ANURAG SANGEM

+1 (551) 220-7098 || ansangem@iu.edu || LinkedIn || Website || GitHub

EDUCATION

Indiana University, Bloomington (Aug 2022 – Jul 2024) Bloomington, US.

Master of Science in Data Science.

Coursework: Data/Social Media Mining, Deep Learning Systems, Statistics and Random Variables, Elements of Artificial Intelligence, Machine Learning, Cloud Computing, Data/Scientific Visualization, Big Data Applications, and Intelligent Systems.

Vellore Institute of Technology (Jul 2016 – Jun 2020) Chennai, India.

Bachelor of Technology in Electronics and Communication Engineering.

WORK EXPERIENCE

Research Data Scientist (Aug 2023 – Dec 2023) Bloomington, US.

Department of Medical and Molecular Genetics - Indiana University

- Proficiently developed R Scripts and executed computational pipelines and workflows for processing a range of biological data types, with a specialization in Nextflow-based pipelines for processing NGS data, including single-cell and bulk RNA-seq. Ensured the efficient and accurate analysis of genomics, transcriptomics, proteomics, and metabolomics datasets.
- Successfully managed the acquisition and utilization of high-quality single-cell RNA sequencing (scRNA-seq) and omics data for Alzheimer's Disease (AD), resulting in the integration of these datasets into approximately 10 Central Nervous System (CNS) papers. Oversaw the compilation and verification of data lists to efficiently downloading and processing of the large datasets.
- Orchestrated the identification and preparation of data for this initiative. Successfully completed similar data preparation tasks for liver diseases, impacting a total of five research tasks. The resulting datasets contributed significantly to ongoing research endeavors.

Graduate Assistant Instructor (Jan 2023 – Jun 2023) Bloomington, US.

Luddy School of Informatics - Indiana University

- Mentored over 60+ students in a graduate-level course, Data Mining (B565), providing guidance and clarification on Machine Learning and Data Mining concepts through doubt-clearing and code review sessions.
- Developed and implemented compelling R and Python scripts focusing on key topics such as Principal Component Analysis (PCA), Classification, Regression, Dimensionality Reduction, K-Nearest Neighbors (KNN), etc. on real-time data sets. These scripts enhanced understanding and facilitated the practical application of complex concepts, elevating the learning experience for students.
- Recognized with the 'Informatics Excellence as an Assistant Instructor' award by the department for exceptional dedication and contributions to the student's learning journey.

Data Scientist (Aug 2020 – Jul 2022) Bangalore, India.

Oracle

- Focused primarily on building propensity to buy models using algorithms like Random Forest, SVM, etc.
- Modeled and scored more than 40 Oracle-owned products like ACONEX, NETSUITE, etc. on the historical data and updated the ‘potential customers to target’ to the sales teams this resulted in a 1.8x lift in opportunity win rate for top-ranked accounts.
- Utilized multi-channel attributions along with an Attention-based RNN and a fully connected neural network, which aims to find out the right marketing channels that ultimately lead to a sale.
- Designed a dynamic attribution framework using the Markov chain algorithm to measure the impact of the marketing campaign on pipeline creation.
- Developed three ranking algorithms namely Sequential Ranking, Weighted Ranking, and Ranking using a generalized linear model (GLM) on the archival data of the Tech Cloud products and presented the top 50 potential customers in each region of the LAD (LAD Territory Account Ranking).
- Leveraged the inbuilt functions of OML4py to improve the efficiency of Python scripts by more than 35%
- Performed Hypothesis Testing for Fusion ERP, EPM, SCM & HCM: Impact of Marketing Touches on Win Rate and Average Won Pipe occurring at different times in the B2B sales funnel.
- Performed statistical analysis on the Oracle CX Sales Data which is being advanced to a new tool called Datafox (acquired by Oracle).
- Accelerated the CXD to Datafox matches by 17% by using Jaro-Winkler similarity.

Data Analyst Intern (Jun,2019 – Jul,2019) Chennai, India.

Appyhub Technology Solutions

- Created illustrative dashboards using Tableau, SQL, MS Excel, and Power Query based on the requests from the BA teams, saving approximately 8 hours of manual reporting work per week.
- Interpreted and visualized data from 6 business campaigns, and user responses and provided weekly reports with insights and outcomes.
- Conducted market research for sales and procurement data which resulted in a 2% increase in the contracts in 1 month.

PERSONAL PROJECTS

Covid-19 Vaccine - Tweets Sentiment Analysis (Deployed in AWS)

Kaggle, Tweepy, NLP, TF-IDF Vectorizer, Supervised model, MLFlow, Docker,Design, AWS (Sage Maker, S3 bucket, BOTO3).

