score, precision, and recall, resulting low scores for all three.

Kaggle Competition: WiDS Datathon 2022 Feb 2022
- Constructed gradient boosting model using XGBoost to predict Site EUI (Site Energy Usage Intensity) over 100k observations of building energy usage records.
- Attained 35.227 RMSE test set score for our team's final submission.

AWARDS

Jennifer Battat Scholarship – for excellence in the field of mathematics