

Vinay Kiran Raju

Data Science

Email: raj.vinay2408@gmail.com | Phone: +1 (201)-492-8306 | LinkedIn: [Vinay Kiran Raju](#)
Portfolio: vinaymeesaraganda.github.io | Blog: [Medium - Vinay Kiran Raju](#)



TECHNICAL SKILLS

- Programming Languages:** Python (Pandas, Beautiful Soup, NumPy, Matplotlib, Seaborn, scikit Learn), Scala, SQL.
- Machine Learning Algorithms:** Random Forest, Fb Prophet, Naive Bayes, Logistic Regression, XG Boost, Decision-Tree.
- Big Data Technologies:** Hadoop, Apache Spark, Apache Kafka, HDFS, Sqoop, Hive, Cloudera, HBase.
- Databases:** MySQL, PostgreSQL SQL Server, Snowflake, Data Modeling, Data Warehousing.
- Visualization Tools:** Tableau, Power BI, MS EXCEL, Google Sheets.
- Orchestration Tools:** Apache Airflow, Ni-Fi.
- Cloud Services:** Amazon Web Services, Microsoft Azure.
- Version Control:** Git, GitHub.

EDUCATION

Pace University, Seidenberg School of Computer Science and Information Systems	New York, USA
MS in Data Science Concentration: Data Analysis, Visualization & Machine Learning GPA: 3.71/4	May 2023
Gandhi Institute of Technology and Management	Andhra Pradesh, India
Bachelor of Technology in Electronics and Communication Engineering GPA: 8.38/10	Apr 2021

WORK EXPERIENCE

Company: Natsoft Corporation	Feb 2023 – Present
Role: Data Engineer	New Jersey, USA
<ul style="list-style-type: none">Performed ETL pipelines to extract data from different sources and load into HDFS and Hive tables using SGOOP.Conducted data transformations and aggregations using PySpark on EMR clusters, ensuring data quality by handling null values and anomalies.Leveraged S3 for data storage and built a data warehouse in AWS Redshift by identifying different dimensions and fact tables for optimized data organization.Generated data visualizations using Tableau to communicate actionable insights, Key Performance Indicators (KPIs), and analytics findings.Implemented Git version control to manage the source code and integrated Git with Jenkins to support build automation.	
Company: Natsoft Corporation	Sep 2020 – Dec 2021
Role: Data Engineer	Hyderabad, India
<ul style="list-style-type: none">Established and managed clusters on Amazon EC2 and deployed EMR to establish big data environments for developing ETL pipelines and workflow.Developed ETL processes for ingesting data from diverse sources into HDFS and Hive using Sqoop. Handled importing and exporting data between HDFS and RDBMS.Processed web URL data using Scala and transformed into Spark Data Frames for analysis using Spark SQL queries.Deployed Spark RDD and Data Frames to rapidly process large datasets, transforming, filtering and analyzing data leveraging Spark's in-memory and lazy evaluation capabilitiesLoaded and processed semi-structured data such as XML, JSON, Avro and Parquet and optimized Spark SQL queries to enhance data access.Orchestrated and scheduled data pipelines with Apache Airflow, utilizing concepts like DAGs, operators, hooks to customize pipeline behavior.Engaged in the development of Code & peer review of assigned tasks and Bug fixing.	

PROJECTS

<u>Building an End-to-End Automated Zillow ETL Pipeline</u>	Jan 2024
<ul style="list-style-type: none">Built a Zillow ETL pipeline using Python, Apache Spark, AWS, and Apache Airflow, seamlessly extracting real estate data from Zillow's API, transforming data on an EMR cluster, and loading the refined dataset into Amazon S3.Automated the ETL workflow with Apache Airflow, integrated Tableau for data visualization, and activated a PySpark script for efficient data transformation, showcasing expertise in end-to-end data engineering and analysis.	
<u>Credit Card Fraud Detection using ML</u>	Dec 2023
<ul style="list-style-type: none">Engineered a machine learning model in Python using Random Forest algorithm to detect fraudulent credit card transactions from a dataset of 550,000 entries. Achieved 99.99% accuracy and 99.98% cross-validation score in identifying fraud.	

- Executed extensive data preprocessing including handling missing values, removing duplicates, standardizing features, and addressing class imbalance.
- Leveraged exploratory data analysis and visualization techniques to uncover patterns and correlations between the 28 principle features provided in the dataset.

British Airways Analysis - Web Scraping

Jun 2023

- Scraped over 1000 customer reviews from Skytrax website using Python library BeautifulSoup and conducting sentiment analysis, revealing trends: 48.9% positive and 43.2% negative and 7.9% neutral sentiments.
- Prepared customer booking dataset for predictive modeling by identifying key variables like purchase lead time, route, flight hour, and length of stay.
- Crafted a Random Forest model in Python predicting customer bookings with 85% accuracy, evaluated model and feature importance, identified improvements by adding promotional offers, payment type.

CERTIFICATIONS

• Snowflake -BUILD 2023 LLM Bootcamp – Snowflake	Dec 2023
• Agile Methodology Virtual Experience Program – Cognizant	Jun 2023
• Data Science Virtual Experience Program – British Airways	Jun 2023
• Azure Data Fundamentals - Microsoft	Apr 2023
• Tableau Desktop Specialist - Tableau	Mar 2023
• Google Data Analytics -Coursera	Nov 2022