Muhammad Mujtaba Mir

(812) 606-9565
mir.m.mujtaba7827@gmail.com
github.com/mirmmujtaba
linkedin.com/in/mirmmujtaba

Education

Master of Science in Data Science, *Indiana University - Bloomington* — May 2023
Post Graduate Diploma in Data Science, *International Institute of Information Technology - Bangalore* — September 2020
Bachelor of Technology in Electronics and Communication, *Islamic University of Science and Technology* — September 2017

Skills

Programming Languages	Python, SQL, R
Database	MySQL, PostgreSQL, MongoDB
Tools	Numpy, Pandas, Scikit Learn, Matplotlib, Seaborn, Scipy, Keras, Tensorflow, OpenCV, RStudio, PySpark, NLTK, Statsmodel, Generalized Linear Models, VSCode, GitHub, Microsoft Office, Neural Networks, A/B Testing, Tableau, Power BI
Courses	Data Visualization, Statistics, Machine Learning (k-NN, SVM, Decision Forests, Naive Bayes, NGBoost etc.), Advanced Database Concepts (PostgreSQL), Data Mining, Applied Algorithms, Social Media Mining, Intelligent Systems (Linear Algebra, Probability, Artificial Neural Networks etc.), Scientific Visualization

Professional Experience

DATA SCIENCE INTERN — **May 2022 — August 2022**
Blue Cross Blue Shield of Michigan — *Detroit, Michigan*

- Implemented an end-to-end business critical regression based predictive model using state-of-the-art machine learning on a large dataset having around **1 million** data points.
- Wrote Spark SQL queries, performed exploratory data analysis, data cleaning using PySpark, feature selection and engineering.
- Experimented on decision trees, boosting algorithms etc. for baseline.
- Trained a deep learning model to predict distributions, wrote the code to automate hyper-parameter tuning, **improving the performance compared to the baseline** with respect to **2** KPIs.
- Presented the solution to the stakeholders by creating visualizations and wrote a detailed report about the project.

Projects

FACEBOOK FRIEND RECOMMENDATION (click here)
Technologies used: *Python, Numpy, Pandas, Matplotlib, Seaborn, Scikit-Learn, XGBoost, Networkx, Jupyter Notebook*

- Built a machine learning model to provide relevant friend recommendations to Facebook users.
- Data is a directed graph having **1,862,220** nodes obtained from Meta's recruiting challenge.
- Used the Networkx library to engineer features like jaccard similarity, cosine similarity, shortest path etc.
- Applied boosting algorithm. Achieved the accuracy score of **0.98** and **0.97** on train and test data respectively.

QUORA QUESTION PAIR SIMILARITY (click here)
Technologies used: *Python, NumPy, Pandas, Matplotlib, Seaborn, Scikit-Learn, Keras, Tensorflow, LSTM, Decision Trees, Jupyter*

- Trained a multi-input LSTM based neural network to predict whether a pair of questions is similar or not.
- Data consisted of **404,290** pairs of questions obtained from Quora.
- Performed data cleaning, feature engineering and converted the text data into vectors.
- Achieved an AUC score of **0.9160** and **0.9124** as compared to the baseline AUC of **0.77** and **0.75** on train and test data respectively.

IMAGE SEGMENTATION USING U-NET (click here)
Technologies used: *Python, Keras, Tensorflow, PIL, OS, NumPy, Pandas, OpenCV, segmentation models, UNet, CNN, Computer Vision*

- Trained a UNet model to segment street-traffic images into **21 classes**.
- Unstructured data consisted of **4008** images along with JSON files containing labels and the list of vertices of different classes.
- Prepared segmented images for output using PIL library based on information in JSON files.
- Used data augmentation techniques like flip, sharpen etc. Achieved IOU score of **0.43** and **0.42** on train and test data respectively.

CLUSTERING OF COUNTRIES (click here)
Technologies used: *Python, Scikit-Learn, Pandas, KMeans, Matplotlib, NumPy, Pandas, Seaborn, PCA, Unsupervised techniques*

- Applied machine learning techniques to cluster the countries to find the cluster which is most in need of aid.
- Performed data visualization, outlier analysis, feature scaling, dimensionality reduction techniques.
- Used KMeans to cluster the data and identify the developed, developing and under-developed countries.
- Suggested **5** countries most in need of aid.

Achievements And Extra-Curricular

Runner-up in the Data Science Hackathon organized by Indiana University Data Science club on a NLP based project — **Spring 2022**