Ullas Umesh

Phone: (+44) 07721972469 (Mobile) | **Email:** ullas0210@gmail.com | **LinkedIn:** www.linkedin.com/in/ullas-umesh/ | **GitHub:** https://github.com/Ullas-Umesh?tab=repositories | **Address:** Newcastle upon Tyne, United Kingdom (Home)

ABOUT MYSELF

I'm a data scientist with strong expertise in building and deploying machine learning solutions within Databricks using PySpark and Spark SQL. My work focuses on developing end-to-end model workflows that are scalable, reliable, and production-ready. I have a proven track record of leveraging Power BI to deliver real-time, interactive dashboards that translate complex data into actionable insights for business.

WORK EXPERIENCE

AKZONOBEL – NEWCASTLE UPON TYNE, UNITED KINGDOM
DATA SCIENTIST – 01/07/2024 – CURRENT

Spearheaded the development of scalable data pipelines using Databricks with PySpark and Spark SQL, enabling real-time analytics and ML experimentation on large-scale coating performance data.
Implemented the Medallion architecture (bronze, silver, and gold layers) to ensure clean, reliable, and query-optimized datasets across Azure Data Lake.
Orchestrated end-to-end workflows and ML model testing within Databricks notebooks, supporting rapid experimentation and deployment of predictive models for product performance forecasting.
Integrated with Azure DevOps for source control, CI/CD of notebooks, and reproducible pipeline runs.
Built Power BI dashboards on top of curated Gold layer tables, reducing reporting turnaround by 30% and helping R&D identify underperforming products 2x faster.
Automated repetitive workflows using Power Automate, streamlining manual reporting and data refresh processes.
Collaborated closely with R&D and product teams to translate analytical insights into data-backed development strategies across 3 business units.

TATA CONSULTANCY SERVICES – BENGALURU, INDIA
DATA ANALYST – 02/10/2021 – 21/08/2023

Utilized PostgreSQL and SAP HANA to extract, transform, and load (ETL) large datasets, ensuring data accuracy and consistency across multiple platforms. Conducted statistical analysis using Pandas, NumPy and Matplotlib to identify trends, correlations, and outliers, driving data-informed decisions.
Built interactive dashboards and visualizations in Power BI, enabling stakeholders to explore data and derive actionable insights in real time.
Cleaned and preprocessed raw data in alignment with DAMA-DMBOK standards, enhancing data quality and analytical reliability.
Optimized API data ingestion by implementing Python multiprocessing, reducing data retrieval time by over 60% and significantly accelerating downstream processing in Databricks pipelines.
Employed A/B testing and other experimental design methodologies to assess the impact of process changes and new implementations on system performance.
Conducted data mining and exploratory data analysis (EDA) to uncover hidden patterns and insights, directly contributing to the optimization of trade operations.
Automated routine data analysis tasks using Python scripting and Excel macros, reducing manual workload and increasing efficiency.

SKILLS

Python (pandas, numpy, matplotlib) | R | SQL | Databricks/Pyspark | ETL | Power BI | Machine Learning | Scikit-learn | TensorFlow | SAP - HANA | Microsoft Azure | Data Processing | Deep Learning | Data Mining | Data Modelling | Statistical Analysis | Data Visualisation | Data Cleaning | PostgreSQL | Azure DevOps | Python Multiprocessing | Microsoft Excel | Power Automate | Git | Big Data Analytics

EDUCATION AND TRAINING

03/10/2023 – CURRENT Nottingham
MSC DATA SCIENCE Nottingham Trent University

Final grade DISTINCTION

07/07/2017 – 30/08/2021
BE MECHANICAL ENGINEERING BMS Institute of Technology and Management

Final grade 7.77 CGPA

PROJECTS

Fetal Health Prediction using Machine Learning

Developed and deployed machine learning models for predicting fetal health status using cardiotocography (CTG) data, contributing to a 30% improvement in early risk detection.

Performed data preprocessing including outlier detection, feature scaling, and stratified data splitting, ensuring clean and balanced input for model training.

Implemented and evaluated SVM, MLP, CNN, Random Forest, and XGBoost models; achieved 95.3% accuracy and 0.96 F1-score with XGBoost after hyperparameter tuning.

Leveraged Azure AutoML to streamline model selection and optimization, improving pipeline efficiency and model performance.

Conducted clustering analysis using Gaussian Mixture Model, KMeans, and Agglomerative Clustering, achieving a silhouette score of 0.67 with the latter.

Addressed AI ethics in healthcare, focusing on GDPR-compliant data privacy, fairness audits, and transparent model interpretability using SHAP and LIME.

Health Insurance Data Analysis

Conducted a comprehensive analysis of health insurance data, examining the impact of demographic and lifestyle variables on insurance charges across 1,300+ individuals.

Performed data cleaning and preprocessing using missing value imputation, outlier detection (IQR method), and log transformations, improving data quality and reducing skewness by 25%.

Used descriptive statistics to summarize continuous variables, identifying high variance in charges (std dev ≈ 12,000) and outliers primarily linked to smoking status and BMI.

Applied Kolmogorov-Smirnov tests for normality, Pearson/Spearman correlations ($r > 0.7$ for BMI and charges), Chi-square tests, and Mann-Whitney U tests to assess statistical associations.

Executed multiple linear regression, achieving $R^2 = 0.74$, indicating that age, BMI, smoking, and region explained 74% of charge variation.

Used t-tests and Kruskal-Wallis tests to validate significant differences in charges across smoker vs. non-smoker groups ($p < 0.001$) and across age brackets.

Analyzed regional BMI differences using ANOVA ($F(3, 1296) = 5.32$, $p < 0.01$), with post-hoc Tukey tests confirming significant gaps between the Southeast and other regions.

Summarized findings show smoking increases insurance charges by over 100%, with policy recommendations focused on preventive care and lifestyle interventions.

CERTIFICATIONS

AZ - 900 Azure Fundamentals

AI - 900 Azure AI Fundamentals

Databricks Fundamentals

MongoDB Python Developer

DATA VISUALIZATION

Dashboards

Marine Product Performance Dashboard: Designed a specialized Power BI dashboard for analyzing the performance of Marine Protective and Yacht Coatings across nine global trade routes. This dashboard assists the marketing and R&D teams in determining the most suitable products for different trade routes, optimizing product marketing strategies, and solving customer issues with tailored solutions.

Raft Management Dashboard: Created a comprehensive Power BI dashboard for managing test panels on rafts deployed across various environmental conditions. This tool allows raft owners to efficiently track available versus occupied slots, facilitating optimal utilization of resources and providing an overview of raft deployment status across multiple locations.

Weather Data Tracking Dashboard: Developed an interactive Power BI dashboard to visualize and track weather data during scientific experiments. This dashboard enabled researchers to correlate experimental results with specific weather conditions, enhancing the accuracy of their findings.