Shubham Pant

+1-917-938-5504 | pant.shubham95@gmail.com | Github | LinkedIn

Professional Summary

Data Engineer with expertise designing and optimizing scalable data pipelines on Azure. Skilled in building interactive dashboards, automating ETL workflows, and enabling real-time analytics across enterprise platforms, reducing manual reporting by 70–80%. Proficient in Azure Data Factory, Power BI, SQL, Python, and cloud-based ETL, with a track record of transforming complex, high-volume datasets into actionable insights that drive measurable business decisions.

Technical Skills

Programming Languages: Python, SQL, JavaScript, TypeScript
Cloud Platforms: AWS , Microsoft Azure
Data Engineering & ETL: Azure Data Factory, AWS Glue, dbt (Data Build Tool), Data Pipeline Design, Data Modeling, Query Optimization
Databases: MSSQL, PostgreSQL, Elasticsearch
Visualization & BI: Power BI, Tableau, Kibana, Microsoft Excel, Vega
Machine Learning: Linear/Logistic Regression, Decision Trees, Random Forest, XGBoost, SVM, Naive Bayes, KNN, KMeans, TensorFlow, Scikit-learn
Python Libraries: NumPy, Pandas, Matplotlib, Seaborn, SciPy
Tools & Methodologies: Git, GitHub, Jupyter Notebook, VS Code, Google Colab, Agile, Waterfall, SDLC

Work Experience

Data Engineer — Nov 2024 – Present

MarketOnce — *Jacksonville, FL*

- Designed and delivered 15+ interactive Power BI dashboards and paginated reports, reducing manual reporting effort by 70% across cross-functional teams
- Built and optimized end-to-end data pipelines in Azure Data Factory using complex SQL queries to extract, transform, and load large-scale structured and semi-structured datasets
- Partnered with Customer Experience and HR stakeholders to transform survey data using Elasticsearch, SQL, and Vega, delivering real-time interactive dashboards that reduced manual reporting by 80%
- Integrated and consolidated data from CRM systems, REST APIs, and marketing platforms to enable centralized, scalable analytics and reporting
- Improved data modeling, indexing strategies, and query performance tuning to enhance reporting efficiency, data reliability, and pipeline scalability

Data Science Intern — May 2024 – Aug 2024

ZSAnalytics, Inc — *New York, NY*

- Engineered Random Forest and Logistic Regression models achieving 80% accuracy by identifying 15 key predictive features, including rolling purchases and regional purchase rates, to enhance targeted marketing strategies
- Designed and implemented an NLP model using TF-IDF vectorization and cosine similarity to identify the top 5 semantically closest sentences to a given input, improving text analysis efficiency
- Built a custom GPT-based model capable of analyzing full-length books and generating contextually accurate responses, improving content accessibility and enabling interactive, insight-driven exploration

Data Processing Analyst — Jun 2017 – Aug 2023

Toluna India Pvt Ltd — *Gurugram, HR*

- Led a team of 4 through the full survey lifecycle, delivering complex online surveys and accurate, on-time data processing for enterprise clients
- Applied advanced data processing techniques using Excel and SPSS, improving analytical efficiency and accuracy for large-scale market research projects
- Built complex, multi-level (up to level 5) web-based surveys for market research using Confirmit, Qualtrics, Decipher, Toluna Start, HTML, CSS, JavaScript, jQuery, and SQL
- Managed 4 monthly and 2 annual data trackers, providing strategic recommendations on questionnaire design and validation logic
- Optimized questionnaire design in collaboration with research teams, leveraging deep survey programming expertise to improve data quality

Projects

Analyzing Chicago Crime Dataset Using Machine Learning Models

- Led analysis of an 800,000-row Chicago crime dataset using Google BigQuery to uncover crime trends and behavioral patterns
- Built and evaluated predictive models (Logistic Regression, Random Forest, XGBoost) plus an AdaBoost ensemble model, achieving 68% accuracy in predicting crime type with XGBoost
- Performed statistical analysis with SciPy and geospatial visualization using Matplotlib, Seaborn, and Folium to identify crime hotspots for targeted prevention strategies

Unveiling New York Air and Water Quality Dynamics with AWS Insights

- Analyzed air and water quality trends across New York (2020–2023) by integrating data from multiple public sources
- Automated data processing and storage into Amazon RDS using AWS Glue for ETL workflows, with AWS SNS configured for real-time email notifications
- Developed interactive Tableau visualizations to surface trends and support data-driven decision-making

Lead Scoring Case Study

- Performed data cleaning and exploratory data analysis using Pandas and NumPy on a lead-conversion dataset
- Built and tuned a predictive model using a Generalized Linear Model (GLM) with feature scaling and Recursive Feature Elimination (RFE), achieving 88.67% test accuracy
- Evaluated model performance via confusion matrix, ROC curve, precision, recall, sensitivity, and specificity, improving lead conversion prediction accuracy to 80%

Education

Yeshiva University Aug 2023 – Aug 2024
Master of Science, Data Analytics and Visualization *New York, NY*

Uttarakhand Technical University Jun 2012 – Jun 2016
Bachelor of Technology, Computer Science *India*

Certifications & Achievements

AWS Certified Cloud Practitioner – Amazon Web Services
Advanced Certificate Program in Data Science – IIIT-Bangalore & upGrad
Core Java Certification – HCL CDC