Sourish Ghosh

📞 (+91) 9163431542 — ✉ ghoshsourish168@gmail.com — 🔗 linkedin.com/in/sourish-ghosh-49aa451ba/ — 🐙 github.com/Sourish168

Experiences

Sravathi AI Technology Pvt Ltd [Bangalore]   **Nov 2024 - Present**
Associate Data Scientist
- Making a **multimodal Transformer** model for ADME prediction using Chemical and Physical Information in PyTroch.
- Collected data from **EcoTox** for Environmental **Toxicity Prediction** and organized them using Pandas operations.
- Made a **Streamlit app** from CAS ID or Name to Structure Generation of molecules using **Python API Handling** methods.
- Used **Prompt Engineering** and finetuned **Gemma-2** on chemical data from **HuggingFace** & **SFT Trainer** for Causal LLM.
- Collecting data using web-scraping tools and used **pandas** to clean and filter the data for further use in our database.
- In **PK profiling**, collected data using **PostgreSQL**, and performing **ML(Ensemble Methods)** and **DL(GenAI)** modeling.
- In **MOPRP(Molecular Property Prediction)**, fixed the whole ML pipeline for a proper output using **Python** and **Pandas**. And finally, **Optimized the Python codes** in the pipeline which became **40% faster** than earlier for toxicity prediction.
- In **Toxicity Prediction**, I have made a Birds and Bees toxicity **(ML)** model to predict the toxicity level of a molecule. For this regression problem, **feature selection, data cleaning, data pre-processing, and data visualization** are performed.

IonCure Tech Pvt Ltd [Remote]   **Jun 2024 - Jul 2024**
Data Science Intern
- Created a database of **5-6 lakh** contacts by web-scraping Cosmetics, Educational, Manufacturer, and Political websites.
- Raw data was cleaned (removing duplicates and invalid emails) to make a further usable database of **3-4 lakh** contacts.
- Mass mailed to all the collected contacts from the database, approaching "to buy products" using Excel and Python.

Projects

Generating Realistic Images using Generative Model   **Jan 2024 - May 2024**
- Stanford Cars dataset is used for the Denoising Diffusion (DDPM) task on **GPU RTX-2080Ti, 12GB**.
- *Training Batch Size:* **16**, *Steps:* **70000** & FID as metric, *Training Time:* **11h 25min 32s** & *Generation Time:* **1.25s**.
- ADD implemented with BSR(Blind Super Resolution) generating images from LR to HR, **Generation Time: 3-4s**.
- Developed a Streamlit web application for generating HR images from LR images.

Exploring Relationship Between Consumer Price Indexes (CPI)   **Sep 2023 - Dec 2023**
- Finding relations between CPIs based on historical time series data with missing values at the time of COVID19.
- Visualization in RStudio and Analysis of Time-Series Models in Python. Generating DF Distribution for **AR(0)** to **AR(15)**.
- Checked **MA, AR, ARMA, ARIMA, SARIMA, DF & ADF** tests. Unit Root stationary and co-integrated series are found.

Analysis of Churning Customers in a Telecom Company

Data Visualization   *Nov 2022 - Dec 2022*
- Data visualization is done with relevant plots on the customer churn data of a US telecom company in RStudio.
- Finding different types of conclusions about churning customers based on the plots.

Machine Learning   *Sep 2023 - Nov 2023*
- Analysis of different types of Machine Learning algorithms and comparison between those algorithms in Python.
- **Logistic Regression:** 0.5692, **KNN:** 0.5725, **Decision Tree:** 0.5451, **Random Forest:** 0.6153, **Support Vector Machine:** 0.5881, **Multilayer Perceptron:** 0.5891 & **XG Boost:** 0.6205

Skills

- **Languages:** Python, R
- **Databases:** MySQL, PostgreSQL
- **Tools:** MS Excel, MS PowerPoint, MySQL Workbench
- **OS Used:** Windows, Linux, MacOS
- **Python Libraries:** NumPy, Pandas, Scikit Learn, Matplotlib, PyTorch

Relevant Courses

- Python Programming
- Machine Learning
- Deep Learning
- Statistics
- Time Series Analysis
- Computer Vision

Educations

Ramakrishna Mission Vivekananda Educational and Research Institute [Deemed University]   **Aug 2022 - Jul 2024**
Course: MSc. in Big Data Analytics

Chandernagore Government College [under The University of Burdwan]   **Aug 2019 - Jun 2022**
Course: BSc. in Mathematics

Certifications

- Hackerrank Python
- Hackerrank SQL

Achievements & Responsibilities

Placement Representative of RKMVERI CS-BDA Batch 2022-24   *Jan 2024 - Jul 2024*
Achieved the First place in Science of BSc. Degree College   *Jun 2024*
Secretary of RKMVERI's CS Departmental Tech-fest Perceptron 2.0   *Apr 2023 - May 2024*
Qualified GATE 2024 in Data Science and Artificial Intelligence Paper   *Feb 2024*