

UDIT GOEL

[linkedin.com/in/goeludit7rma/](https://www.linkedin.com/in/goeludit7rma/) . github.com/goeludit . goeludit7rma@gmail.com . (+1)7325589842

EDUCATION

Rutgers University

Masters in Data Science and Statistics

Coursework: Regression and Time Series, Deep Learning, Database Systems, Stat Modelling, Probability, Convex Optimization for Applications, NLP, Statistical Inference, Data Structures and Algorithms

New Brunswick, NJ

August 2023 - Dec 2024 (E)

University of Mumbai (R.A. Podar College of Commerce and Economics)

Bachelor in Commerce

Coursework: Statistical Tech, Business Economics, Financial Mgmt, Marketing Research

Mumbai

June 2016 - April 2019

PROFESSIONAL EXPERIENCE

Bright Power

Data Analyst Intern

New York City, NY

April 2024 - June 2024

- Managed a 20+ million data point database, ensuring data accuracy and consistency using Python and SQL.
- Developed Python scripts for seamless integration of APIs with databases, optimizing data workflows for over 10,000 utilities.
- Automated data flows with Rundeck and SQL, enhancing data integrity and reducing manual effort by 40%.

Healofy

Data Scientist

April 2021- July 2023

- Implemented AutoML NLP-based machine translation transliteration and spell correction model for 8 vernacular languages.
- Built a hybrid filtering content recommendation system using matrix factorization for 7+ million daily users.
- Automated ETL marketing data pipeline using Branch, and Google APIs for scheduled updates hosted on AWS S3.
- Implemented digital media compression algorithms decreasing cost by 10% and load times by 5% using AWS lambda.

Data Analyst

Dec 2019 - April 2021

- Implemented and automated data collection methods on Google Cloud Platform and Metabase.
- Developed web scrapers for a competitive pricing strategy leading to growth in revenue by 15%.
- Reduced script's (data collection) runtime by 60% using multithreading utilizing Python's ThreadPool
- Developed data visualization dashboards for daily KPIs using Google Data Studio, BigQuery connector, and Tableau.
- Devised data collection scripts using third-party APIs and integrated third-party tools to analyze data of 20+ million users.

RELATED EXPERIENCE

NJ Department of Health (Rutgers University)

Research Assistant, School of Public Health and School of Social Work

New Brunswick

February 2024 - Present

- Developed scripts to collect gambling data from Reddit and used BERT for sentiment analysis to interpret user opinions & trends.
- Applied statistical methods, executed scraping techniques for US census data, and researched synthetic control methods.
- Implemented data cleaning and processing techniques using Python/R to aid the research on the effects of lead on pregnancy.

TECHNICAL SKILLS

Languages: Python, R, Java, SQL, HTML, CSS, Javascript

Cloud and Software: Reddis, Postgres, MySQL, BigQuery, Tableau, SageMaker, Agile, Azure, MongoDB, Jira, Snowflake, Postgres

Packages: Pandas, NumPy, SciPy, Requests, Selenium, Matplotlib, Seaborn, TensorFlow, Tweepy, PyTorch, Airflow, SQLite, Langchain

RESEARCH PROJECTS

RAG-based LLM Document Analyzer: (Packages: Langchain, ChromaDB, OpenAI):

[Link](#)

Developed a tool leveraging the OpenAI API to vectorize and embed input, providing precise responses with sources and references and resolving the problem of ghosting in LLMs to an extent.

Energy Consumption Forecasting using LSTM: (Packages: Keras, Sklearn, Matplotlib)

[Link](#)

Utilized Adaptive Moment Estimation within the Long Short-Term Memory framework to forecast energy consumption at 10 epochs. The model yielded an R^2 value of 0.95.

Uber Price Prediction Model (Packages: Statsmodel, Sklearn, Scipy, Seaborn, Matplotlib, Pandas, Numpy)

[Link](#)

Applied XGBoost, Random Forest, and bagging regressor to forecast Uber and Lyft ride prices. Boosted accuracy to 97% using bagging regressor and hyper-tuned the model with the most optimized parameters.

Twitter Database and ETL Pipeline (Packages: Snowflake, MongoDB, AWS RDS, Tweepy)

[Link](#)

Developed a cloud-native storage solution for real-time Twitter data using Snowflake, MongoDB Atlas & Amazon RDS. Managed user and session cache and implemented indexing for faster querying.

Semantic graph embedding for text representation: (Packages: Python-Louvain, PyVis)

[Link](#)

Created a program to establish semantic relationships between words using word embeddings, visualizing an interactive, color-coded semantic graph based on similarity and frequency