# VISHAL BAKSHI

Arlington, Virginia | (571) 579-0298 | Mail | LinkedIn | GitHub | LeetCode | Daily Work: Progress

## EDUCATION

**The George Washington University** — Washington, DC
*MS, Data Science* — Exptd May 2025
*Relevant Coursework: Intro to Data Science, Data Mining, Data warehousing*

**Jawaharlal Nehru Technological University** — Hyderabad
*B.Tech. Electronics and Communication Engineering* — Aug 2022

## TECHNICAL SKILLS

**Skills:** R, Python, SQL, MongoDB, Neo4j, Apache Spark, Dockers, NumPy, SciPy, Pandas, Matplotlib, OpenCV, Seaborn, Beautiful Soup, PySpark, ScikitLearn, AWS (EC2, S3, SageMaker, Quicksight), PyTorch, Plotly, Tableau, Linux, PyQt5, Robot Operating System, Git.

## ACADEMIC PROJECTS

**K-means Clustering on Spotify Data using Hadoop, PySpark, and SQL: Link**

- Implemented big data infrastructure using Apache Hadoop and PySpark on a multi-node cluster.
- Built ETL pipelines using Spark SQL module to clean and organize data containing over 20 million records.
- Employed SQL analytics on Spotify 20 million albums, artist dataset, analyzing ratings, trends, genres for actionable insights.
- Established a PySpark session for seamless data access and transformation, implementing K-means clustering with PySpark MLlib to categorize artists.

**NASA's Asteroid Classification: Link**

- Analyzed NASA's real-time asteroid dataset to identify trends and potential threats, applying statistical tests like SMOTE to address data imbalances and enhance predictive accuracy, contributing to risk management strategies.
- Developed and deployed logistic regression, random forests, and decision tree models, highlighting a 6% increase in AUC-ROC score (98%) for balanced datasets compared to unbalanced dataset.

## RELEVANT WORK EXPERIENCE

**HBots Robotics: Machine Learning Engineer** — **Jun 2022 - Aug 2023**

- Engineered and deployed an efficient object detection model using MobileNet V1 within the PyTorch framework, facilitating real-time detection of objects based on volume and live camera feed.
- Key contributor in developing an algorithm for automated image labeling, cutting time by 90% and reducing labor costs by 20% for SkipMyLine. Notably, the algorithm labels over 10,000 images in under 5 minutes, enhancing efficiency.
- Researched and implemented innovative methods to freeze trained layers during new data batch training, enhancing model performance and scalability by reducing training time by 40%.

**The UnEarth: Associate Research & Development Intern (Non-Profit)** — **Jan 2021 - Jan 2022**

- Developed and implemented a high-accuracy student performance prediction model using linear regression and scaling techniques, achieving a remarkable 98% adjusted R-squared with an MAE of 1.6.
- Identified study hours as the most influential feature directly proportional to performance index, utilizing real-time student data pooled from The UnEarth Collaborated schools to drive actionable insights and inform decision-making processes.
- Conducted workshops for 200+ underprivileged students, enhancing their skills in machine learning, and robotics.

## CERTIFICATIONS

- DataCamp: Introduction to R *[Show Credential]*
- Google Cloud: Diwali Game 1 *[Show Credential]*
- Google Cloud: Diwali Game 2 *[Show Credential]*
- Google Cloud: Diwali Game 3 *[Show Credential]*

## ADDITIONAL INFORMATION

- **Organization:** Led Sahara Research & Development Club as Team Lead for a tenure of 3 years.
- **NGO:** Hosted emerging tech workshop with Telangana Gov. & Deloitte, aiding 800+ underprivileged students. *[Link]*
- **Award:** Best Social Cause of the Year by BITS PILANI *[Link]*
- Achieved victory in over 20 technical competitions in robotics at esteemed institutions such as IITs and NITs, reflecting dedication and proficiency in these fields. *[Link]*