Shubham Pramod Sangole

sangoleshubham20@gmail.com | Portfolio | GitHub | LinkedIn | +91 720 864 4023

PROFESSIONAL SUMMARY

A Data Scientist with 1.5 years of relevant experience, specializing in leveraging advanced analytics and machine learning techniques to extract meaningful insights from complex datasets. Proficient in designing and implementing end-to-end machine learning solutions, from data acquisition to model building, process optimization, and deploying models on cloud services for widespread accessibility.

SKILLS

Languages: Python, SQL, HTML, CSS
Technologies and Tools: AWS, EC2, GCP, NumPy, Pandas, Scikit-learn, NLTK, Flask, Beautiful Soup, Docker, MySQL, TensorFlow, Keras, OpenCV, PyTorch
Miscellaneous: Deep Learning, Optimization, NLP, Recommender Systems, Matrix Factorization, Data Visualization, Feature Engineering, Hyperparameter Optimization, Web Scraping, Time-series Analysis and Forecasting

PROJECT EXPERIENCE

Autonomous Driving | Python, NumPy, SciPy, Matplotlib, TensorFlow, Keras, OpenCV **March 2024**
GitHub Repository | Presentation Video | Deployment

- Ideated an Autonomous Driving System using a Convolutional Neural Network (CNN) trained on a dataset of 43,000 images. Leveraged OpenCV for Image Processing and conducted Exploratory Data Analysis (EDA) to extract valuable insights from steering angles.
- Utilized Rectified Linear Unit (ReLU) activation and employed Kaiming Initialization for weight initialization. Integrated Data Normalization techniques like batch normalization and a dropout rate of 0.2 to augment the model's overall stability and performance.
- Optimized model convergence by implementing Stochastic Gradient Descent (SGD) and Adam Optimizer, resulting in a reduction of the training loss from 6.4698 to 0.1350.

Human Activity Recognition | Python, NumPy, Pandas, Seaborn, Scikit-learn, Streamlit **January 2024**
GitHub Repository | Presentation Video | Deployment

- Devised a Human Activity Recognition system, utilizing gyroscope and accelerometer data collected from the smartphones of 30 individuals. Evaluated time-series data with 561 features, employing fixed window widths of 2.56 seconds with a 50% overlap, to classify human activities into one of the six predefined categories.
- Performed Exploratory Data Analysis on 10k motion sensor readings that helped identify and categorize stationary and moving activities.
- Implemented Logistic Regression and hyperparameter tuning techniques, such as grid search and random search, to enhance the F1-score from 89% to 96%, resulting in an 8% increase.

Forecasting Taxi Demands in Manhattan | Python, NumPy, Dask, Graphviz, Folium, Scikit-learn, Beautiful Soup **December 2023**
GitHub Repository | Presentation Video | Deployment

- Designed a web application to optimize the workday of a taxi driver by calculating the best pickups based on location and time in high-demand areas of Manhattan, resulting in an 8% increase in monthly earnings.
- Applied K-Means clustering algorithm to segment New York City into 40 clusters. Used dask to load 146 million data points, enhancing operational efficiency. Executed time binning to organize the data efficiently into 10-minute intervals for improved analysis.
- Designed Linear Regression and fine-tuned hyperparameters, reducing Mean Absolute Percentage Error by 27% from 0.1821 to 0.1335.

Personalized Cancer Diagnosis | Python, NumPy, Pandas, Scikit-learn, Matplotlib, Regular Expressions **October 2023**
GitHub Repository | Presentation Video | Deployment

- Developed a web application to classify genetic mutations into 9 different classes based on evidence extracted from clinical literature.
- Analyzed individual features through univariate analysis and applied TF-IDF vectorization to generate 100-dimensional word embeddings.
- Restructured class balancing via oversampling and implemented Logistic Regression to achieve a 59% reduction in log loss, decreasing from 2.4659 to 0.9862.

Quora Duplicate Questions Pair Identification | Python, MySQL, NumPy, Pandas, Scikit-learn, NLTK **September 2023**
GitHub Repository | Presentation Video | Deployment

- Built a Machine Learning web application to empower users in identifying duplicate questions on Quora.
- Executed SQL queries to extract 0.4 million rows from a database file and stored the data in an SQL table. Enhanced Natural Language Processing (NLP) using TF-IDF weighted Word2Vec and integrated Fuzzy features to achieve robust semantic understanding and vector representations of words.
- Enhanced the performance of a baseline random model by implementing XGBoost and reduced the log loss by 60% from 0.8854 to 0.3509.

CERTIFICATIONS AND AWARDS

- **Collaborated in co-authoring a paper that achieved successful publication in IEEE Xplore.**
- Applied Artificial Intelligence (AI) / Machine Learning Course on Applied Roots
- Complete Python Mastery on Code with Mosh

EDUCATION

Ramrao Adik Institute of Technology, Navi Mumbai, India **June 2019**
B.E. (Bachelor of Engineering) in Information Technology
Relevant Coursework: Object Oriented Programming, Linear Algebra, Vector Algebra, Differential Equations, Image Processing.