Yashwanth Vijayaragavan

yavijay@iu.edu +1 8129742092 Bloomington, Indiana Linkedin Github

PROFILE

Data Science graduate from Indiana University with 2+ years of experience in developing scalable ETL pipelines, automating reporting workflows, and supporting performance metric analysis. Worked across academic and industry settings on data quality automation, Snowflake migration, and NLP-driven pipelines for financial research. Proficient in Python, SQL, Power BI, AWS, and Spark, with a focus on delivering actionable insights to support strategic decisions and operational efficiency.

EDUCATION

Master of Science in Data Science, Aug 2023 – May 2025 | Bloomington, Indiana
Indiana University Bloomington, Luddy School of Informatics, Computing, and Engineering

SKILLS

Programming Languages: Python (NumPy, Pandas, SciPy, Scikit-learn, PyTorch, BeautifulSoup, Requests), R, SQL

Databases: MySQL, BigQuery, Databricks, Snowflake, MongoDB, Azure Data Lake, SQLite

Data Visualization Tools: Power BI (Certified), Tableau, SAP Analytics Cloud, Microsoft Excel, Python (Matplotlib, Seaborn, Plotly), Streamlit

Machine Learning & NLP: Hypothesis Testing, PCA, SMOTE, Clustering (K-Means, DBSCAN), Anomaly Detection, Association Analysis, Fraud Detection (Ensemble Methods), TF-IDF, Dimensionality Reduction, Large Language Models (LLMs), Document Semantic Search (FAISS)

PROJECTS

Scalable Data Pipeline Development for Real-Time Traffic Analytics, *Azure Databricks | SparkSQL | CI/CD | ETL*
- Engineered a real-time ETL pipeline handling **100,000+ traffic records/day**, using Spark SQL and Medallion Architecture on Azure.
- Implemented and deployed Power BI dashboards that improved decision-making speed by **30%**.
- Conducted deployments via Azure DevOps, reducing manual deployment time by **80%**.

NYC Crime Analytics: End-to-End Data Pipeline, *Google BigQuery | SQL | Looker Studio | API Integration*
- Created a serverless ETL pipeline to ingest and process **1.2M+ crime records** from the NYC Open Data API into BigQuery.
- Applied **12+ SQL transformations** for data cleaning, time-based features, demographics, and geolocation mapping.
- Created a **3-page Looker Studio dashboard** visualizing crime trends, hotspots, and severity for public safety insights.

K-Means RFM Customer Segmentation on E-Commerce Data, *Python | RFM analysis | clustering| anomaly detection*
- Analyzed **21,500+ transactions**, segmented customers into 4 key personas using K-Means, improving targeting precision.
- Detected outliers with DBSCAN, highlighting **2% of customers** with atypical high-value behaviors.
- Leveraged Seaborn and Matplotlib to create visualizations of customer behavior, resulting in the identification of key trends.

Fraud Detection with PCA, SMOTE, and Ensemble Models, *Python | PCA | SMOTE | ensemble learning | evaluation metrics*
- Analyzed **1M+ transaction records** with severe class imbalance (~1.21% fraud) using EDA and category-level fraud patterns.
- Applied **SMOTE** post-split to balance classes and engineered features for age, category, and transaction types. Trained and tuned **Logistic Regression, KNN, Random Forest, XGBoost**, and ensemble models; **XGBoost** achieved highest recall and ROC-AUC with lowest false negatives.

NYC Motor Vehicle Collision Analysis with Spatial Insights, *Python | EDA | spatial mapping | data cleaning*
- Processed **1.2M+ rows** of collision data, pinpointing top 5 high-risk intersections in NYC using spatial mapping.
- Produced actionable heatmaps and visuals that could inform city safety initiatives.

Sales and Customer Analytics Dashboards, *Tableau | data visualization | trend analysis*
- Developed Tableau dashboards analyzing **$2.5M+ in sales** across regions, surfacing key profit trends and customer behaviors.
- Simplified strategic planning by enabling 3x faster access to category-level insights.

PROFESSIONAL EXPERIENCE

Data Scientist, *Indiana University Bloomington* May 2024 – Dec 2024 | Bloomington, Indiana
- **Spearheaded end-to-end development** of NLP-driven pipelines to extract and classify content from **5,000+ SEC 10-K filings**, enabling **revamped sector tagging** for financial analysis.
- Implemented **Python (BeautifulSoup, Requests)** and **EDGAR APIs** for large-scale document scraping and preprocessing; integrated **MongoDB** for structured storage of unstructured text.
- Implemented a clustering framework using **TF-IDF**, **cosine similarity**, and dimensionality reduction to group companies by disclosure themes; evaluated clustering techniques including **K-Means** to ensure meaningful groupings and improve research efficiency.
- Collaborated with **Professor Leslie Hodder** to fine-tune classifications and support publication of insights on **corporate reporting standards**.

Data Engineer, *Cognizant* Feb 2022 – Jun 2023 | Chennai, India
- **Led development of ETL pipelines** for BHP Group using Informatica PowerCenter, processing **500K+ records daily** across 10 source systems with built-in integrity checks.
- **Migrated data infrastructure to Snowflake**, cutting compute costs by **30%** and improving query performance and scalability.
- Composed **modular, reusable pipeline components**, reducing new data source onboarding time by **50%** through automated schema detection and logic blocks.
- Developed a **Python-based data quality suite** with **95% anomaly detection accuracy**, resolving **80+ critical issues** and enhancing regulatory compliance.

CERTIFICATION

- Microsoft Certified: Power BI Data Analyst Associate (PL-300)
- SQL Associate by DataCamp
- PwC Switzerland - Power BI Job Simulation
- Robotics And Control : Theory and Practice (NPTEL)