LAKSH KATARIA

Jersey City, NJ | (908) 340-8342 | lvk8525@nyu.edu | Github | LinkedIn

EDUCATION

New York University - Tandon School of Engineering, New York, NY **May 2023**
Master of Science, Computer Science GPA: 3.9/4.0
Relevant Coursework: Design and Analysis of Algorithms, Machine Learning, Big Data, Deep Learning

Thadomal Shahani Engineering College, Mumbai, India **Jun 2021**
Bachelor of Engineering, Computer Engineering CGPA**:** 8.81/10
Relevant Coursework: Machine Learning, Database Management Systems, Natural Language Processing

SKILLS

- **Programming Languages** - Python, R, Tableau, SQL, JavaScript, HTML5/CSS, C/C++, Haskell, Scheme, Julia
- **Frameworks/Tools** - Docker, Git, Azure, Selenium, Scikit-learn, PySpark, NLTK, PyTorch, TensorFlow, Keras, OpenCV, Pandas, NumPy, BeautifulSoup, Node.js, Hive, Sqoop, VueJs, D3Js, ElectronJs
- **Databases** - MySQL, MongoDB, Firebase

EXPERIENCE

Data Science Intern, *Analysis Group*, Boston, USA **Feb 2023 – Apr 2023**
- Collaborated with a team of 5 members to develop a healthcare analytics application using Vue 3, Vite, and the composition API for project setup, Vuetify for template generation, D3.js and Pinia stores for data and event handling, and developed features such as modals, webpage to pdf generation functionalities amongst others.
- Used ElectronJs to create a desktop executable for the application, and Electronegativity plugin for testing. Used Git and Jira to help manage the development lifecycle.
- Created an online dashboard for a Haitian clinic to visualize and streamline patient data from 2019 to 2022 using Python Dash. Used dash bootstrap components to create visualizations and data tables to display raw data with carefully crafted callback functions to identify and represent errors directly on user side
- Assisted analysts with Python Dash usage, acclimatization with Git, and docker environment setups
- Executed web scraping projects to collect unstructured data and performed extensive pattern and string matching amongst other analyses to create deliverables

Office Assistant*, NYU Tandon School of Engineering***,** New York City, USA **Oct 2021 – Dec 2022**
- Executed various administrative tasks including managing department databases, files, and inventory, in addition to client-side responsibilities such as answering emails and resolving inquiries. Coordinated student employment, adjunct faculty filing

Data Science Intern, *Analysis Group*, Boston, USA **Jun 2022 – Aug 2022**
- Structured and maintained two data engineering pipelines, to ingest over 11 million data points from Reddit and 40k profiles from LinkedIn, leveraging BS4, Selenium, and API manipulation. Parallelized data parsing over an HPC system through bash commands, improving run time by over 100%, and implemented version control using Git
- Leveraged nltk library to implement keyword matching with inflected words, and generated word counts, in text extracted from 578 PDF files leveraging Azure's Computer Vision API
- Developed a web dashboard operating Python Dash, to visualize the company's HPC resource usage and allocation, optimizing allocation by 5% by identifying unnecessary ongoing jobs

PROJECTS / ACADEMIC RESEARCH

Text Summarization - RNN vs Transformers **Oct 2022 - Dec 2022**
- Conducted text summarization on news articles operating Recurrent Neural Networks with LSTM logic, and T5 Transformer model, and compared performances using ROGUE scores
- Devised an LSTM Neural Network using Tensorflow Keras library, trained model on over 75k data points, and validated results on over 10k data points
- Performed hyperparameter tuning and produced an average ROGUE-1 recall score of 0.26, comparable to pre-trained Transformer score of 0.41

SUD Treatment Result Predictor **Oct 2022 - Dec 2022**
- Performed downsampling and data cleaning to deal with imbalanced classes, along with extensive data analysis to identify most lethal drugs, at-risk ethnic and age groups, and American states with highest drug-related death rates
- Pipelined predictive models such as XGB, RandomForrest, LogReg, and Decision Trees to anticipate the outcome of a treatment regime. Performed hyperparameter tuning using GridSearchCV and achieved a precision score of 88% and 87% accuracy using RF model

YouTube Video Analysis **Mar 2022 – May 2022**
- Collaborated in a team of 3 to conduct analysis using PySpark, on 835Mb YouTube Video data from 10 countries
- Performed EDA, AWS - Rekognition, and MinHashLSH with TF-IDF and Jaccard Distance based text matching to identify valuable metrics such as optimal viewing and posting times to maximize viewership and bring videos onto YouTube's trending page

Machine Learning in Statistical Arbitrage (Research) **Feb 2022 – May 2022**
- Conducted a literature review on statistical arbitrage to understand co-integration in stock prices
- Carried out statistical tests such as augmented dickey-fuller on data from 442 stocks (S&P 500) to identify co-integration; performed linear regression, and PCA, building a covariance matrix for further application. Studied the DDPG algorithm

Crop Prediction using ML **Mar 2020 – May 2020**
- Constructed a Machine Learning model based on RandomForestClassifier, KNN, and Decision Trees to predict which crop to grow in a specific geographical area in agricultural India
- Deployed a Multi-Variable Linear Regression algorithm to predict profit generated for a particular crop taking the crop and space in hectares as independent variables and performing feature transformations