SAI SRUJAN KUMAR MAJETI

(+1) 857-891-6994 | majeti.sa@husky.neu.edu | LinkedIn: srujanmajeti | GitHub: srujanmajeti

Education

Master of Science in Analytics (Specializing in Statistical Modeling) *(Exp.) May 2020*

Northeastern University, Boston

- Large-Scale Parallel Data Processing, Data Management and Big data, Machine Learning, Probability and Statistics

Bachelor of Technology in Electronics and Communication Engineering *Jul 2012 - May 2016*

Vellore Institute of Technology, India

- Object Oriented Design, Algorithms & Data Structures, Computer Organization and Architecture, Database Management

Professional Experience

Graduate Teaching Assistant *Jan 2020 - Present*

- Worked as a TA for the courses 'Probability and Statistics' and 'Intermediate Analytics'
- Helped graduate students in lab sessions that emphasize hands-on machine learning algorithms and problem solving with R

Data Analyst Intern, All the Way Live Analytics *Sep 2019 - Dec 2019*

- Architected models to predict difference of home and away team scores in NBA by performing complex feature engineering
- Developed an API to scrape data in JSON format from different webpages and created pre-processing scripts to clean the data
- Hosted the model in AWS using EC2 and Kinesis by creating a streaming service to ingest the data for stream processing

Big Data Analyst, Accenture *Aug 2016 - Jul 2018*

Clinical Ingestion

- Identified data requirements and performed data mapping with the target table format in Teradata
- Applied transformations on Hive tables by performing lookups on views that are fetched from Teradata using SFTP
- Migrated 2 TB of customers claim data from AWS S3 bucket and performed Spark batch processing
- Performed Sqoop Export from Hive to Teradata, and integrated analytic datasets used for modeling

ETL Pipeline

- Extracted raw data from different source systems (Teradata, SQL Server, and Oracle DB) to HDFS using Sqoop
- Responsible for creating Hive external and internal tables using partitioning and bucketing techniques with ORC format
- Developed shell scripts to automate creation of Hive DDL statements that are compatible with the source system
- Implemented CDC and scheduled cron jobs to capture changing data once a day for multiple systems

Skills

Languages	Python, Java, R, SQL, NoSQL
Big Data	Hadoop, Spark, Sqoop, Kafka, Hive, AWS
Data Libraries	Spark MLlib, Spark SQL, Pandas, NumPy, SciPy, scikit-learn, Beautiful Soup, Keras
Data Visualization	Tableau, matplotlib, seaborn
Tools	Git, Maven, Jupyter Notebook

Projects

Credit Card Fraud Detection - (Apache Spark, Apache Kafka, AWS, Python)

- Established an Ingestion platform on AWS EMR through Kafka to stream data of 2,84,808 records in real-time
- Addressed imbalanced data; detected anomalies; performed cross-validation and feature selection
- Implemented fraud detection models using Random Forest and Gradient Boosting Trees Classifier
- Predicted the class with an accuracy of 98.82% using scikit-learn and Spark MLlib

Million Songs Data Warehouse on AWS Redshift - (Apache Airflow, Python, AWS)

- Created a Data Warehouse for a subset of Million songs data and stored on the staging area on AWS S3 using Python
- Loaded JSON formatted files from S3 to AWS Redshift to create Star schema tables in aiming to align with OLAP and BI processes
- Configured DAG workflows in Airflow to maintain and schedule sequential jobs for the pipeline
- Built four different operators that will stage the data, transform the data, and run checks on data quality

MNIST Digit Recognizer - (Python, TensorFlow, Keras)

- Developed models to recognize grayscale images of handwritten digits taken from MNIST dataset
- Built on TensorFlow, the models included: Basic Neural Network, Deep NN, RNN, CNN
- Achieved best accuracy of 99% (approx.) with CNN model; where accuracy is percentage of correct classifications

Taxi Trip duration prediction - (Python)

- Performed Feature Engineering, feature selection, Exploratory Data Analysis on 1200k+ data points to predict Trip
- Duration; Implemented multiple models and performed hyper-parameter tuning which improves performance
- Techniques Used: K-means clustering, PCA, Linear Regression, Random forest, XGBoost