

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 of experience building **forecasting**, **NLP**, and **predictive analytics** solutions across research, agriculture, and large-scale consumer software. Applies **Python**, **SQL**, **XGBoost**, **Random Forest**, **TF-IDF**, and **cloud workflows** to deliver reproducible data pipelines, rigorous model evaluation, and stakeholder-ready insights with measurable impact.

SKILLS

- **Programming:** Python, SQL, PySpark
- **Machine Learning:** XGBoost, Random Forest, Scikit-learn, NLP, Forecasting, Recommender Systems
- **Statistics / Evaluation:** Model Evaluation, A/B Testing, Hypothesis Testing, Classification, Regression
- **Data:** EDA, Feature Engineering, Data Validation, ETL, Data Pipelines, TF-IDF
- **Visualization:** Power BI, Tableau, Looker
- **Cloud / Platforms:** AWS, GCP, Databricks, BigQuery
- **Tools:** Git, Jupyter, GitHub, spaCy, PRAW

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

PROFESSIONAL EXPERIENCE

Research Assistant, *Drexel University College of Computing and Informatics* Jun 2025 – Present

- Built a **CommonCrawl WARC.WET** preprocessing pipeline in **Python** for **LLaMA-inspired** corpus preparation to extract 30K+ conversion records and retain 10K+ quality-filtered documents through staged cleaning, deduplication, and page filtering
- Expanded corpus analysis with **langid**, **NLTK**, **TextBlob**, **textstat**, and **MinHash** to detect mixed language, script and encoding issues, readability, sentiment, POS patterns, lexical richness, and near duplicates for research-ready text datasets
- Hardened **JSONL** streaming and reporting workflows to skip malformed rows, surface drop reasons, and visualize retention across extraction, filtering, line deduplication, page filtering, and 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 to strengthen reproducible analytics, **KPI** reporting, and model-evaluation practices across course projects
- Reviewed student code, queries, and dashboards to improve analytical accuracy, debugging discipline, and data storytelling across technical assignments

Data Scientist, *SmartAgri AgroVille – Smart Agricultural Solutions* Jun 2024 – Dec 2024

- Deployed an **XGBoost** forecasting pipeline on **GCP** to reduce crop spoilage by 30% across multi-region agricultural planning workflows
- Engineered **SQL** and **Python** feature pipelines from soil and weather data to lower prediction error by 18% and improve forecast reliability for operational stakeholders
- Automated **Power BI** dashboards across soil, weather, and crop-health metrics to enable self-service reporting and accelerate planning decisions across teams
- Integrated extraction, transformation, and validation workflows across multi-source datasets to improve data quality for ad hoc analysis and predictive modeling

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 by 70% for 100K+ active users
- Improved page-load performance by 83% through query tuning and delivery optimization, strengthening customer experience and production reliability at scale

- Led an **Angular-to-React** migration for customer-facing workflows to resolve iOS deployment blockers and improve release velocity across distributed teams
- Conducted production-grade code reviews and release coordination to improve deployment safety, code quality, and system reliability

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** on **TF-IDF** unigram-to-trigram features, and reached **0.39 MSE**, **0.36 R²**, and **91.62%** within ±1 severity point with **Random Forest** severity modeling
- Generated weak-supervision labels from disorder keywords, intensifiers, qualifiers, and medication mentions to scale experimentation without clinician-annotated data

Hybrid Recommendation System

- Built a hybrid recommender with **ALS**, **SVD**, **PCA**, **k-NN**, and **Jaccard similarity** on MovieLens data to address sparse interactions and cold-start recommendation scenarios
- Achieved **RMSE 0.668–0.912** across comparative validation runs on **250K+ ratings** and **59K movies**, improving personalization through model tuning and offline benchmarking

CERTIFICATIONS

- | | |
|--|----------|
| ● OCI 2025 Certified Data Science Professional (Oracle University) | Nov 2025 |
| ● SQL for Data Science (UC Davis, Coursera) | May 2025 |
| ● Exploring Data Transformation with Google Cloud (Google Cloud, Coursera) | Apr 2025 |