

Aman Ostwal

Philadelphia, PA (Open to Relocation) • ostwalaman7@gmail.com • (215) 669-1393

[LinkedIn \(https://www.linkedin.com/in/ostwalaman/\)](https://www.linkedin.com/in/ostwalaman/) • [GitHub \(https://github.com/ostwalaman/\)](https://github.com/ostwalaman/)

PROFESSIONAL SUMMARY

Data Scientist and Software Engineer with 3+ years building forecasting, NLP, predictive analytics, and production software solutions. Skilled in Python, SQL, PySpark, machine learning, cloud workflows, and stakeholder reporting, with measurable impact across research, agriculture, and consumer technology through scalable pipelines, model evaluation, automation, and performance optimization.

EDUCATION

Master of Science in Data Science, *Drexel University* – GPA: 3.76

Sep 2023 – Jun 2025

Bachelor of Engineering in Computer Engineering, *SPPU*

Aug 2017 – May 2022

SKILLS

Languages & ML: Python, SQL, PySpark, R, Scikit-learn, XGBoost, Random Forest, NLP, TF-IDF, spaCy, Forecasting, Model Evaluation | **Data, BI & Cloud:** Power BI, Tableau, Looker, BigQuery, Databricks, AWS, GCP, ETL, Data Pipelines, Data Validation, Feature Engineering | **Developer Tools & Methods:** Git, GitHub, Jupyter, APIs, A/B Testing, Hypothesis Testing, Regression, Classification, Recommender Systems, EDA

PROFESSIONAL EXPERIENCE

Research Assistant, *Drexel University College of Computing and Informatics*

Jun 2025 – Present

- Built a Python-based CommonCrawl WARC.WET preprocessing pipeline for LLaMA-inspired corpus preparation, processing 30K+ raw records into 10K+ quality-filtered documents through staged extraction, cleaning, deduplication, and page filtering
- Expanded corpus diagnostics with langid, NLTK, TextBlob, textstat, and MinHash to evaluate mixed-language content, encoding issues, readability, sentiment, POS patterns, lexical richness, and near-duplicate text across research datasets
- Hardened JSONL streaming and reporting workflows across 5 filtering stages, improving pipeline reliability by skipping malformed rows, surfacing drop reasons, and visualizing retention from extraction through exact deduplication

Teaching Assistant, *Drexel University College of Computing and Informatics*

Jun 2024 – Jun 2025

- Led Python, SQL, and Power BI labs for 60+ students, helping students build reproducible analytics workflows, KPI reports, and model-evaluation practices across applied data science coursework
- Reviewed student code, SQL queries, dashboards, and analytical writeups to improve debugging discipline, data storytelling, metric interpretation, and technical accuracy across assignments
- Supported hands-on instruction in analytics, reporting, and data science workflows by explaining implementation logic, correcting errors, and reinforcing reproducible project practices

Data Scientist, *SmartAgri AgroVille – Smart Agricultural Solutions*

Jun 2024 – Dec 2024

- Deployed an XGBoost forecasting pipeline on GCP that reduced crop spoilage from ~18% to ~12.5%, a ~30% reduction across multi-region agricultural planning workflows
- Engineered SQL and Python feature pipelines using soil and weather data, lowering RMSE from 0.52 to 0.43 and improving forecast reliability by ~18% for operational stakeholders
- Automated Power BI dashboards for soil, weather, and crop-health metrics, reducing manual reporting from ~12 hours/week to ~3 hours/week and enabling faster self-service decision-making
- Integrated ETL and validation workflows across multi-source agricultural datasets, reducing missing and invalid records from ~14% to ~5% and significantly improving data quality for modeling and analysis

Software Engineer, *gurado GmbH – Digital Gift Card Platform*

Sep 2021 – Aug 2023

- Optimized PHP APIs, SQL queries, caching layers, and CDN usage to reduce server response time from ~900 ms to ~270 ms, a ~70% improvement supporting 100K+ active users
- Improved page-load performance from ~6.5 seconds to ~1.1 seconds, an ~83% reduction, through query tuning, caching, and delivery optimization for customer-facing workflows

- Led Angular-to-React migration for customer-facing workflows, resolving iOS deployment blockers and improving release cycles from biweekly to weekly across distributed teams
- Conducted production-grade code reviews and release coordination, improving deployment success rates from ~85% to ~97% and reducing post-release defects by ~25% across e-commerce and user-facing systems

PROJECTS

NeuroNexus - Reddit Mental Health NLP

- Built a canonical Reddit NLP pipeline with PRAW, spaCy, NLTK, TF-IDF, and scikit-learn to ingest, clean, merge, and model 28K+ posts from r/anxiety, r/depression, and r/ADHD for disorder classification and severity scoring
- Achieved 0.827 weighted F1 and 82.87% accuracy with XGBoost using unigram-to-trigram TF-IDF features, improving disorder classification performance across noisy, weakly labeled Reddit text
- Built a Random Forest severity model with 0.39 MSE, 0.36 R², and 91.62% of predictions within ±1 severity point, providing a practical accuracy measure for weak-supervision labels
- Generated weak-supervision labels using disorder keywords, intensifiers, qualifiers, and medication mentions, enabling scalable experimentation across 28K+ posts without clinician-annotated training data

Hybrid Recommendation System

- Built a hybrid recommender using ALS, SVD, PCA, k-NN, and Jaccard similarity on MovieLens data to address sparse user-item interactions and cold-start recommendation scenarios
- Benchmarked recommendation models on 250K+ ratings and 59K movies, achieving RMSE between 0.668 and 0.912 across comparative validation runs
- Improved recommendation experimentation by comparing collaborative filtering, dimensionality reduction, nearest-neighbor similarity, and hybrid modeling approaches across offline evaluation workflows

CERTIFICATIONS

- | | |
|---|----------|
| ● OCI 2025 Certified Data Science Professional (Oracle University) | Nov 2025 |
| ● OCI 2025 Certified Generative AI Professional (Oracle University) | Oct 2025 |
| ● SQL for Data Science (UC Davis, Coursera) | May 2025 |