Varsha Vattikonda

vv2116@nyu.edu | +1-805-743-9755 |Linkedin

Education

M.S. in Data Science, New York University, USA Sep 2021 - *May 2023
Coursework : Natural Language Processing, Deep Learning, Big Data (Spark), Probability and Statistics, Linear Algebra

BTech & MTech in Electrical Engineering, Indian Institute Of Technology Madras (IITM) Aug 2011 - May 2016

Skills

Programming Languages	Python, SQL, C# and R
Data Science	Statistical Models, Machine Learning Models, BI Models, Identify KPIs
Software Development	ETL , Object Oriented Design, Data Structures and Algorithms, Git
Cloud Services	Databricks, Azure (Data Factory, Analysis Services), GCP (Vertex AI, Earth Engine, BigQuery)
Reporting	Power BI, Splunk and Salesforce

Work Experience

Data Science Intern | MSCI Inc. ESG NYC, USA May 2022 - Aug 2022
Developed a PoC evaluating the capabilities of **Google Cloud Platform** and **Google's Earth Engine** in Spatial Finance that has a potential of **$1M** in annual recurring revenue

- Delivered an asset level flood forecast model using **KMeans Clustering** and **Random Forest**
- Deployed and automated the model to update daily flood probability in BigQuery using **Vertex AI**

Designed an enterprise architecture facilitating **Data Catalog** for around **10GB** data flowing in daily from around 2000 vendors

- Delivered a prototype for pluggable and extensible pipeline for data quality assurance using **Great Expectations** and **Airflow**
- Implemented a custom integration that pushes the metadata, lineage and data quality metrics of Delta tables to **Datahub** using **Spark** and **Databricks**

Data Science and Machine Learning Engineer |MSCI Inc., Mumbai, India Jun 2016 - July 2021
Multifaceted role in building a centralized Data Mart and pipelines for mining actionable insights from client's usage data. This project has earned an additional **$1.2M** of annual recurring revenue

- Achieved MAE of **5mins** wrangling around 10M rows of data to predict the duration of a simulation on the clients' portfolio using **K-Means Clustering and Linear Regression**
- Built a **Unigram (NLP)** model on 10K email conversations between the clients and consultants assessing the sentiments using **SentiWordNet**
- Built a **Churn Prediction** model using 10-15 features like magnitude of spikes, dips and trends of Client Engagement KPIs for 4 different product lines with around 5000 clients each using **Ensembled decision tree algorithms**
- Reduced churn rate by **10%** using **Statistical Data Modeling** and categorizing the users as DAU/WAU/MAU and thereby underlining the dips in usage
- **Collaborated** with Product teams to identify KPIs for products' usage and built a **Product Recommendation Model** underlining the potential cross-sells using **PCA** followed by K-Means Clustering
- Designed **ETL and QA Framework (C# API)** that extracts around **10GB** data each day from **10** different sources (eg. Splunk, Salesforce) and runs quality checks against self-learned **dynamic thresholds** using **Azure cloud services**
- Increased Client Insights Reporting platform adoption by **40%** by creating an intuitive and actionable **PowerBI** reporting solution in addition to firm-level trainings on the same

Projects

- **Precipitation Nowcasting** - A deep learning model based on **UNet (Pytorch)** using Negative Gaussian log-likelihood as the loss function with the model generating two output images - Mean and Standard deviation
- **Movie Recommendation Engine**- Comparative study on the performances of Movie Recommendation Engine using **Latent Factorisation** model (Alternating Least Squares for optimization) using **SparkML** (distributed system) vs Lenskit(single machine) for different data sizes
- A study on **The Evolving Sentiment of Work From Home** based on 3M samples of daily tweets over the last 3 years using **Naive Bayes Topic Model** to understand the topics of concern related to WFH and how the popularity and polarity of the topics have changed over time