Saisamrit Surbehera

+1 (206) 612 7767
ss6365@columbia.edu
saisamrit-surbehera
saisurbehera

Academia

Education

Columbia University

Columbia Engineering ○ Masters of Science in Data Science ○ Class of 2022

University Of Michigan Ann Arbor

School of Information ○ Bachelors of Science in Data Science ○ Class of 2021

Graduated with James Scholar and a GPA of 3.9/4.00.

Research

From Human Days to Machine Seconds: Automatically Answering and Generating Machine Learning Final Exams
Text to graphics by program synthesis with error correction
A dataset for learning university STEM courses at scale and generating questions at a human level

Industry

Walmart Search
2022 and 2023-present

ML Research Scientist, Hoboken, NJ

Search Experience Team focusing on Guided Navigation and Facet Curation

- Developed synthetic data generation pipelines for customer queries using fine-tuned LLMs and RAG-based tools to train production-ready distilled models for significantly improving search refinement accuracy
- Designed and deployed an item attribute matching system using open-source LLMs with Chain-of-Thought reasoning, enhancing indexing precision and ranking relevance

Algorithmic Re-ranking Team focusing on cold start and boosting

- Implemented an transformer based price prediction model based on item level features including item text, images and engagement
- Developed and deployed multiple cold start pipelines with PySpark to identify underexposed cold start items and generate embeddings using item information
- Worked on multiple Click Through Rate(CTR) Predictions XGBoost models fine-tuned for newly added items

American Family Insurance
2021

Machine Learning Intern, Madison, WI

- Created an Information Extraction pipeline (IE) to extract legal definitions and entities from AmFam Policy Form contracts with an 83%. Pipeline turned forms into structured knowledge graphs which is used for answering queries in internal Knowledge App tool.
- Designed and trained multiple co-reference resolution transformer models using semi-structured policy forms.

University of Michigan, Giving
2020

Data Scientist Intern, Washington, DC

- Developed an Entity Resolution (ER) system and knowledge graphs for 700,00 living alumni across multiple databases with different level of information for donation campaign targeting
- Integrated Graph Embeddings using DeepWalk / Node2Vec, Tf-idf and word embedding to beat existing internal ER benchmark by 16%.
- Engineered a full stack search application to find alumni's interests and developed unique knowledge graphs with over 1 million nodes to increase alumni reach.

Academic Projects

AAA Insurance

Implemented ensemble logistic regression, Modified Multi-variate Gaussian (MVG), Modified Randomized Under sampling (MRU), and Random Forest to select the best model to predict the number of breakdowns for a model in a zip-code with 85% accuracy

University of Stellenbosch

Lead a team of four members for over nine months to build a re-modeled business conflict barometer, an automated and analysis tool of textual data and variables within date, location, etc. to determine when and how conflict levels are rising or falling based on a given situation. Pipeline included several scrapers and multiple BERT models to predict conflict scores

Skills

Technical

General Programming (Java, Python) ○ Data (SQL, PySpark) ○ Deep Learning (PyTorch)