Tushar Sharma

tusharmahalya@gmail.com +91 7652064884 tushar-5harma.netlify.app linkedin.com/in/tushar-5harma
github.com/tushar-mahalya kaggle.com/tushar5harma

Data Scientist with expertise in gathering, cleaning, and organizing data for both technical and non-technical stakeholders. Proficient in advanced statistical, algebraic, and analytical techniques, demonstrating a comprehensive understanding of data analysis. Highly organized, motivated, and diligent, leveraging a strong background in analytics to drive insightful and actionable results. Committed to applying my skills and knowledge to contribute to the success of a company's mission.

PROFESSIONAL EXPERIENCE

Data Analyst - Intern Nov 2022 – Jun 2023 | Kanpur Nagar, India
QualDigiIn Technologies® Pvt. Ltd.

- Managed and optimized **SQL databases** for **25+** clients, ensuring efficient data storage and retrieval.
- Gathered, cleaned, and organized data using **MS Excel** extensively, ensuring data integrity and facilitating analysis.
- Utilized **Tableau** to create dynamic dashboards connected to SQL databases, providing live feeds and insights for multiple clients.
- Led **5 analytical projects**, extracting valuable insights and identifying patterns from primary data sources, resulting in enhanced business strategies and increased revenues for clients.
- Implemented a **machine learning**-based recommendation engine for an e-commerce website, achieving **5%** (estimate) projected sales growth and improving user experience.

Data Science & Analytics - Virtual Intern Jul 2022 – Sep 2022 | Remote (Online)
Forage® Pvt. Ltd.

- Prepared insightful report and strategic plan for 'Zilinka' at **Quantium**, utilizing PowerPoint for data visualizations and recommendations.
- Presented findings to 'Social Buzz' at **Accenture**, leveraging Python, SQL, and Tableau to increase engagement by 20% and improve content conversion rates by 15%.
- Analyzed natural gas prices for **JP Morgan Chase & Co.**, applying time series analysis, pricing model development, machine learning, and quantization techniques.
- Identified key churn factors for 'PowerCo' in **BCG**, achieving 85% accuracy and potential $10K cost savings using Random Forest model.
- Analyzed supply chain data for 'Gala Groceries' in **Cognizant**, providing stock optimization recommendations and developing a strategic plan using sales and sensor data. Created business-friendly PowerPoint slides and a production-ready Python module for the machine learning algorithm.

EDUCATION

Indian Institute of Technology, Madras
BS in Data Science and Applications
2021 – present | Chennai, Tamil Nadu

Dayanand Anglo-Vedic College
B.Sc in Electronics
2018 – 2021 | Kanpur, India

Gulmohar Public School
Senior High School
2017 – 2018 | Kanpur, India

Gulmohar Public School
High School
2015 – 2016 | Kanpur, India

ACCOLADES

- Achieved first rank individually in a private Kaggle competition conducted by Univ.AI (AI2 Cohort 4), outperforming 21 teams and 32 competitors with the highest accuracy on the Kannad MNIST dataset using a simple ANN model with regularization.
- Received the "Academic Excellence" award from O.P. Verma, District and Session Judge, Kanpur, for achieving a perfect 10 CGPA in high school.

SKILLS

Techinical Stack
Python, SQL Database (MySQL), HTML, Conda, Version Control Systems (Git), CI/CD Pipeline (Github Actions), Natural Language Processing (NLP), Time-Series Forecasting & Feature Engineering

Platforms & Tools
Windows, Linux (Debian based), Microsoft Office, Tableau, AWS Sagemaker Studio Lab, Microsoft Azure & JupyterLab

Data Science Algorithms

- **Supervised Learning** - Linear Regression, Logistic Regression, k-NN, Decision Trees, Random Forests, Boosting (Gradient, AdaBoost, XGBoost), SVM & Naive Bayes
- **Unsupervised Learning** - Clustering (k-means, Hierarchical, Agglomerative) & Principal Component Analysis (PCA)
- **Deep Learning** - ANN, CNN, RNN, LSTM, GRU, GAN, Transformers & Autoencoders
- **SOTA** (**State-of-the-Art**) **Algorithms** - YOLO, StackGAN, BERT, SAM, ResNet, EfficientNet, VGGNet & GPT

Libraries & Frameworks
pandas, matplotlib, seaborn, scikit-learn, TensorFlow, Keras, spaCy, NLTK, Streamlit, Flask, FastAPI, BeautifulSoup, LangChain

INTERESTS & HOBBIES

Coding • Reading Books • Indian Mythology
Equity & Derivative Trading • Artificial Intelligence

PROJECTS

Spotify Song Recommender System
github.com/tushar-mahalya/Songs-Recommender-System

