Vivek Gupta

vivekgup@cs.cmu.edu | (412) 626-8773 | www.linkedin.com/in/vivekgup96

SUMMARY

A top-ranked CMU Masters candidate in **Data Science** with a strong **Computer Science** background & deep expertise in **enterprise level application** of **machine learning techniques** through internships and projects. Highly skilled in **algorithms, databases, cloud tech, ML, neural-networks & python** with a stellar academic record. And a **published computational linguistics** (NLP) researcher.

EDUCATION

Carnegie Mellon University - School of Computer Science — Pittsburgh, PA
Master of Computational Data Science | **GPA: 3.92** — Dec. '20
Coursework: Deep Learning, Distributed Systems, ML for Large Datasets, Cloud Computing, Machine Learning, Interactive Data Science, Deep Reinforcement Learning*, Probabilistic Graphical Models*, Capstone Project* *(*ongoing)*

Indian Institute of Technology - Kharagpur — Kharagpur, India
Bachelor of Technology in Computer Science and Engineering | **GPA: 9.58 / 10.0** (Dept Rank - **4/105**) — Aug. '19
Selected Coursework: Algorithms, ML, NLP, Data Analytics, Data Mining, Database Mgmt. Sys, Probability & Statistics, SW Engrng
Certificate of Excellence in Academics by the Department of Computer Science and Engineering, IIT Kharagpur

EXPERIENCE

SDE Intern (remote) | **Adobe Inc.** — Pittsburgh, PA| May '20 - Aug. '20
- Implemented a RESTful microservice with Azure cosmos db backend to act as a registry for storing, updating & retrieving 'metric-definition' documents with robust exception handling & validation. **Deployed the service to the production server**
- Metric definition standardizes the way platform services store and query monitoring metrics. The registry service will allow controlled ingestion reducing metric volume saving resource cost; enforce consistency in metric properties avoiding high cardinality time series; constrain request parameters when querying metrics via API preventing unauthorized operations.

Summer Analyst | **Goldman Sachs Pvt. Ltd.** — Bengaluru, India | May '18 - Jul. '18
- Expedited the compilation process of proprietary modeling language by identifying optimization points and pruning the compilation graph; improved compilation-time by appx. 58%. Developed a testing framework to verify the modified compiler

ACADEMIC PROJECTS & ASSIGNMENTS

Exaggeration Detection in Health Research (Thesis Project) — IIT Kharagpur | Jan '18 - Apr. '19
- Performed temporal analysis of exaggerated content dissemination in a social network (Twitter)
- Classified Twitter users according to their tendency to spread exaggerated news using profile & tweet based features
- Compared ML algorithms (SVC, XGBoost, Random Forest) to classify users & obtained a final F1-score of 0.83
- Published **"Characterizing the spread of exaggerated health-news content over social media"** in **ACM HT '19** [paper]

End-to-end 2D to 3D Video Conversion [video] — CMU | Spring '20
- Trained an end-to-end deep CNN to predict the right image of a stereo image pair given the left image as input
- Infused input left image with depth and segmentation information to reduce the pixel-wise mean absolute error to 6.84

Predicting Local Business Popularity during Olympics using Geospatial features [datafolio] — Datathon | July '20
- Identified factors affecting popularity of local vendors during Olympics: station connectivity, diversity of neighbourhood, etc.
- Predicted change in popularity (measured by #check-ins) of businesses using ML models achieving AUC score of 0.69 (RF)

NYC Taxi Fare Prediction [link] — CMU | Spring '20
- Trained distributed ML models (linear regression, decision tree, RF, GBT) on large-scale NYC-taxi data (~95 GB) to predict cab fare using engineered features like trip distance, duration, #passenger, time etc. Used Pyspark to clean data on the cloud.
- Evaluated and compared the ML models along 4 dimensions - RMSE, model size, training time, inference latency

Yelp Data Insights - Interactive Analysis of Restaurant Data [website] — CMU | Fall '19
- Created a website to interactively analyze Yelp restaurant data along different data dimensions using Tableau visualizations

Iterative Graph Processing with Spark — CMU | Fall '19
- Analyzed a large Twitter social graph with Apache Spark using PageRank algorithm to rank users by their influence
- Configured Spark application to optimally utilize Azure HDInsight cluster resources to meet the performance objective

Twitter Analytics Cloud-hosted Web Service — CMU | Fall '19
- Designed, deployed & optimized functional web-servers that can handle a high load reliably under a constrained budget
- Implemented Extract, Transform and Load (ETL) on a large data set (~ 1 TB) to load clean data into MySQL database
- Designed schema as well as configured and optimized MySQL database to deal with scale and improve throughput

TEACHING & RESEARCH ASSISTANTSHIP

- **Research Assistant at Language Technologies Institute, CMU** — CMU | Spring '20
  - Member of the development team of "Foundations of Computational Data Science" course offered by LTI, CMU
  - Developed auto-gradable assignment on solving a movie recommendation problem using collaborative filtering
- **Teaching Assistant for Machine Learning for Large Datasets (10-605)** — CMU | Fall '20
  - Hold recitations and office hours; Develop and update course projects; Evaluate assignments [100+ enrolment]

SKILLS

Languages: Python, Java, C++, JS, Scala | **Technologies**: SQL, Mongo, AWS, Pytorch, Pandas, ScikitLearn, Spark, Tflow, Docker