KABIR THAKUR

(571) 591-8688 • kathakur@syr.edu • kabirthakur.github.io/portfolio/ • LinkedIn

EDUCATION

Syracuse University, School of Information Studies, Syracuse, NY — Sep 22 – May 2024
M.S. Applied Data Science (Specialization: NLP) *Coursework:* NLP | Machine Learning | Big Data Analytics Artificial Intelligence Algorithms | Text Mining with LLM | Business Analytics | Statistics | Advance Database Management | Data Science

Central University of Punjab, Department of Computational Sciences, Punjab, India — Nov 2020 – June 2022
M.S. Physics (Computational Physics) *Coursework*: | Python programming | FORTRAN | Mathematics for Computational Sciences

Shiv Nadar University, Department of Physics, NCR, India — Sep 14 – May 2018
B.S. Physics (Research) *Coursework*: Linear Algebra | Calculus II | Data Management and Analytics | Python for Physics

SKILLS

Programming Languages: Python, R, SQL (MS SQL, MySQL, SQLServer, HQL, NoSQL), Bash Scripting, FORTRAN, MATLAB
Machine learning: Regression, Random Forests, LGBM, XGBoost, Time Series Forecasting, SVM, Decision Trees, kNN,
Deep Learning: CNN, RNN, LSTM, Huggingface Transformers, Transfer Learning, Reinforcement Learning
Big Data: Hadoop, Hive, Spark, Cassandra **Streaming:** Kafka, **Cloud Services:** AWS (Certification), Azure Data Studio, GIT
Libraries: TensorFlow, PyTorch, Pandas, scikit-learn, PySpark, NumPy, NLTK, Spacy, ggplot2, dyplr, caret, matplotlib, seaborn
Certification: AWS Cloud Practitioner

WORK EXPERIENCE

Tutor for Student Athletes – Stevenson Educational Center, Syracuse University — Aug 2023 – Present
- Tutored 12 undergraduate student athletes in courses on **Data Analytics in R, probability, statistics, and calculus**.
- Facilitated an average grade improvement of 25% among tutored students by developing tailored learning strategies.

Data Science Researcher – Decision Science, JPMorgan Chase & Co, London — Feb 2023 – Jun 2023
- Collaborated with 2 members to integrate algorithmic decision making with expert opinions using a **Bayesian Framework**.
- Tested 5 different methods of sharing information between human experts and ML models monitoring performance indicators.
- Showcased superior performance of information sharing through Bayesian learning by improving F1 score by 7%.
- Built a deferral system where algorithms can defer to expert when they have low confidence in an outcome.
- Co-Authored a peer reviewed tiny paper with the team for ICLR 23 - Dynamic Human AI Collaboration

PROJECTS

Skillspotter: Named Entity Recognition on Job Descriptions – Python, PyTorch, NLU, NLI — Sep 2023 – Dec 2023
- Created a dataset of 100K+ rows by web scrapping job portals. Cleaned and tokenized job descriptions for **BERT** model.
- Built a taxonomy of 8000+ soft and tech skills. **IOB tagged** skills using pattern matching and regular expressions.
- Trained a distilbert-base-cased model from HuggingFace to identify skills from job descriptions achieving 98% accuracy.
- Cumulated 34 sets of required skills for different tech roles and built a recommender system based on similarity score.

Yelp Recommendation System – Python, PySpark, Collaborative Filtering — Mar 2023 – May 2023
- Spearheaded a team of 4 to develop a recommender system using Yelp customer reviews dataset and Spark.
- Cleaned and transformed 1M+ rows of data followed by feature engineering to implement **K-means and ALS algorithm**.
- Developed scalable framework to recommend 2 similar restaurants for each restaurant and 2 similar users to each user.
- Increased number of relevant recommendations by 60% through integration of a hybrid K-means and ALS model.

HealthCost Insight: Reducing Healthcare Cost – Rstudio, dyplr, ggplot — Sep 2022 – Dec 2022
- Led a 4-member team to pinpoint primary expense drivers, resulting in a 20% cost reduction for a Health Management Org.
- Performed extensive data cleaning, segmenting dataset at 75% cost quantile for precise binary classification.
- Implemented 3 ML models (**Linear Regression, Tree Bag, SVM**), boosting predictive accuracy by 15% for healthcare costs.
- Designed an interactive Shiny App Dashboard for 4 types of visualizations-Histograms, Scatterplot, Boxplot, Map Plots.

LEADERSHIP EXPERIENCE

IIT2024 Global Conference, Washington DC – Lead Volunteer Team — Jan 12-14, 2024
- Helped organize a team of 50+ volunteers to coordinate 1500+ attendees for a 3-day conference in Washington D.C.
- Managed LinkedIn and Instagram for the conference, growing social media outreach by over 50%.

QuantumCuse, Quantum Computing Club, Syracuse University – Director of Education — Jan 2023 – Dec 2023
- Initiated creation of 5 educational resources and 5 reusable modules for quantum computing beginners.