- Worked on the Restaurant Reviews data set from Kaggle to train a supervised model with sentences of positive and negative emotions, used several NLP techniques/libraries such as stop words, PorterStemmer, and TfidfVectorizer for data processing.
- Leveraged the open-source MLFlow platform to keep track of my model performance on various parameters and had all the instances logged to mlruns.
- Pushed the docker container image of the model and loaded it to the S3 bucket in AWS, using the mlflow. sage maker module to provide an API to deploy my MLFlow models.
- Using BOTO3 as the AWS SDK the endpoint of my model in Amazon Sage Maker is invoked to make predictions on the recent tweets related to 'Covid-19 Vaccine' extracted using Tweepy.

IU Grad School Admission Prediction and University Recommendation System Web application.

Beautiful Soup, Scrapy, Weighted KNN, Random Forest, HTML, JavaScript, Flask, Model Deployment, AWS, Heroku, GIT.

- Collected and analyzed admission decision data from multiple sources, including web scraping using beautiful soup and Scrapy, Google forms, and pre-existing datasets on Kaggle to develop a university recommendation system, incorporating factors such as program, university, work experience, research publications, GRE scores, TOEFL scores, and US NEWS Rankings.
- Calculated the Grad admission probability at IU by utilizing the Random Forest algorithm and evaluating metrics such as precision and F1 score.
- Implemented a University Recommendation System based on a weighted KNN algorithm to provide personalized recommendations for applicants.
- Developed a feature-weighting system based on major requirements and utilized the KNN algorithm to identify similar applicants. Implemented an algorithm, 'filter50plus,' to filter the top 50 universities with the highest acceptance rates in each cluster, achieving an accuracy of 68.83% for MIS and 54.57% for computer science.
- Deployed a web application on Heroku connecting it to a Git repository, configuring environment variables, and utilizing a CI/CD pipeline. Ensured correct deployment and scalability of the app on Heroku's cloud platform.

Analysis of emotion/sentiment in Taylor Swift's lyrics in her Discography (NLP)

Pandas, Data Cleaning, Data analysis, NLP, VADER, Data Visualization, NLTK, Sentiment Analysis, data driven.

- Gathered all the song lyrics of Taylor Swift's new album (my personal favorite) Midnights, and all her album lyrics since 2006 (Taylor Swift, Speak Now, Red, 1989, etc.).
- Removed the stop words and tokenized the data to prepare it for analysis. Visualized the most frequent words she used in each album.
- Used VADER (Valence Aware Dictionary and sentiment Reasoner) from the NLTK Python Library to label the words in the lyrics into these 3 categories Negative, Neutral, and Positive using the 'polarity_scores' method available in the library.
- Overall, Taylor Swift's albums returned a 'positive' sentiment, and in contrast to what any 'Swiftie' would expect the album 'Midnights' is not the most negative/sad album in her discography!

PROFESSIONAL CERTIFICATION AND COURSES

Oracle Machine Learning using Autonomous Database 2021 Certified Specialist

Oracle Machine Learning, Autonomous Database, SQL PL/SQL, OML4PY, OML4SQL, Scheduling Jobs, Predictive Analytics, Data Mining, Model Building, Feature Selection, Model Evaluation, Clustering, Classification, Regression.

Stanford NLP By Prof.Chris Manning (CS224N-Spring'19)

Word2Vec, N-Gram, Attention, Transformers, BERT, GloVE, Recurrent Neural Networks and Language Models, Hugging Face Transformers, Large Language Models, Generative Adversarial Networks (GAN).

SKILLS

Programming Languages: Python, C++, R, PL/SQL.

Developer Tools: Oracle machine learning notebooks, Docker, MLflow, Kuberflow, GCP, Kubernetes, AWS Sage maker, Tableau, GIT version control, Heroku, PowerBI, Excel, Oracle Sql developer, Linear programming.

Data science /Model Deployment skills: Statistical modeling, Hypothesis Testing, EDA, Predictive Modelling, Machine Learning, Decision Trees, XGBoost, LightGBM, optimization, Pandas, Matplotlib, NumPy, Weights & Biases, Deep Learning, TensorFlow, Pytorch, Hugging Face, Natural Language Processing, Neural networks, Topic Modelling, Sentiment Analysis, dplyr, scikit-learn, a/b testing, Autoencoders, Airflow, Flask, YAML, Boto3, and Prometheus, Statistical Analysis System (SAS), Data modeling.

Big Data: Map Reduce, Hadoop, Apache Spark, Data Warehousing, Data Visualization, ETL (Extract, Transform, Load), NoSQL, Postgres, Distributed Computing.