SAI SANTOSH EEDUPUGANTI

saisantosheedupuganti@gmail.com| +1 5184969581|LinkedIn | Github | saisantosh97.github.io

EDUCATION:

Master's in Data Science
University at Albany, SUNY — May 2020

Bachelor of Mechanical Engineering
Jawaharlal Nehru Technological University, Hyderabad — June 2018

TECHNICAL SKILLS:

Programming Languages	**:** Python, R Programming, HTML, CSS
Frameworks/Libraries	**:** Hadoop, PySpark, Hive ,Pandas, NumPy, Scikit, BeautifulSoup, Selenium
DevOps/BI/Tools	: AWS S3, AWS EMR, AWS Batch, Databricks, Docker, Jenkins, Tableau, Qlik sense, Apache Airflow
Databases and repositories	**:** Snowflake, MySQL, MongoDB, GIT, BitBucket, Jira

WORK EXPERIENCE:

Verizon, Data Analyst/ Data Engineer — **January 2021 – Present**

DENA Team -Collects all the Events related to Verizon FIOS media Service and provide Business analytics service to build visualizations provide insights from data.

- Building end to end pipelines using Python scripts with Spark to collect the events from Kafka and land them in Snowflake database.
- Designed and deployed Snowflake schemas and created scripts to perform continuous data loads using snow pipes and Streams from AWS S3(Staging)
- Performed analysis like Peak viewership of channels, Churn rate analysis and figured out the hotspot location of the webpage UI to improve the placement of Advertisements by using PySpark and AWS EMR notebooks.
- Involved in developing Business reports by writing complex SQL queries and building interactive dashboards using Qlik sense
- Introduced and configured Apache Airflow to develop ETL pipelines required for Analytical reporting and Dashboards.
- Participated in solving various Data Quality issues and monitoring data traffic using Splunk.

Technologies: Python, Snowflake, Qlik sense, PySpark, Apache Airflow, Docker, Splunk, AWS S3, AWS EMR, Bitbucket, JIRA

Capital One, Data Analyst/ Data Engineer — **August 2020 - January 2021**

Fraud analytics- Monitors Fraud's in Transactions, identify anomalies developing solutions to solve them

- Performed data migration from SQL Server to Snowflake.
- Automated a vital functionality which identifies data gaps in transactional volumes and raise a Jira ticket and assign it to respective teams.
- Came up with an analytical approach to Flag a transaction as a anomaly by using geo-referenced data and determine if activity is indicative of previous pattern.
- Developed interactive dashboards using Tableau to monitor the spikes and volumes of Transactions
- Reduced cost of manual incident data abstraction by>10% by designing an end-to-end ETL Pipeline using Python and Airflow.

Technologies: Python, Snowflake, Qlik sense, PySpark, Apache Airflow, Docker, Splunk, AWS S3, AWS EMR, Bitbucket, JIRA

Data Analyst, State university Of New York, Albany — **May 2019 - May 2020**

NYSDEC Team- Perform water quality analysis of New York state's water bodies laid throughout the state.

- Build data pipeline from Extracting data through API's using Python and created visualizations over 20 years of data using Tableau
- Created multiple relational database schemas for storing the extracted data (JSON) from API into MySQL server.
- Performed Exploratory data analysis on historical data and came up with insights which helped to identify and solve few waters quality issues.
- Build a Regression models to predict chemical value in water for safer consumption using Linear Regression and Tree based model.
- Supported Multiple data requests like managing, Organizing, and creating easy access data frame for Stakeholders and peers.

Technologies: Python, SQL, MySQL, Pandas, NumPy, Tableau, Scikit learn, GIT

ACADEMIC PROJECTS:

Facebook Friend Recommendation using Graph Mining (Python, Machine Learning)

- Developed a model to predict link whether two users are going to be friend in future or not.
- Created graph-based features like shortest path, Jaccard, Cosine distance, PageRank, Adar index, Katz Centrality and built random forest model, achieved a f-1 score of 93 and performed hyper-parameter tuning for best scores (Precision and recall)
- Tools: Python, NumPy, pandas, Scikit-learn, NetworkX, Random forest, Xgboost

Quora Question Pair Similarity (Python, NumPy, Pandas, Scikit)

- Developed a model to predict whether a pair of questions are duplicate or not, which is useful to instantly provide answer to questions thathave already been answered
- Performed text feature engineering and implemented random forest, logistic regression, linear SVM model, XGBoost with Hyperparameter tuning and achieved a log loss of 0.313.
- Tools: Python, NumPy, Pandas, Scikit-learn, Plotly, NLTK, Beautiful Soup, XGBoost

Taxi Demand Prediction in NYC (Python, SQL, Dask, Pandas)

- Cleaned and performed Exploratory data analysis and feature engineering on New York city taxi data using Python and Dask
- Segmented the data by using K-means clustering and applied baseline models, Linear Regression, Random Forest Regressor, XGBoost Regressorto predict the number of taxi pickups in a given location at a particular time interval and computed a Mean Absolute per error of 11.57
- Tools: Python, Pandas, NumPy, Matplotlib, Sqlite3, Dask, Folium, Gpxpy, Scikit-Learn, XGBoost

Stack Overflow Tag Prediction (Python, NLTK, SQL, Scikit-Learn)

- Collaborated with a team of 3 to design a model to suggest tags for questions posted on Stack Overflow.
- Interpreted data, created statistical analysis of content of questions, framed a multi class classification problem and build logistic regressor and
- support vector regressor to predict tags for content of question.
- Achieved a average F-1 score of 0.51 after necessary hyper-parameter tuning.
- Tools: Python, pandas, NumPy, sqlite3, matplotlib, seaborn, scipy, NLTK, Scikit-learn.