Vaibhav Jain

✉ vaibhav.jain174@gmail.com   linkedin.com/in/vaibhav174   github.com/vaibhav174

I am a M.Sc. Data Engineering and Analytics(major in NLP) graduate from Technical University of Munich with about 2 years of hands on experience working in the field of data science. I love bringing data to life through analytics and visualizations. I want to further expand my practical data science and engineering knowledge by working on cutting edge research and real-world problems.

EXPERIENCE

Siemens — Apr 2022 – Present

Data Scientist — *Munich, Germany*

- Developed, deployed and maintained Machine Learning and NLP based research prototypes that solve business problems for other internal units. (Python, ML, NLP, EC2)
- Performed large-scale Batch data pre-processing and ETL on data from various sources.(MySQL, S3, Athena)
- Developed and Maintained a python library which holds the functionalities to domain adapt pre-trained neural models to downstream task domains(Siemens industrial/technical domains) using Supervised and Unsupervised Learning techniques. (Python, Git, Linux, CI/CD)
- Developed a new metric to measure the domain divergence of Language models reducing the running time of the library functions by 56%.(Semantic Analysis, Hypothesis testing)
- Cloud deployment of domain adapted models for ease of access by other teams within Siemens and presented result dashboards to multiple global heads.(AWS, Flask, Tableau)
- Got 8 team lead approvals for the library based on the improvements in AI products resulting from library usage.(Stakeholder collaboration, Presentation)

Horváth-Steering Labs — Apr 2022 – Sept 2022

Application Project — *Munich, Germany*

- A 6 months project jointly supervised by TUM DI lab and Steering Labs at Horváth.
- Utilizing Self supervised learning approach to build an end-to-end pipeline to generate Knowledge graphs from PDF's taken from German Environment Agency (Umweltbundesamt)
- Project details and report: LINK

Volunteering — Jul 2021 – Present

Data Science Tutor — *Munich, Germany*

- Volunteering to teach Python and basic Machine learning free of cost to students living in Studentenstadt Munich (Largest student dormitory in Munich)

EDUCATION

Technical University of Munich — Oct 2020 – Jul 2023

M.Sc Data Engineering and Analytics — *Munich, Germany*

Birla Institute of Technology, Mesra — Aug 2016 – Jul 2020

B.E Computer Science — *Ranchi, Jharkhand*

Ryan International School — Apr 2015 – May 2016

Senior Secondary School(High school) — *Jaipur, Rajasthan*

TECHNICAL SKILLS

Programming Languages	Python, C, SQL, R
Data Science Frameworks	Pandas, NumPy, Keras, Tensorflow, Pytorch, HuggingFace, SpaCy, Scikit-Learn, Flask
Analytics tools	Tableau, Excel, Tensorboard
Version Control	Git(Proficient)
AWS Services	S3, EC2, Athena

PROJECTS (https://vaibhav174.github.io/)

Effect of Vocabulary overlap and dataset size on domain adaptation of LLM's Master's Thesis

- Worked on mitigating the effect of Out of Vocabulary words and low resource setting condition on domain adaptation of LLM'S
- Goal is to build the most effective domain adaptation pipeline based on the downstream dataset properties.
- Developed a new metric to measure the divergence between two text datasets.
- Developed a new metric which is independent of type of downstream task to measure the impact of domain adaptation(Under review for Siemens Patent).

Predicting optimal Health insurance premium price using Machine Learning(Regression).

- **Task:-** Create and test different prediction models that predicts the optimal premium price of health insurance policy for the customers.
- Data Exploration and Feature Analysis to perform feature selection and feature engineering using ColumnTransformer from Scikit-Learn.
- Buld a pipeline that test multiple regression models including linear Models, Tree-based models and Boosted tree using Light Boost and XG-Boost using scikit learn and pick the best based on R2 score and adjusted R2 Score.
- Grid search for hyper parameter tuning integrated with the model pipeline using GridSearchCV from Scikit-Learn.
- Best trained model achieved R2 score of 96,3% and ajusted R2 score of 94.5%
- Deployment and prediction using flask web application.

Movies and Songs Recommendation system

- **Task:-** Implement different recommendation algorithms for songs and movies recommendation.
- popularity based recommendation.
- Content based filtering using movies genres for movies recommendation.
- User and Item based collaborative filtering using movie genres and co-occurrence matrix.
- Pearson's correlation to find similar users and items.

MCQ AND T/F Question generation using Transformers

- **Task:-** Building an end to end pipeline that help primary school teacher in generating the MCQ and T/F questions from the given correct answer and paragraph from which correct answer was extracted.
- Fine tune the T5 transformer model using SQuAd Data set to generate questions from one word answer,context(word sense) and text from which question to be generate from.
- Adapt BERT to perform Word sense disambiguation using using positive negative context-gloss pair.
- Generate wrong choices for MCQ using co-hypernyms of the correct answer in the WORDNET.
- Generate False statements for T/F type questions by removing ending verb phrase or noun phrase from the sentence and completing the sentence using by wrong verb/noun phrases generated using GPT3.

PUBLICATION

Image and Video colorization system **Jan 2020**

International Journal for Research in Applied Science and Engineering Technology (IJRASET)

PAPER LINK

PARTICIPATION/VOLUNTEERING

- Participated in Rajasthan Hackathon 5.0 aimed to developed a strategic solution to help small shopkeepers increase business through IT solutions.
- Practical training to manage medical and technical portals by Government of Rajasthan.
- Achieved All India Rank 674 in Graduate Aptitude Test in Engineering(GATE).
- Tutor at studentenstadt Munich House 3, responsible for handling public events for the house.
- Volunteered to assist for StuStaCulum 2023 (festival organised by students).