

# Harnish Shah

Data Scientist / Data Analyst / ML Engineer  
(732)351-3241 | [harnishshah25@gmail.com](mailto:harnishshah25@gmail.com) | [LinkedIn](#) | [Github](#)

## **SUMMARY**

I am an aspiring data scientist with 9+ years of experience as a biomedical engineer, where I designed, tested, and improved innovative and affordable medical devices that can enhance the health of communities worldwide. My previous experience working with statistical models and machine learning helped me discover that I love jumping into problems to help drive businesses decisions or find insight into what may look like messy unstructured data. I also developed strong communication skills alongside learning, investigation, and collaboration. I look for opportunities to leverage my knowledge of statistics and machine learning to boost biomedical device efficiency and revenue.

## **SKILLS**

Languages: Python | SQL | R | Scala | Spark

Technologies: Github | Databricks | Tableau | Google Suite | VS Code | Command Line | AWS EC2 | Flask | Jupyter Lab

Skills: Web Scraping | Data Cleaning | Linear Models | Tree Models | Support Vector Machines | Predictive Modeling | Time Series / Seasonality Analysis | Neural Nets (DNN, RNN, CNN) | Natural Language Processing (NLP) | Principal Component Analysis (PCA) | Pipelines | Data Visualization | Data Analysis | Data Management |

Python libraries: Pandas | NumPy | Matplotlib | Seaborn | Scikit-learn | Natural Language Toolkit (NLTK) | Beautiful Soup | TensorFlow / Keras | PySpark

## **DATA SCIENCE PROJECTS**

### **Recommender System, Traditional Retail Company**

Jupyter Lab, Tableau

- Audited, validated, and scrubbed 10M + rows of sales data and performed the RFM analysis
- Utilized supervised models to get the accuracy of over 97% taking users and the products they purchased
- Built a Recommender System based on user and the products purchased
- Performed seasonality analysis to predict the upcoming sales using the system prepared

### **Classification model with NLP to Sentiments of two subreddits**

Reddit PRAW API, Jupyter Lab

- Created an unstructured dataset by scraping over 12,000 posts from both subreddits
- Transformed data in preparation for modeling using feature engineering and NLP
- Identified topics that were strongly associated with subreddits using a logistic regression classifier
- Achieved a recall score of .6944 to address the class imbalance in the dataset

### **Hotdog or Not Hotdog**

Jupyter lab, Streamlit

- Built and trained a convolutional neural network on ~ 4,200 image files in 6 hours
- Co-created a Streamlit app for the CNN that can accurately classify an uploaded picture as a hotdog or not

### **Successful Song Prediction with Features**

Jupyter Lab, Github

- Cleaned and analyzed 250,000+ rows of Spotify data and built Random Forest, DNN, Logistic Regression, and Decision Tree Classifier models
- Trained a dense neural net model with 79% accuracy in predicting popular Spotify songs
- Collaborated with 3 data scientists over Github for a 20-minute stakeholder presentation

## **Regression model to predict home sale prices in Ames, IA**

- Explored a Kaggle dataset of home sales from 2006 to 2010 provided by the Ames's Assessor's office
- Used data visualization and feature engineering techniques to analyze and manipulate data into a format usable for modeling
- Evaluated several models using linear regression and LASSO or Ridge regularization
- Developed a model that was able to predict housing prices with an R2 score of .97
- Stood 1<sup>st</sup> of all model submissions on a Kaggle competition

## **PROFESSIONAL EXPERIENCE**

### **Soma Technology Intl**

May 2018 – Jan 2023

*Biomedical Engineer, Bloomfield, CT*

- Analyze large amounts of data and develop statistical models to find patterns and solve problems that will help drive strategic business decisions.
- Design and implementation of machine learning solutions using regression, model, clustering (KNN, K-means, Bayesian, Mean-shift), statistical profiling, inference, classification, and predictive analysis.
- Working knowledge of medical device and combination product regulations and standards such as 21 CFR Part 4, 21 CFR 820, ISO 13485, and ISO 14971
- Experience with change controls, deviations, CAPA Experience with Medical Device or Combination products.
- Understanding of the six-sigma process, applying statistical tools to solve statistical problems.

## **EDUCATION**

### **General Assembly, NYC, USA – Remote**

Jan 2024 - April 2024

*Certificate of Completion, Data Science Immersive*

### **Saurashtra University, Gujarat, India**

April 2007 – April 2012

*Bachelor of Engineering, Biomedical and Instrumentation Engineering*

## **CERTIFICATES**

### **Databricks**

*Certificate of Completion, [Databricks Fundamentals](#)*

### **Databricks**

*Certificate of Completion, [Databricks Lakehouse Fundamentals](#)*

### **Databricks**

*Certificate of Completion, [Generative AI Fundamentals](#)*