Sidharth Gupta

☎ +917543898755 • ✉ sidhukumar805140@gmail.com
in linkedin.com/in/gupta-sidharth

Professional Summary

Detail-oriented data enthusiast with a Master's degree in Statistics and a Diploma in Data Science, equipped with a solid foundation in statistical methodologies, machine learning, and data analytics. Proficient in Python, R, and SQL, with extensive hands-on experience in data preprocessing, exploratory data analysis, and predictive modeling. Adept at tackling real-world challenges such as class imbalance, regression, and classification problems through innovative techniques and advanced algorithms. Passionate about deriving actionable insights from data and leveraging AI-driven solutions to address business needs. Strong communicator and a quick learner, eager to contribute to impactful data science or analytics roles.

Projects

Master's Thesis: Regression Models Using Composite Weibull-WTP Model

- Developed an innovative composite statistical model by integrating the Weibull distribution with the tempered Pareto distribution, addressing limitations in modeling heavy-tailed data and providing enhanced flexibility.
- Designed a regression framework tailored for the composite Weibull-WTP model, enabling robust prediction capabilities and superior analysis of real-world datasets.
- Conducted rigorous evaluations on diverse datasets, including financial risk assessment and survival analysis, to showcase the model's adaptability and performance improvements over existing methodologies.
- Incorporated advanced numerical optimization techniques for parameter estimation, ensuring model convergence and accuracy under various conditions.
- Authored a detailed academic report emphasizing theoretical contributions, practical applications, and model limitations, and delivered a professional presentation using LaTeX, demonstrating clear and concise communication of complex statistical concepts.

Machine Learning Project: Classification Model for Bank Telemarketing Campaign Success Prediction

- Developed a comprehensive machine learning pipeline to predict the success of bank telemarketing campaigns.
- Implemented advanced imputation techniques using Random Forests for categorical and mixed-variable features.
- Engineered feature transformations with one-hot encoding, min-max scaling, and custom correlation analysis for mixed data types.
- Addressed class imbalance using techniques like SMOTE, Borderline-SMOTE, and Edited Nearest Neighbors.
- Trained multiple classifiers, including Random Forest, XGBoost, AdaBoost, and Voting Classifier, achieving a macro average F1 score above 0.75.
- Built and evaluated neural network and ensemble models, optimizing hyperparameters using GridSearchCV.
- Deployed a final CatBoost model for full data prediction and generated submission files for competition evaluation.

Data Processing Pipeline: Automated Video-to-Excel Data Extraction and Cleaning

- Designed a pipeline for processing videos, extracting frames, and performing OCR to gather structured data from video content.
- Utilized FFmpeg for frame extraction and Tesseract OCR to extract text data from processed images.
- Developed a Python script for data cleaning, handling noisy text, standardizing units, and managing invalid or duplicate entries.
- Implemented Excel-based data storage, ensuring clean, validated output with detailed logs of dropped or corrected rows.
- Automated batch processing for multiple videos, integrating file handling, and output optimization using shell scripting.
- Enabled scalable and efficient extraction and organization of nutritional data for further analysis.

Education

Indian Institute of Technology Madras
Diploma in Data Science, CGPA: 7.43 (as of last completed semester) *2023 – Present*

Central University of Rajasthan
M.Sc. in Statistics, CGPA: 7.1 *2022 – 2024*

Patna Science College
B.Sc. (Hons) in Statistics, Percentage: 74% *2018 – 2021*

Skills

Programming and Query Languages: Python, R, SQL

Data Science and Machine Learning: Statistics, Regression Models, Hypothesis Testing, Probability Theory, Data Preprocessing, Imputation Techniques, Class Imbalance Handling, Feature Engineering, Model Evaluation, and Optimization

Libraries and Frameworks: Pandas, NumPy, Matplotlib, Seaborn, SciPy, Scikit-learn, Imbalanced-learn, Tesseract OCR, FFmpeg, OpenCV, XGBoost, LightGBM, CatBoost, and PyTorch (basic) , Tensorflow (basic)

Tools and Platforms: Jupyter Notebook, RStudio, MS Excel, MS Word, MS PowerPoint, LaTeX, Kaggle, and OpenPyXL

Languages: English (Fluent), Hindi (Fluent), and Maithili (Native)

Certifications

Diploma in Data Science: IIT Madras (Ongoing)

Coursework includes Predictive Analytics, Machine Learning, Data Visualization, Statistical Modeling (Linear Models, Logistic Regression, Bayesian Statistics), Business Analytics, and Python Programming. Proficient in libraries such as Pandas, NumPy, Scikit-learn, and TensorFlow.

Awards and Achievements

Awarded for Excellence: In the Essay Competition organized by NSSO Ajmer on the topic of Indian Official Statistics and Sustainable Development Goals.

Additional Information

Career Objective: Passionate about using data-driven insights to solve complex real-world problems, with a strong foundation in statistical analysis, machine learning, and predictive modeling. Aiming to contribute to innovative data science solutions by leveraging skills in data analytics, AI, and advanced statistical techniques. Eager to collaborate in dynamic teams, drive business outcomes, and continuously expand expertise in emerging technologies.

Hobbies: Badminton, FPS games, Chess, Reading tech news.