Harish Surapur

Dallas, TX (Open for Relocation) | (469)-714 3543 | hks210001@utdallas.edu| Linkedin | Streamlit | Github

EDUCATION

The University of Texas at Dallas
Master of Science in Information Technology & Management (MIS)
Specialization: Data Science & Analytics
GITAM University – Hyderabad, India
Bachelor of Technology, Computer Science Engineering

PROFESSIONAL EXPERIENCE

Verizon, Dallas, USA (Prodapt) Jul 2023 - Present
Data Scientist

- Implemented an LSTM model for next word prediction in chatbots using Azure Machine Learning, achieving 66.5% accuracy with sequence padding and word embeddings. Collaborated via GitHub for version control.
- Implemented XGBoost and Random Forest models using Spark's MLlib, improving sales forecast accuracy by 5% with F1 score of 0.79.
- Designed automated data pipelines using Azure Data Factory and Apache Airflow, resulting in a ~4.5% reduction in operational workload, improving enhanced data accessibility and reporting efficiency across multiple teams.
- Optimized data pipeline workflows using Azure Data Factory and Kubernetes (AKS) to orchestrate Docker images, achieving a 30% reduction in SQL query execution time for a multi-terabyte booking database.
- Utilized CausalML for propensity score matching to inform targeted marketing strategies, leading to a 15% increase in conversion rates.
- Developed an ARIMA time series model using Databricks and Azure Synapse for inventory forecasting, leveraging big data techniques to analyze large datasets, resulting in a 30% reduction in stockouts and an 8% MAPE.

Prodapt, Dallas, USA Jun 2022 - May 2023
Data Science Intern

- Improved customer retention rates by 10% through PCA for regression analysis of service adoption in SPSS, visualizing results in Power BI for agile sprint reviews.
- Conducted A/B testing on telecom service features, applying hypothesis testing methods such as T-Tests and Chi-Squared Tests, resulting in a 10% reduction in customer churn.
- Created a topic clustering model with Word2Vec and K-means, achieving a silhouette score of 0.85, which led to a 30% improvement in issue resolution rates.
- Processed large sales billing datasets in Oracle SQL Developer with SQL queries using CTEs and Window Functions, visualizing insights in Tableau dashboards to facilitate data-driven decisions.
- Streamlined ETL processes using Azure Data Factory and Azure Data Lake, automating data preparation for billing, facilitating real-time monitoring and improved efficiency.

Cleartrip., India May 2019 - Aug 2021
Data Analyst I

- Utilized R libraries (dplyr, ggplot2) and UNIX commands for data preprocessing on travel booking data, uncovering trends in peak booking times and popular destinations, contributing to a 6% increase in user engagement.
- Performed exploratory data analysis (EDA) using Jupyter Notebook and Pandas, generating insights contributing to 15% rise in CSAT score.
- Created interactive dashboards in Power BI using DAX to visualize KPIs like booking conversion rates, enhancing data-driven decision-making for stakeholders, while coordinating tasks in JIRA within scrum teams for iterative improvements.
- Used Azure Blob Storage for data storage and Azure Synapse Analytics for fast analytics on large datasets, improving accessibility and insights for travel booking analysis. Utilized MongoDB to handle unstructured data, enabling organization and faster access to insights.
- Automated data processes for customer booking analysis by using VLOOKUPs to retrieve data from tables, pivot tables for summarization, and VBA macros for reporting, achieving a 30% reduction in processing time and saving 10 hours per week.

PROJECTS

Real-Time Feedback Sentiment Classifier Using GRU RNN GITHUB STREAMLIT

- Developed a sentiment analysis application using a GRU-based RNN model with over 1M parameters, achieving 92% accuracy in classifying user feedback through advanced text preprocessing with NLTK and GloVe embeddings.

Customer Salary Prediction Using RNN GITHUB STREAMLIT

- Created an RNN salary prediction model with KerasRegressor, utilizing standardization and log transformations for scaling. Achieved a 21% reduction in MAE through PCA, Lasso regularization, and grid search while analyzing historical employee data for trends.

TECHNICAL SKILLS

Certifications: Machine Learning with Scikit-Learn, Google Cloud Certified - Associate Cloud Engineer.
Programming Languages & Databases: Python, R, SQL, UNIX.
Data Science & Machine Learning Tools: Pandas, NumPy, Matplotlib, Seaborn, Statsmodels, Scipy, NLTK, Gensim, CausalML, Spark, Hadoop, PySpark, Hive, Databricks, Jupyter Notebook, Dataiku, Regression, Classification, Clustering, Causal Inference.
Data Analysis Tools: GitHub, Apache Airflow, JIRA, Slack, R Studio, Advanced Excel, VBA Macros, Tableau, MS Power BI, Talend.