Khanh (Chris) Tran

khanh.tran@simon.rochester.edu | (716) 513-4637 | linkedin.com/in/chriskhanhtran | chriskhanhtran.github.io

EDUCATION

UNIVERSITY OF ROCHESTER   **Rochester, NY**
Master of Science in Business Analytics (STEM); GPA: 3.96/4.00   Dec. 2020 (Available for full-time from May 2020)
- Coursework: Core Statistics, R Programming, Predictive Analytics with Python (Machine Learning), Causal Analytics with R (A/B Testing), Social Media Analytics (NLP), Database Management (SQL, Cypher), Big Data (Hive, Spark), Pricing Analytics

NIAGARA UNIVERSITY   **Niagara University, NY**
Bachelor of Business Administration in Accounting; GPA: 3.99/4.00   2019
- Dean's List (all attended semesters); Top 5 graduated student
- Coursework: Business Analytics, Linear Models, Management Information Systems, Econometrics

EXPERIENCE

Skim AI Technologies, Inc.   **New York City, NY**
Machine Learning Engineering Intern (NLP)   Oct. 2019 – Present
- Trained and deployed MobileBERT for document summarization on mobile devices, achiving 5.5x faster speed while attaining 95% of BERT-large's performance.
- Trained deep learning models (Transformers, CNN, RNN-LSTM) for document classification and named entity recognition tasks to extract information from legal documents, achieving 0.95 F1 score and saving client hundred hours of manual labeling.
- Pretrained Spanish BERT and ELECTRA models from scratch on 18 GB of Open Super-large Crawled Corpus (OSCAR) using multiple GPUs on AWS EC2 instances, achieving state-of-the-art results on Spanish benchmarks.
- Researched latest NLP and CV papers on regular basis, reported findings to CTO and wrote codes to reproduce research results.

Aurubis Buffalo, Inc.   **Rochester, NY**
Data Scientist (Capstone Project)   Feb. 2020 – May 2020
- Developed data preprocessing and regression pipeline to predict yield percentage from production of coils into finished products. The developed LightGBM regressor achieved highly accurate predictions with 4% MAE, improving scheduling efficiency and significantly lowering production and inventory cost.
- Performed data labeling, data exploration, data cleaning, feature engineering on coil history, and hyperparameter tuning with cross validation to optimize accuracy of regressor.
- Built web app to deploy developed pipeline that can quickly generate yield predictions for future production.

Tax Technologies, Inc.   **Buffalo, NY**
Tax Intern   Mar. 2019 – July 2019
- Provided technical supports to Fortune 500 clients utilizing Tax Series – an all-inclusive SaaS global data collection, tax compliance and provision software, and assisted on implementation engagements for new clients.
- Conducted essential application diagnostics on client financial data, including periodically generating technical reports, maintaining data integrity and monitoring client databases.
- Conducted in-depth research on tax regulations and e-file requirements in 32 states and four foreign countries.
- Performed application testing and collaborated with software engineers to build enhancement update for Tax Series.

PROJECTS & COMPETITIONS (more details at https://chriskhanhtran.github.io/)

Social Media Analytics for Airlines: Fine-tuning BERT for Sentiment Analysis
- Vectorized tweets with pretrained fastText word vectors and trained CNN model for sentiment classification using PyTorch.
- Fine-tuned BERT model to detect negative tweets, achieving 10% accuracy improvement over TF-IDF.

Kaggle Competition: Advanced Regression Techniques in House Price Prediction – Top 0.6% on leaderboard
- Performed comprehensive data analysis, data cleaning and feature engineering on Ames, Iowa housing dataset.
- Ensembled Ridge, Lasso, XGBoost, and LightGBM models to predict house prices.

Humana-Mays Healthcare Analytics Competition – Top 50 out of 460 teams
- Preprocessed 7M EHR records of 20K patients, performed feature engineering from past diagnoses and medical claims.
- Built LightGBM model to predict patients with long-term opioid therapy, achieving 0.88 AUC score.

Business Analytics Competition & Conference 2018 @ Manhattan College, NY – 2nd best research poster out of 18 colleges
- Led team of 4 students to analyze 6M records of NYC governmental spending data and present key findings.
- Built statistical models to predict governmental key performance metrics and educational quality.

SKILLS

Programming: Python (NumPy, Pandas, Scikit-learn, PyTorch, TensorFlow), Big Data (Spark, Hive), R, SQL, Cypher, MATLAB
Others: AWS EC2, GCP, Tableau, Matplotlib, Seaborn, SAS, SPSS, Adobe Suite