

# Sourabh Sonker

+91 8447727696 | [sourabhsonker17@gmail.com](mailto:sourabhsonker17@gmail.com) | [LinkedIn](#) | [GitHub](#) | [Portfolio](#)

## SUMMARY

A highly motivated Data Scientist with a B.Tech in Mechanical Engineering, bringing a rigorous, analytical mindset to the data science domain. After self-directing an intensive, project-based curriculum, I have developed proven expertise in the end-to-end machine learning lifecycle—from deep-dive SQL analytics and building automated Python pipelines to deploying production-grade AI models using modern MLOps principles. Eager to apply a unique blend of engineering discipline and hands-on data expertise to solve complex business challenges.

## TECHNICAL SKILLS

**Languages:** Python, SQL

**Data Analysis & Visualization:** Pandas, NumPy, Matplotlib, Seaborn, Plotly, Power BI, Advanced Excel (Power Query, PivotTables)

**Machine Learning:** Scikit-learn, XGBoost, TensorFlow, Keras, NLTK, Gensim, Transformers (Hugging Face), CausalML, Reinforcement Learning (Gymnasium)

**MLOps & Data Engineering:** AWS (Fargate, SQS, DynamoDB), Docker, CI/CD (GitHub Actions), DVC, MLflow, Streamlit

**Developer Tools:** Git, GitHub, VS Code, Jupyter Notebook

## PROJECTS

### Predictive Maintenance & Explainable AI (XAI)

*Python, XGBoost, Scikit-learn, SHAP, Streamlit*

- Engineered over 40 time-series features from raw sensor data to predict component failure in a manufacturing setting, handling a highly imbalanced dataset of 60,000 readings.
- Trained and fine-tuned an XGBoost classifier, achieving 99%+ precision on the minority (failure) class and identifying failures up to 7 days in advance.
- Developed a cost-analysis model demonstrating an estimated **\$1.6M in savings** on the test set by preventing unplanned downtime.
- Deployed the model and SHAP-based explanations in an interactive Streamlit dashboard for non-technical stakeholders.

### Reproducible Causal Inference Pipeline (MLOps)

*Python, DoWhy, DVC, MLflow, AWS S3*

- Architected a reproducible MLOps pipeline to measure the causal impact (uplift) of a marketing campaign, integrating DoWhy for causal modeling.
- Implemented DVC for data versioning with an S3 remote and MLflow for tracking experiments, parameters, and model artifacts, ensuring full reproducibility.
- Performed sensitivity analyses (e.g., random cause and placebo treatment) to validate the robustness of the causal estimate.
- Demonstrated a **7% increase in model-estimated lift** (from \$3.41 to \$3.65) by versioning data and code to correctly handle outliers.

### End-to-End Vendor Performance BI Dashboard

*Power BI, DAX, Power Query, SQL, Python*

- Developed a comprehensive ETL process using Python and SQL to ingest, clean, and model vendor purchasing data from multiple CSV sources into a central SQLite database.
- Engineered key business metrics (e.g., Purchase Dependency %, Margin %, Turnover) using complex DAX measures in Power BI.
- Designed an interactive, multi-page dashboard to visualize vendor performance, revealing that **66% of purchases depended on the top 10 vendors** and identifying high-margin, low-turnover brands for strategic review.

## EDUCATION

### B.Tech in Mechanical Engineering

Delhi Technological University, Delhi —Graduated August 2024

## CERTIFICATIONS

- Data Science & Machine Learning Track (Kaggle):** Completed 10+ micro-courses including Python, Pandas, Advanced SQL, Feature Engineering, Deep Learning, and ML Explainability.
- Excel Skills for Business Specialization (Coursera):** Essentials.
- Excel Skills for Business Specialization (Coursera):** Advanced.
- Basic Statistics (Coursera):** University of Amsterdam.