Aim: *To develop a web application that emulates the UI of Spotify, integrating a robust recommender system that utilize the audio features of selected songs to provide personalized recommendations and **Analytical Engine** capable of analyzing songs & genres.*

- Scraped the list of songs in **Billboards Hot 100** Charts spanning from 1946 to 2022 (77 **years**) using **BeautifulSoup** and fetched metadata, including audio features, for each song using the **Spotify Web API**, creating dataset with over **6500+** songs.
- Preprocessed relevant features using **TF-IDF**, **TextBlob**, and **OHE** techniques to generate a Song Summarization Vector, and employed **Cosine Similarity** to recommend songs from the selected songs list.
- Conducted thorough **Exploratory Data Analysis (EDA)** on the collected data and developed an interactive **Tableau** dashboard.
- Web Application was successfully deployed on **Streamlit Cloud**, ensuring easy access and smooth usage for all users.

Custom ChatGPT
github.com/tushar-mahalya/Custom-ChatGPT

Aim: *Develop a customized chatbot integrated with the **GPT-3 model**, trained and fine-tuned on a custom text corpus. This tailored chatbot will generate contextually appropriate, human-like responses closely aligned with the training data.*

- Collected comments from the top **1000** posts of the three leading data science communities on Reddit using the official **Reddit API**. This resulted in a text corpus of approximately **12 million** words from around **223k** comments.
- Performed **EDA** using **NLTK** & **spAcy**, and pre-trained **Hugging Face** models for **Sentiment and Emotion Analysis**.
- Created indexes of GPT-3 model embeddings using the **LangChain** and **FAISS** framework to achieved contextually appropriate responses to user queries/prompts.
- Developed a simulation web application using **Streamlit** for a user-friendly chatbot experience.

Krishi: Detect Plant Disease with Ease
github.com/tushar-mahalya/Krishi

Aim: *To Develop a centralized platform for farmers to efficiently detect plant health issues by creating a disease classifier capable of accurately identifying **21 diseases** across 9 different plant species, resulting in a total of **30 distinct categories**.*

- *Utilized the 'Plant Village' dataset consisting of approximately **67k images** of diseased and healthy leaves from 9 plant species and* conducted **EDA** to gain insights into the dataset.
- Developed an optimized **CNN architecture** specifically tailored for plant disease classification, achieving a validation accuracy of **+95%** for all 9 plant species. Visualized the training history of all CNN models to get overview of their performance.
- Designed a responsive frontend using **HTML**, **CSS**, and **JavaScript**, integrated with a **Flask framework** backend.
- Deployed the application on **MS Azure**, leveraging a **CI/CD** pipeline (**GitHub Actions**) for integration and deployment updates.

Racial Bias Detection
github.com/tushar-mahalya/Racial-Bias-Detection

Aim: *The objective of this study was to investigate potential disparities between African-American and Caucasian defendants by analyzing the data used by the **COMPAS** software of Equivant, an American tech company specializing in law enforcement.*

- A fine-tuned **Logistic Reg.** & **Decision Tree** model with accuracy of ~70% was built to predict recidivism and draw inferences.
- **ROC curves** were generated for both models to showcase their ability to discern high-risk individuals, highlighting the balance between True-Positive rates (**TPR**) and False-Positive rates (**FPR**).
- The findings from both models revealed discrepancies in the data. African-Americans had a **1.3x** higher likelihood of being inaccurately identified as high-risk, whereas Caucasians had a **1.6x** higher probability of being incorrectly labeled as low-risk.

My GitHub profile features additional projects centered around various areas of interest such as Forecasting, Predictive Statistical Modeling, Sentiment Analysis, and the implementation of various state-of-the-art (SOTA) models.

COURSES & CERTIFICATIONS

Foundational Level Certificate | Data Science
Indian Institute of Technology Madras | 2022-23

Machine Learning | Andrew NG
Stanford University | Coursera
Aug 2022 – Oct 2022

AI-2 : Convolutional Neural Networks
Univ.AI
Jun 2022 – Jul 2022

IBM Data Science
IBM | Coursera
Oct 2022 – Dec 2022

Google Data Analytics
Google | Coursera
May 2021 – Sep 2021

VOLUNTEERSHIP

Academic Societies
Volunteer
Jul 2022 – present

Active participant in the Apostrophe Oratorary Club, Corbett House #206 & Ramanunjan Society for Research at IIT-M, actively participating in group activities and fostering a sense of community.

Volunteer for Cause (VFC) India
Scribe
May 2022

Served as a scribe for 4 visually impaired students in the Pen Pals Programme, providing support during their academic exams.