OJAS SHARMA

✉ ojassharma16@gmail.com ☎ +1(929)6847733 Linkedin Github Website

EDUCATION

Rutgers University – New Brunswick, NJ – | **MS – Statistics, Data Science | GPA: 3.80** **Sep 2023 - May 2025**

Relevant Coursework: Probability and Stat Inference, Regression and Time Series, Data Structure and Algorithm, Database Management System, Stat Model and Computing, Deep Learning, Financial Time Series, Algorithmic Trading & Portfolio Management

Manipal University Jaipur – Jaipur, IN – | ***B. Tech Computer* Engineering | GPA: 3.61** **Aug 2019- May 2023**

Relevant Coursework: Engineering Mathematics, Data Structures, Data Science with Python, AI/ML, Deep Learning

SKILLS

- **Languages/Framework:** Python, PySpark, SQL, R, C, C++, Beautiful Soup, Selenium, MATLAB, Flask, SreamLit
- **Software:** Microsoft Power BI, Tableau, Microsoft Excel, Git, Databricks, SPSS, BLOOMBERG TERMNINAL
- **Database/Cloud:** NoSQL, PostgreSQL, MySQL, Hadoop, Hive, Snowflake, Apache Spark, AWS, GCP
- **Machine Learning Frameworks and Libraries**: PyTorch, TensorFlow, Keras, NLTK, Scikit-Learn, spaCy, Regex, SciPy, LLMs, RAG, Generative AI, Transformers, LangChain

WORK EXPERIENCE

Data Scientist(Research Assistant) | **Center Of Gambling Studies, Rutgers University** | New Brunswick | **Apr 2024 – Present**

- Currently conducting extensive data analysis on a **4.5 TB dataset** encompassing **13.6 billion records** to identify 25+ predictive variables for problem gambling severity, utilizing PySpark and SQL
- **Architected a scalable machine learning pipeline** leveraging **Isolation Forest, SVM, and XGBoost** with 50+ engineered features for high-risk behavior prediction. Applied advanced hyperparameter optimization techniques, including **GridSearchCV**, achieving a **20%** reduction in false positives and enhancing model reliability for predictive behavioral analysis.
- Optimized code for data extraction and modeling by leveraging Python frameworks like **PySpark and Pandas**, resulting in a 90% increase in data processing speed, reducing processing time from **90 seconds** to **9 seconds**.
- Developing end-to-end machine learning pipelines, streamlining data preprocessing, feature selection, model training, and deployment. Achieved an **AUC of 0.93** in identifying high-risk behaviors with a **25%** improvement in precision-recall over baseline models, optimizing for production-level performance and continuous monitoring.
- Designed machine learning algorithms like **XGBoost and Isolation Forest** to detect **chasing patterns** like loss chasing bet size escalation, prolonged sessions, and riskier bets using behavioral data and anomaly detection, achieving **90%** accuracy in identifying **high-risk behaviors.**

Data Analyst Intern | Omalco Extrusion, New Delhi, India **Aug 2022 – July 2023**

- Improvised business teams' efficiency by 80% to validate each month end data flow from upstream systems by designing automated PowerBI reports; revamped logic of reporting tables reduced ETL jobs run time by **50%**
- Integrated real-time data pipelines for dynamic market analysis and implemented time-series forecasting algorithms like **ARIMA and Prophet models**, driving actionable insights to support data-driven strategies across departments.
- Built and optimized Power BI dashboards leveraging **DAX and Power Query**, improving data processing efficiency by **25%.**
- Conducted statistical analysis using **linear regression** and **hypothesis testing** to uncover key business trends, contributing to a 15% increase in operational efficiency.
- Utilized **Apache Spark's** distributed computing capabilities to parallelize data processing and model training tasks across multiple nodes, reducing computation time and improving scalability for handling large datasets.

ACADEMIC PROJECTS

LLM-Powered Product Recommendations **(Github Code)**

PyTorch, HuggingFace, RAG, Quantized Low-Rank Adaptation, NLP, Prompt Engineering

- Fine-tuned LLMs (Mistral-7B, TinyLlama) for product recommendation using Amazon Review Dataset. Achieved 98%+ accuracy in next-purchase prediction through novel prompting and efficient fine-tuning (LoRA, QLoRA).

Twitter Data Analysis(Search Application for Twitter) **(Github Code)**

- Parallelized Data Processing: Leveraged **joblib's Parallel and Delayed** for multi-core execution, reducing tweet processing time by **60%**, optimizing large-scale data extraction tasks.
- Implemented LRU Caching: Developed an in-memory **LRU cache**, reducing retrieval time **from 5 seconds to 0.4 second** and optimizing database load. Integrated **eviction policies** and **TTL** for maintaining cache freshness and performance.
- Hybrid Data Architecture (PostgreSQL + Couchbase): Built a hybrid solution with PostgreSQL for relational user data and **Couchbase** for schema-less tweets, using **ACID transactions** and **NoSQL** sharding for scalability and data consistency.
- **Integrated Elasticsearch** to index JSON-formatted Twitter data, enabling fast queries with **relevant sorting** based on engagement metrics. Implemented **drill-down features** for exploring tweet metadata and improved query performance by 50% through optimized indexing and partitioning.

Explore Crime Data with the Campus Watch Dash App: Rutgers University–New Brunswick **(Github Code)**

- Implemented dynamic web scaping using **Selenium** to extract daily crime records from SpotCrime.
- Integrated real-time data sources and developed an automated data retrieval pipeline, ensuring the dashboard is continuously updated with the latest crime information.
- Leveraged **Plotly library** to craft interactive charts and geospatial maps, significantly enhancing the user experience with rich data visualizations.
- Designed interactive dashboards and visualizations using PowerBI to monitor and analyze real-time customer churn insights