Siddhant Thakur

sidthakur08@tamu.edu | 206-480-8468 | LinkedIn:siddhant-thakur-data | Github:sidthakur08

EDUCATION

Texas A&M University, College Station
M.S. in Computer Science, GPA - 3.9 *2022 – 2024*

SRM University, Chennai
B.Tech. in Computer Science w/ Big Data Analytics, GPA - 3.87 *2018 – 2022*

WORK EXPERIENCE

Stealth Startup **Remote, United States**
Machine Learning Engineering Intern, Health *December 2024 – Present*

- Automated **caregiver filtering workflows** by designing custom decision logic integrated with WellSky filters, resulting in **25% reduction** in manual scheduling time
- Designed a caregiver profile database schema in **PostgreSQL** by analyzing key attributes, reducing multi-agency onboarding duration for **30+ caregivers**
- Currently implementing an open-source **Retrieval-Augmented Generation (RAG)** pipeline using **domain adapted Mistral 7B & FAISS** vector store to streamline caregiver inquiries and scheduling

Detroit Tigers **Lakeland, Florida**
Data Science Associate, Analytics *January 2024 – December 2024*

- Analyzed ground reaction forces' impact on fastball velocity using **GPBoost and SHAP** with pitchers as a random effect, achieving an **RMSE** of **1.22 mph**
- Quantified pitching acceleration shapes for **550 MLB pitchers** through **time series clustering** in R to investigate different kinematic sequencing patterns
- Engineered analytical workflows using **XGBoost** to **identify key biomechanical metrics** influencing in-game velocity, creating data-driven player plans for **50+ pitchers** across the minor league
- Automated **athlete jump profile generation** via an **R Markdown** pipeline for **180 Amateur players**, utilizing RM-MANOVA to assess variations across age & position groups

Airtel Telecom **Haryana, India**
Data Science Intern, Engineering *June 2021 – August 2021*

- Enhanced **customer satisfaction score** by **30%** through implementation of an ETL pipeline targeting address resolution within Airtel's support system
- Streamlined address extraction from OpenStreetMaps using **fuzzy matching** with **Levenshtein** similarity, expanding the address dictionary by **40000 residential entries** nationwide

Bajaj Health Insurance **Remote, India**
Data Science Intern, Finserv *March 2020 – August 2020*

- Engineered **full-stack recommendation system** providing health-related suggestions to **5000+ users**, leveraging patient assessments to tailor recommendations
- Modeled a **logistic regression** algorithm to estimate at-risk probability of Diabetes or Hypertension with recall score of **0.74**
- Deployed model in a serverless scaled environment on **AWS Lambda**, scaling to **10000 concurrent users** under Bajaj

PERSONAL PROJECTS

Natural Language Understanding *November 2024 – Present*

- Built a transformer-based sentiment analysis pipeline using **BERT-mini**, fine-tuned for **multi-domain sentiment** patterns to capture domain-specific contextual representations

Sports Analytics *August 2019 – Present*

- Implemented **computer vision** with **YOLOv8** to track player and ball movement for Arsenal's 2022/23 season, improving pass analysis through **Kernel Density Estimation**
- Developed a **mixed effect model** for Induced Vertical Break and Horizontal Break to estimate the impact of **dew point** on **pitch movement** by identifying unlikeliness of a residual from a prediction
- Analyzed **shot tendencies** of penalty takers in soccer using **Conditional GMMs** and **LightGBM** to predict shot direction

Group Recommender System *December 2022 – May 2023*

- Designed a group recommender system utilizing LinkedIn profiles and Slack bios for optimal **pairings of 3-4** students, leveraging **GloVe** embeddings, **TF-IDF** scores, and **Zero-Shot transformer** to calculate user-pair similarities
- Achieved **63% preference rate** for the model-generated optimal pairings over randomly selected pairings through independent observer evaluations

SKILLS & TOOLS

Skills: Statistics, Natural Language Processing, Computer Vision, Time Series Analysis, Deep Learning, Transformers, Generative AI, MLOps (CI/CD), Data Modeling & Schema Design, Mixed-Effect Modeling, A/B Testing

Tools: Python (Pandas, NumPy, Scikit-Learn, Flask), R (Shiny), SQL, PyTorch, TensorFlow, PySpark, Docker, Kubernetes, Git, AWS, Databricks, Snowflake, Tableau, Power BI