Viraj Shetty

viraj.shetty@utdallas.edu · linkedin.com/in/virajshetty47/ · github.com/VirajVShetty· @virajshetty47
Dallas, TX · +1 (469)388-7873

Skills

Programming: R, SQL, API, JavaScript, Scala, Python (Pytorch, Keras, Pandas, Matplotlib, Numpy).

Tools: PowerBI, Tableau, Alteryx, MySQL, MongoDB, Microsoft Excel (VLOOKUP, Macros, VBA, Index-Match)

Big Data: Hadoop, Sqoop, Hive, Impala, Pig, Flume Kafka, PySpark, AWS (S3, Athena, Redshift, Glue)

Miscellaneous: Linux, Git, Snowflake, Databricks, Big-Query, JIRA, Visual Studio, VMWare, HIPAA Training

Experience

Loopback Analytics, *Dallas, TX* June 2022 - December 2022
Healthcare Data Analyst Intern

- Collaborated with a cross functional team, identified **5 key business metrics**, and conducted Hypothesis (t-tests) on patient data to identify differences in treatment outcomes and improving healthcare solution efficiency by **20%**
- Utilized **PySpark** to manage and update NLP pipelines for EHR data and used **Databricks & R** to perform **Data Cleaning & ETL** on socio-economic data from **Census API** reducing the null value in the existing data by **19%**
- Analyzed client feasibility requirements using **Snowflake**, **JIRA**, and wrote stored procedures using **JavaScript** to optimize execution time and space by **20%**, used **PowerBI** to optimize healthcare reports by **30%**

m-Uni Campus, *Mumbai, IN* December 2019 - May 2020
Data Scientist

- Collected and integrated data from multiple sources including academic performance and lecture data using **Apache Airflow & AWS S3** and redesigned **API** services for greater scalability, reducing manual effort by **50%**.
- Conducted Root cause analysis for factors affecting student performance using regression analysis and identified factors that impacted test scores, used correlation analysis to detect patterns for developing effective solutions.
- Developed a **Python-based** web scraper with **Selenium** to extract **100K** YouTube video links based on search keywords, assigning relevance scores based on metrics like views, likes/dislikes to enhance search result accuracy.
- Visualized Data and implemented machine learning algorithms **Random Forest** & **KNN** and generated reports. Achieved an average precision of **92%** and recall of **89%** for student performance prediction.

Projects

Truck Risk Factor Analysis, *Big Data Project* HDFS | Tableau | PySpark

- Created and executed a data pipeline to ETL **100M** rows of truck data into HDFS and import it into a HIVE table
- Analyzed large datasets connected through **HiveQL** and used **Tableau** to create real time visualization, leading to **28% reduction** in truck risk factors. Created and trained Logistic Regression model with **83%** accuracy

Temperature Anomaly Detection, *Predictive Forecasting Project*, [Code] MongoDB | Python

- Collected temperature readings in India using **MongoDB** as the data storage system and designed an ETL process.
- Applied **Holt-Winter** model with exponential smoothing to capture temperature data components and enhance forecast accuracy. Evaluated model's performance resulting in a high accuracy score - **MAE 2.5** & **RMSE 3.8**.

Fraud Transaction Detection Well Fargo, *ML Project*[Code] Tableau Prep | Python

- Cleaned and prepared **1M** rows of transaction data from Wells-Fargo for analysis by **one-hot encoding** categorical variables and conducted **feature engineering** to identify and select relevant variables
- Implemented **Xgboost** model for fraud detection by using cross-validation techniques such as **grid & random** search to obtain the best hyperparameters, resulting in a robust model with an **RMSE** of **0.459 and 93%** accuracy

Alzheimer's Prediction, *Deep Learning Project*, [Viz] TensorFlow | Matplotlib

- Utilized a dataset of brain MRI images, split into training, validation, and test sets with an **80:10:10** ratio.
- Applied image preprocessing techniques to resize images to **128x128** pixels. Employed dropout regularization method to improve the model's ability to generalize and handle potential overfitting.
- Developed a deep learning model using a 2D CNN to accurately classify brain MRI images and detect dementia with 92% accuracy on test set.

Education

The University of Texas at Dallas, *Dallas,TX* **(GPA: 3.63/4.0)**
MS in Information Technology Management *(Business Analytics Specialization)* August 2021 - May 2023
Relevant coursework: Machine Learning, Advance Statistics for Data Science, Database Foundations, Big Data

The University of Mumbai, *Mumbai, India* **(GPA: 3.5/4.0)**
BE in Computer Engineering June 2016 - November 2020
Relevant coursework: Analysis of Algorithms, Cloud Computing, Natural Language Processing, Data Warehousing