SATYAM MISHRA

Boston, MA | P: +1 857-701-3819 | mishra.sat@northeastern.edu

EDUCATION

Northeastern University – Boston, USA *August 2020 - July 2022*
Master of Science, Data Analytics Engineering
Relevant Coursework: Machine Learning, Neural Networks, Data Mining, Data Management, Statistics, Probability

Medicaps University – Indore, India *August 2016 - July 2020*
B.Tech, Computer Science Engineering
Relevant Coursework: Python Programming, NLP, Data Visualisation, C Programming, Discrete Mathematics

TECHNICAL SKILLS

Programming: Python (Numpy, Pandas, Tensorflow, Matplotlib, OpenCV, Scikit-learn, PySpark, Flask), R
Tools/Frameworks: Docker, AWS, Kubernetes, Tableau, Alteryx, Spyder, Jupyter, Git, BERT, VGG-16, InceptionV3
Databases and Tools: S3, MySQL, Postgres, SSMS, MongoDB, Snowflake, Apache Cassandra, Redshift, Airflow
Analytics: Flourish, ETL, Hypothesis Testing, A/B Testing, Market Basket Analysis

PROFESSIONAL EXPERIENCE

Unum Group - Portland**,** ME *July 2021 – December 2021*
Data Science and Analytics Co-op

- Automated an **ETL** pipeline to stage data from **S3 buckets** on **Redshift**, transform, and divert to **Tableau** for **ad-hoc analysis** of mental health status of users as well as strategic plan for marketing team
- Assisted in writing python **DAG** to schedule ETL operations on data pipelines every 12 hours using **Airflow**
- Deployed **Random Forest** and **Naive Bayes classifiers** to minimise false negative users with **AUC of 0.8**
- Programmed PoC NLP classifier on **RoBERTa** on 56000 journal entries to identify users with critical symptoms

Eletech Engineering - Indore, India *January 2020 – June 2020*
Data Science Intern

- Performed **data quality checks** on operations data using pandas followed by **EDA** and statistical analysis
- Used **PyTesseract** to create an OCR model to digitise database, migrated it to **S3** buckets for optimal **ETL**
- Implemented dimensionality reduction with (**PCA**, **Lasso**) on impact metrics data with **55 features** to build **K-Means** clustering model to segregate faulty manufactured components
- Analysed feedback of **1Mn** customers, deployed **BERT base** classifier to assist marketing team campaigns

Aam Aadmi Party - Delhi, India *July 2019 – September 2019*
Data Analysis and Machine Learning Intern

- Coded **Beautifulsoup** and **Selenium** script to extract data of **70** assemblies from Election Commission's website
- Worked with a team of 2 developers to assist with **data warehouse** for querying data, thus eliminating work of a DBA. The playground ran on **Docker**, **Flask** and **SQLAlchemy** on local server
- Created machine learning models (**XGBoost** and **Ridge Regression**) on electoral data of past 15 years and 4 elections which resulted in **84%** and **76%** accuracy which classified wins and number of favoured votes

ACADEMIC PROJECTS

Caption Generation With CNN-GRU Encoder-Decoder [Tensorflow | Docker | MS-COCO]

- Applied VAE and Convolutional VAE architecture with tensorflow 2 on **50000** images
- Implemented the architecture on Google Colab with **InceptionV3** encoder and **CNN-GRU** decoder networks
- Evaluated performance against current SOTA, with a **BLEU** score of 32.5 on the best model
- Containerized the model with **Docker** and further hosted it on AWS

Music Platform Data Pipelining and EDA [S3 | BOTO3 | Psycopg2 | Redshift | PySpark SQL]

- Migrated JSON data from local directory **AWS S3** bucket with **BOTO3** and **psycopg2**
- Staged the data from S3 to **AWS Redshift** with SQL queries written in python script via **configparser**
- Integrated data pipeline from **Redshift to Tableau** after engineering features with **PySpark SQL**

Anomaly Detection On Time Series Data With SVM and DBSCAN [Tensorflow | Flask | EC2 | Docker]

- Employed SVC and DBSCAN for detecting the anomalies on a **16000** cases and 4 predictor variables
- Performed **Radial Based Oversampling** and Synthetic Minority Oversampling (**SMOTE**) for class imbalance
- Engineered a feature (**weighted absolute difference**) to minimise false negatives
- Achieved an **ROC** score of **0.88** and 0.71 and an **F-1** score of **0.23** and 0.14 on both models