Yongrui Chen

Email: yc910@georgetown.edu | Tel: +1 (703)229-7278 | LinkedIn | Portfolio

SKILLS

Analytical Techniques:

- Machine Learning and Deep Learning Algorithms: SVM, Naïve Bayes, Decision Tree, Random Forest, XGBoost, Regression (Linear, Ridge, Lasso, Logistics), ANN, CNN, RNN, LSTM, Transformers
- Principal Component Analysis (PCA), K-Means, Clustering, Associate Rule Mining (ARM), Feature Selection, Model Evaluation
- A/B Testing, Hypothesis Testing, Data Wrangling, Exploratory Data Analysis

Programming Skills: Python (TensorFlow, PyTorch, PySpark, SciPy, scikit-learn, Pandas, etc.), R (tidyverse, caret, etc.), SQL, C, HTML

Tools: Jupyter Notebook, MySQL Server, Hive, AWS, Hue, GitHub, Tableau, SPSS, CLion, MS Office

EDUCATION

Georgetown University, Washington, DC, US — Aug. 2021 - May 2023

Master of Science in Data Science and Analytics — GPA: 4.0/4.0

- **Courses:** Natural Language Processing, Big Data and Cloud Computing, Statistical Learning, Advanced Data Visualization, Optimization
- **Achievements:** Returning Scholarship Competition (2nd place),

Beijing Jiaotong University, Beijing, China — Sept. 2016 - Jul. 2020

Bachelor of Economics in Finance — GPA: 3.63/4.00

- **Courses:** Calculus, Linear Algebra, C Language, Applied Statistics, Data Analysis, Financial Analysis

WORK EXPERIENCE

Rise Against Hate, Data Analyst Intern; Washington, D.C., USA — Jun. 2022 - Now

- Constructed Database of New Jersey Arrest and Police Use of Force Data, extracted and conducted statistical analysis using MySQL
- Conducted a Hypothesis Test to verify the significant difference of arrest rate among different races and visualized the results and collaborated with other colleagues to output the Arrest Justice Report of New Jersey
- Built a regression model to analyze the impact of different features on the level of use of force and designed a dashboard via Tableau
- Visualized the development of reproductive rights of all states in USA and assisted the research team to deliver the report

Kuaishou Technology, Data Analyst Intern; Beijing, China — Dec. 2020 - Jun. 2021

- Analyzed more than 280M users' behaviors using Hive, to calculate key indicators including retention rate, online duration, etc. and constructed a regression model to predict the CTR and Completion Rate of the Useful Content (5-fold CV Accuracy: 82%)
- Designed A/B testing on the landing page to find out the impact of different landing pages on the user's following video consumptions to provide optimization suggestions on products, strategies, and commercialization
- Classified more than 50M E-commerce customers into five groups and defined the level of trust for each group based on K-means clustering and Hierarchical clustering
- Conducted a Hypothesis Test to find out the significant difference of User's Search Behavior between loyal and disloyal users
- Tracked metrics for user activity, delivered weekly and monthly reports and conducted attribution analysis for abnormal changes

Shenwan Hongyuan Securities, Research Analyst Intern; Beijing, China — Nov. 2019 - Jan. 2020

- Tracked the commercial retail sector, compiled the Third-quarter Industry Key Company Performance Summary, and completed the quarterly reports of listed companies
- Built a Mixed Effects model to predict the profit, collaborated with analysts to write and deliver tailored reports for senior leadership, including 2021 Business and Retail Strategic Planning and Forecast Report
- Collected emerging market data, and compiled them into more than 20 daily and weekly reports to provide a perspective on the commerce and retail industry sectors to key external stakeholders

PROJECT EXPERIENCE

Reddit Sentiment Analysis [More] — Jan. – Apr. 2022

- Set up environments in AWS including launching EC2 instances, creating EMR clusters, etc. to enable big data analysis
- Conducted EDA, and built a pipeline to normalize, tokenize and lemmatize more than 9 million reddit posts using PySpark
- Assembled a pipeline to conduct sentiment analysis using a pre-trained sentiment deep learning model from John Snow Labs
- Classified posts into 5 topics with Latent Dirichlet allocation (LDA) and extracted keywords of each topic
- Constructed a Long Short Term Memory model with Early Stopping using different look-back windows to predict Nasdaq Index using post sentiment, topic, and other related variables, and the R-Squared of the best model (60 days look-back) is 0.9881

ML-Powered Heart Disease Screening [More] — Jan. - Apr. 2022

- Conducted data munging and worthwhile EDA on more than 300 thousand rows of data collected from CDC
- Constructed pipelines of different ML models including Logistic Regression, Random Forest, XGBoost, etc., performed grid search for the best parameters, and evaluated each model via 10-fold Cross Validation using J Index, Kappa, ROC, etc.

Steam Game Review Classification Using NLP and Text Modeling [More] — Nov. - Dec. 2021

- Explored the data through different visualizations and conducted data wrangling on more than 330 thousand rows of records to remove the meaningless reviews and stop words in the contents. Created a preprocessor class to strip, stem and tokenize the data
- Utilized BERT transformer to encode the data and create attention masks and batches and constructed a Recurrent Neural Network model to train on the batches. The average training loss descended from 0.12 to 0.03 in ten epochs
- Defined a function to conduct dimensionality reduction using Principal Component Analysis (PCA) and visualize all the layers of the Neural Network in every epoch and evaluated the model in terms of accuracy, F-1 score, etc.