

Zara Sarkar

✉ sarkar.z@northeastern.edu ☎ 978-908-8840 🇺🇸 US Citizen 📍 7 Country Road, Westford MA 01886 🌐 github.com/zara-sarkar
🔗 luxxify.streamlit.app

🎓 EDUCATION

MS in Computer Science - Data Science Concentration, *Northeastern University*

Present

BS in Data Analytics - Minors in CS and Business, *Bentley University*

Aug 2020 – Dec 2023

Technical Skills, *Python, R, Spark, Java, SQL, Postgres, AWS Glue, AWS Athena, Databricks, Excel, Tableau, D3*

Relevant Courses, *Data Mining, Machine Learning, Causal Inference, Data Structures and Algorithms, Databases, Web Dev, Linear Algebra*

📁 PROFESSIONAL EXPERIENCE

Data Science Intern - Management Solutions, *May 2023 - Aug 2023*

- Developed time series regression **AR, ARIMAX**, enhancing **ARIMAX** accuracy by **10%** in **Python**
- Performed **causal analysis** on financial Bayesian credit risk models and presented climate risk features to **Director of Data Science**
- Predicted likelihood of credit default with **logistic regression** in **Python** and improved **F1** score by **15%** with **feature selection**

Data Analytics Intern - American Tower, *May 2022 - Aug 2022*

- Managed live **business intelligence** dashboards using **D3, SQL, JavaScript** and introduced **VIEW** to combine data from different sources
- Designed **materialized views in SQL** to consolidate data and increased reporting speed by **50%** using **CTE, INNER JOIN, UNION**
- Accelerated dashboarding response and stability by **50%** and showcased dashboards to **VP of Shared Services**
- Developed custom interactive **data visualizations** for **live business intelligence dashboards** with **D3**
- Validated datasets in **Excel** using **VLOOKUP, HLOOKUP, XLOOKUP**

📁 LUXXIFY: REALTIME MAKEUP RECOMMENDER

LINK TO SITE 🔗

- Designed **machine learning personalization model** to recommend makeup products using **RandomForest**
- Developed **end-to-end async web scraping pipeline** with **Python, Selenium, BeautifulSoup, Asyncio** extracting over **300k customer reviews** from Ulta with **40% increase** in efficiency for **data lake**
- Integrated **data engineering ETL pipeline** to **Databricks data warehouse** to clean data from **multiple JSON sources** using **Postgres**, eliminating **98% of data inconsistency and defects**
- Trained custom **Word2Vec word embedding model (natural language processing)** on **100MB of Reddit makeup data** to create interpretable features
- Deployed **Apache Spark UDFs** for feature engineering via **Databricks**, improving processing speed by **40%**
- Boosted recommender **F1 score by 30%** with domain specific natural language processing features from embedding model
- Applied **oversampling** and **undersampling**, improving the **F1 score by 12%** and preventing data leakage

📁 ACADEMIC PROJECTS

Causal Inference and Generalized Linear Models with R

- Designed, powered, randomized an **A/B test** with **ANOVA** to **model different advertising techniques for a business**
- Tested assumptions and hypotheses to assess validity of **A/B test** with **Shapiro-Wilks test, QQ plots, and T-tests**
- Designed a **difference-in-differences model** in **R** to test effectiveness of advertising models
- Predicted rate of fissures for a wind turbine dataset by constructing and tuning a **binomial regression statistical model**
- Analyzed distributions to apply link functions and evaluated link functions with residuals to improve accuracy by **30%**
- Determined relationships of **categorical variables** in advertising model with **chi-squared test**
- Assessed model with **stepwise selection to select best features**

Predicting Customer Attrition with RandomForest in Python

- Incorporated **AWS Athena, AWS Glue, AWS S3, boto, AWS QuickSight, SQL, Python** to interpret customer data from a bank
- Predicted customer attrition using **machine learning RandomForest models** and **AdaBoost** to optimize accuracy by **20%**
- Developed model selection process using **CrossValidation** and **GridSearch**, leading to adoption of **RandomForest**

Data Structures and Algorithms with Python

- Programmed **LinkedLists, Hash Tables, Trees, and Queues** and utilized **algorithms**: backtracking, BFS, DFS, recursion
- Employed **Binary Search** to improve algorithm performance by **25%**

Unsupervised Machine Learning in Python

- Developed **custom KMeans clustering** algorithm for 20NewsGroups dataset
- Calculated **Gini index, Purity score, objective score** for KMeans
- Implemented **DBSCAN clustering class** on MNIST and housing data
- Coded and finetuned **custom PCA** on image data

Web Development

- Created interactive shopping website using **JavaScript, PHP, MySQL** to display information from custom **MySQL database**
- Designed **normalized relational MySQL database** for interactive site with **DDL keywords: CREATE, INSERT, PRIMARY KEY, FOREIGN KEY**