AREEBAH SUHAIL

DATA SCIENTIST

Contact: +91 9718433064
Email: areebahsuhail4@gmail.com
Website: itsareebah.github.io
GitHub: https://github.com/itsareebah
Address: New Delhi, Delhi

Data-driven Computer Science enthusiast with a passion for data science and machine learning, possessing notable communication and teamwork abilities. Aspiring to secure an internship, combining technical expertise with effective communication to unravel insights and foster innovation. Eager to learn from industry experts while contributing fresh perspectives.

Education

06/2022 – 06/2024	**MCA,** Masters in Computer Application, ***University of Delhi***	Awaited
01/2021 – 03/2025	**B.S**. Data Science & Application**, *IIT Madras***	8.14/10.0
07/2019 – 06/2022	**B.Sc.** (Mathematics) Honors, ***JMI, Jamia Millia Islamia***	8.87/10.0
03/2018 – 03/2019	XII Standard (CBSE), ***HPS, Hamdard Public School***	86.75%
03/2016 – 03/2017	X Standard (CBSE), ***HPS, Hamdard Public School***	9.6/10.0

Skills

- Data Science and Analysis: NumPy, Pandas, Matplotlib, Excel, Scikit –Learn, TensorFlow, Keras.
- Programming Languages: Python (advanced), Java (intermediate), C/C++ (intermediate), R (Beginner)
- Studying currently – Neural Networks, NLP, AI, Deep Learning
- Operating Systems: Linux command line
- Data Structures and Algorithms
- Mathematical Foundation: Linear Algebra, Probability and Statistics, Calculus
- Machine Learning: ML algorithms, Neural Network

Projects

IPL Innings Score prediction using Power Play – ALL INDIA HACKATHON IITM *April 2023 - May 2023*

- Built a machine learning algorithm to predict IPL scores trained on the past 20 years of data, achieving a rank of 39 in the hackathon and ending up in the top 50 All India.
- Fine-tuned the algorithm by reducing MAE by up to 30% and improving the R2 score through refining the estimators and using feature engineering by KMeans Clustering to cluster similar players and utilizing cluster statistics of players.
- Improved the model training time to 15 seconds by employing ridge regression with gradient boosting estimator, allowing for quick and accurate response.

ML Classification model for Ecommerce buyer Intention (Course Capstone) *January 2023 - March 2023*

- Developed a highly accurate classification model for the Online Shoppers Intention Classification competition on Kaggle, achieving a precision rate of 76% and an F1 score of 0.76, and a final rank of 14.
- Examined the dataset using multiple EDA techniques such as analyzing feature spread and skewness in numerical features. Discovered intriguing patterns within the data samples.
- Implemented feature engineering by creating new variables based on user behavior patterns and seasonal purchases, resulting in a 10% improvement in model performance surpassing the 0.62 baseline model
- Optimized the model's hyperparameters through grid search and cross-validation techniques, achieving a final accuracy of 80% on the validation set and 76% on unseen test data.
- **Tech Stack**: T-Sne plot for pattern recognition, Isolation Forest, SMOTE, AdaBoost, XGBoost, and Random Forest.

Data Analysis of Road Accidents in India (2019) using R *January 2023 - April 2023*

- Led comprehensive data analysis on road accidents in India (2019) using R and key libraries like ggplot2.
- Created diverse visualizations (scatter plots, line charts, bar graphs) to explore factors affecting road accidents, including age groups most vulnerable, spread of accidents across the country, and regression for time series accident data.
- Identified significant correlations and trends through regression analysis, providing valuable insights into road safety dynamics.
- Collected and extracted relevant variables from multiple data files sourced from data.gov.in for in-depth analysis.
- Collaborated effectively with team members, managing, and analyzing a substantial dataset of over 10k records.

Business Data Analysis – Retail Business *January 2022 - April 2022*

- Collected and analyzed 2-year's raw business data, identifying key trends and patterns to drive strategic decision-making.
- Derived comprehensive data cleaning procedures, resulting in higher data accuracy and reliability for future analysis.
- Achieved better understanding of customers' behaviors and delivered practical advice to boost profit margins up to 5% and cut losses in the future.
- **Tech Stack**: Python, SQL, HTML, and CSS. Python libraries used: SQL Alchemy, Flask, security, DB browser (SQLite), etc.

Workshops and Certificates

- Data Mining: **NPTEL** (January 2023)
- Natural Language Processing Specialization: **Coursera** (In Progress)
- Soft Skill Training Workshop by **Ms. Priya Dua** (Professional Soft Skill Trainer)
- Implementing Machine Learning Techniques by IIT Madras – **Qualified** Test-Based Workshop