Siddhant Thakur

College Station, TX | https://www.linkedin.com/in/siddhant-thakur-08/ | 206-480-8468 | sidthakur08@tamu.edu

EDUCATION

Texas A&M University, College Station — *M.S. in Computer Science, GPA - 3.87, Graduation: May 2024*
SRM University, Chennai — *B.Tech. in Computer Science Engineering, GPA - 3.9, Graduation: May 2022*
University of Wisconsin, Madison — *Semester Abroad (Statistics), GPA - 4.0, Graduation: May 2021*

WORK EXPERIENCE

Group Recommender System — **College Station, TX**
Graduate Student, CSCE 670 — *December 2022 – May 2023*

- Designed a recommender system to determine **optimal groupings of 3-4** users based on user-pair similarities aggregated using **DESM** and **GloVe embeddings** over student bios from Slack and LinkedIn.
- Leveraged a **Zero-Shot Transformer** from **HuggingFace** to extract category weights, enabling accurate identification and classification of topics discussed in student bios.
- Implemented a comprehensive **survey-based** approach resulting in a **63% enhancement** in group **functionality**, showcasing exceptional analytical prowess, and effective problem-solving abilities.

Bharti Airtel Ltd. — **Haryana, India**
Data Science Intern, Engineering — *June 2021 – August 2021*

- Facilitated addition of **15+** different **streets** for **6** major **cities** in India by developing an **ETL pipeline** for address extraction through **OSMNx** and **OpenStreetMap** in Python.
- Extracted information from **40000 residential addresses** employing natural language models with the **deepparse** library.
- Optimized **average resolution time** across **2 states** for support tickets by deciphering highly unstructured addresses.
- Collaborated with the **Logistics** team, fostering seamless communication and coordination to enhance the feasibility of **network expansion** in **rural** areas

Big Data, Small Bugs — **Madison, WI**
Semester Abroad Student, STAT433 — *January 2021 – May 2021*

- Examined diversity over **263 aphid** species from **49 sites** across **14 years** throughout the Midwestern US landscape visualized using **Shiny in R** web app.
- Established association of aphid diversity count with cumulative **weather** and **crop monocultures** by implementing a **Poisson Regression** model built on environmental features engineered through **PCA**.

Bajaj — **Remote**
Data Science Intern, HealthRx — *March 2020 – August 2020*

- Proposed a proof-of-concept **recommendation engine API** providing suggestions to more than **5,000** users regarding their health, based on different parameters present in lab reports.
- Developed a **logistic regression** model in Python with a **recall score of 0.74** to estimate **at-risk** probability of a patient with Diabetes or Hypertension, while identifying important features.
- Orchestrated serverless deployment of the classification model to **10,000** users on **AWS Lambda** for sustainable scaling.

PROJECT EXPERIENCE

NFL/MLB Analytics — *December 2022 - Present*

- Devised a **random forest** algorithm with **63%** accuracy while predicting money line for the **2019 NFL** regular season and published articles on **Medium** summarizing weekly game winners.
- Implemented the idea of catcher framing by utilizing **mixed effect modeling**, incorporating pitching data from the PITCHf/x and Statcast tools through **pitchRx and baseballR** packages.

March Machine Learning Mania - Kaggle — *March 2022 - Present*

- Attained **2nd position among 50** participants in the ESPN March Madness Bracketology using **XGBoost** achieving a brier score of **0.26**, leveraging key statistics like **FG%**, **ELO** and **+/-** score.
- Improved the 2022 model by incorporating features like **Rebound Differential** and **Assist to Turnover ratio** into an ensemble of models fine-tuned via **GridSearchCV** resulting in a **20%** improvement in brier score.

Soccer Analytics — *August 2020 - Present*

- Utilized **YOLOv5** to identify the players in a football match and in process of applying **Kernel Density Estimation** to this tracking data to analyze the team’s **ball distribution** over the whole game.
- Built an **end-to-end pipeline** consisting of an **analytical panel** to help Fantasy Premier League managers to make informed decisions and choose from **560** premier league **players**.

SKILLS & TOOLS

Skills - Statistics, Machine Learning (Supervised & Unsupervised Learning), Computer Vision, Natural Language Processing, Data Engineering, Time Series Analysis, Deep Learning, Dimensionality Reduction, Feature Engineering, Data Visualization
Tools - SQL, Python (Pandas, NumPy, Scikit-Learn, Flask, Plotly, Dash), R (Shiny, CropScapeR), Tableau, TensorFlow, Pytorch, C++, GIT, AWS, Rest APIs, Ruby, VS Code, Microsoft Excel