Debargha Ghosh

ghoshdebargha2524@gmail.com | 6822461650 | | linkedin.com/in/DGutd/ | github.com/debargha12

SUMMARY

Data Scientist with 6 months experience in Exploratory Data Analysis, Feature Engineering, Machine Learning Model Development, A/B Testing and Model Deployment in a production environment. Self-driven with the ability to quickly learn new domains.

EXPERIENCE

Data Scientist, ***The University of Texas at Dallas*** *May 2020 - Aug 2020*

- Built a **data pipeline** to **model, analyze and visualize** COVID-19 data across 12 counties of Dallas
- Implemented a **geolocation SIR** model proving that mitigation fared better than suppression for **80**% of the cases
- **Collaborated** with the development team to provide **actionable insights** aimed at improving user interaction

Data Analyst Intern, ***Cyber Security Knowledge Sharing & Research Counci***l *Aug 2017 - Oct 2017*

- Performed **network traffic analysis** using raw packet data with **Python reducing** congestion by **5**%
- Analyzed and created a **data visualization** of DNS query logs using **R** for 100% of the incoming queries
- Developed an **anomaly detection pipeline** for firewall data logs using Python **improving** performance by **10**%
- Translated the findings into **accessible visuals** for effectively **communicating complex analysis** to non-experts

PROJECTS

Emotion Detection of Reddit posts using NLP

- Performed **data analysis** using **Python** on Reddit posts to reveal information beyond the formal modeling
- Developed a predictive model using **LSTM** from **Tensorflow.Keras** predicting multi-label emotions with **56**% accuracy
- Determined the best hyperparameters using **GridSearchCV** improving the accuracy of emotion detection by **2**%

Sentiment Analysis of COVID-19 tweets using BigData Technologies

- Built an end-to-end big data solution for **sentiment analysis** on live tweets using **Python, Kafka, Kibana, Spark**
- Analyzed the combined score of **Vader algorithm** to gauge positivity, negativity and neutrality
- Deployed the **web-based visual dashboard** using **Kibana** to show the **distribution of the sentiments**

Search Engine for Cricket using Information Retrieval Techniques

- Collected data of webpages by scrapping 100,000 web pages using Apache Nutch & Solr to build web graphs
- Implemented **clustering** using **Python** and **Scikit-learn** library improving search results by **15**%
- Deployed the search engine by creating **REST APIs** with **Flask** and collaborated with the design team for testing

Character Level Name Generation Model using ML Techniques

- Created our own **Recurrent Neural Network** library using **Python** to correctly generate names of people.
- Implemented gradient clipping and sampled prediction at each step **reducing** the training time by **20**%

Analysis of Car Evaluation Dataset using ML Techniques

- Implemented an **end-to-end ML pipeline** using **Python** and **Scikit-learn** library to predict **evaluation of cars**
- Performed **exploratory data analysis** and **feature engineering** to prepare the dataset for modeling
- Determined the best **supervised algorithm** by creating a data pipeline and using RandomSearchCV on alogrithms like **SVM, KNN, Decision Tree, Random Forest Classifier, AdaBoost Classifier, Gradient Boosting Classifier**

AI Searching Techniques

- Implemented Uninformed Search Strategies like **Breadth-first search, Uniform-cost search, Depth-first search, Depth-limited search, Iterative deepening search** and Informed Search Strategies like **Greedy Best-first search, A* search and RBFS** using **Python** to find the s**hortest path from Seattle to Dallas**

EDUCATION

Master of Science (M.S.), *Computer Science* *Jan 2019 - Dec 2020*
The University of Texas at Dallas GPA: 3.85
Bachelor of Technology (B.Tech.), Computer Science *Aug 2014 - Aug 2018*
West Bengal University of Technology GPA: 3.54

SKILLS

Programing Languages: Python, R, Java
Big Data Ecosystem: Hadoop, Spark, Kafka, Elastic Search, Kibana
Database & Cloud: SQL, NoSQL, MySQL, MongoDB, Firebase, AWS
Data Science Libraries: NumPy, Pandas, Seaborn, Matplotlib, SciPy, scikit-learn, Tensorflow, PyTorch, Keras, LightGBM, Plotly, ggplot
Tools: MS Word, Excel, PowerPoint, Jupyter, Tableau, R Markdown, PyCharm, Git, Docker, Kubernetes