Rahul Veerapur

(215) 651-0869 | mv.rahul9@gmail.com | rxv5218@psu.edu | www.linkedin.com/in/rahul-v-rvce999/

Education

Pennsylvania State University Aug. 2024 – *Expected Dec. 2025*
Master of Data Science, *Philadelphia, PA, United States*
Big Data Analytics , Database Design , Large Scale Database , System Design, Statistics, Data Warehousing , Foundation of AI, Data Mining, Machine Learning, NLP, Predictive Analysis.

R.V. College of Engineering Aug. 2018 – Aug. 2023
Bachelor of Engineering in Computer Science, *Bengaluru, India*
Coursework: Data Structures, Algorithms, Operating Systems, System Design, Software Engineering, Machine learning

Experience

Pennsylvania State University, ***Research Assistant (Financial Analyst)*** Oct. 2024 – Dec. 2025
Python, Deep Learning, sickit-learn , Power Bi. *Pennsylvania, United States*

- **Drove data analysis and predictive modeling initiatives** by developing Spatio-Temporal Graph Neural Networks (GNNs) for financial time-series forecasting, delivering a **15% improvement in prediction accuracy** to inform business strategy and planning. **Translated analytical findings into actionable business insights**, supporting key scorecard metrics and long-term strategy.
- **Engineered and optimized large-scale models (100M+ rows)** using GATv2Conv and LSTM architectures, achieving a **20% reduction in validation loss** through advanced hyperparameter tuning with Ray Tune and Optuna.**Enhanced analytical workflows and process efficiency**, improving model deployment agility and enabling more data-driven decision-making.

CME Group, ***Intern*** Jan. 2022 – Jul. 2022
Java, UC4, SQL, Data Warehouse, Automated Reporting System, Power Bi, Tableau, Random Forest, . *Bengaluru, India*

- **Engineered a high-performance, Java-based search engine** to manage over a million databases, enabling efficient data retrieval and seamless integration with data cubes.
- **Optimized processing in RDBMS environment:** Refined SQL query structure and **implementing automated schema for data warehouse achieving a 14% boost** in server efficiency also reducing database lookup.
- **Dynamic Dashboard:**Desgined a **reporting system that provides statistical tracking** such as data visits, connection time,and the type of data retrieved.
- **Predictive Intelligence:** Integrated a **predictive analysis using Random forest with gradient boosting** and also using DBSCAN to determine the information required by employees before initiating a search.

Projects

Real Time Distributed System for trend Analysis

Python, Kafka, PostgreSQL, Cassandra, MongoDb, Spark, Apache Flink

- **Real-time Distributed Data Pipeline Architecture & Implementation :** Designed a real-time distributed system for news trend analysis using Apache Kafka for continuous ingestion and low-latency processing of high-throughput data. Ensuring fault tolerance and scalability via modular, decoupled components.
- **Polyglot Persistence & Scalable Data Storage Solutions :**Engineered a data storage layer utilizing polyglot persistence by deploying MongoDB and apache flink for flexible archival of 2.5M+ raw, unstructured news articles and Cassandra for high-speed, write-optimized storage of time-series metrics, achieving horizontal scalability and fault tolerance across diverse data workloads.
- **Containerized Data Infrastructure & System Optimization** : Orchestrated a containerized data infrastructure using Docker for isolated deployments of Kafka, Spark, Hadoop, MongoDB, Cassandra, and flink enhancing deployment manageability, testing efficiency, and independent component scaling. For system performance for continuous data flow.

Modeling Opioid-Influenced Crash Risk: Insights for Insurance Risk Assessment

Python, Neural Networks, SVM, Hyperparameter Tuning, LightGBM, Random Forest, SHAP, PowerBi, Scikit-learn, Tableau, Plotly

- **Data Engineering & Database Management:** Designed, implemented, and managed robust Python/MySQL data pipelines, integrating and cleaning over 2.5 million records from 5 state and national opioid overdose datasets.This process involved transforming 130+ raw fields into 70+ analytical features, ensuring high data quality and integrity for advanced analytics.
- **Advanced Analytics & Predictive Modeling:** Developed and validated machine learning models (Logistic Regression, Random Forest, XGBoost) to predict opioid-related crash risk and severity, achieving AUC scores up to 0.99. Conducted in-depth statistical and geospatial analyses of over 16,000 opioid-related crashes, identifying key demographic risk segments.
- **Strategic Insights & Quantifiable Business Impact:** Translated complex analytical findings into actionable business strategies, demonstrating potential cost savings exceeding $114 billion with an ROI up to 8000%. Presented data-driven recommendations for optimizing insurance premiums, underwriting policies, and targeted interventions to mitigate opioid-related crash risks.

Technical Skills

Languages and Frameworks: Java, Python, C, C++, SQL, MongoDB, MATLAB, Pandas, Numpy, Tensorflow, Pytorch, Keras, Seaborn.
Technologies: Big Data(Hadoop, Spark, Hive, Kafka,flink), Cloud (AWS S3, Docker, Kubernetes),Predictive Analytics(Regression,Classification,Time Series Analytics),Reporting System(Power BI, Knime,Tableau),Machine Learning(Pytorch,sickit-learn,), NLP.