Yash Soni

Toronto, ON
(647) 853-8424 yashsonig@gmail.com LinkedIn GitHub yashns.me

Education

Wilfrid Laurier University **Expected Graduation: April 2027**
B.Sc in Data Science | *Big Data Concentration* | *Economics Minor* *Waterloo, ON*

Technical Skills

Languages: Python, SQL, VBA, C/C++, R, Java, JavaScript
Developer Tools: Azure Databricks, GitHub, Excel, VS Code, Power BI, Tableau, Streamlit, AWS, Jupyter Notebook, Eclipse
Data Tools & Libraries: Pandas, NumPy, scikit-learn, LangChain, PyTorch, Matplotlib
Concepts: GenAI, LLMs, Prompt Engineering, Data Pipelines, Financial Analytics

Experience

Manulife Financial **September 2025 – December 2025**
Data Analyst Co-Op *Toronto, ON*

- Leveraged **PySpark and SQL in Databricks** to process large-scale financial datasets for Global Wealth and Asset Management, enabling scalable pipelines that supported reporting, analytics, and **LLM-driven risk workflows** across business units.
- Automated key data ingestion and transformation processes using Databricks notebooks and scheduled workflows, **reducing processing time by 30%** and significantly improving data readiness for machine learning and LLM experimentation.
- Designed and implemented **PySpark-based monitoring** and reporting pipelines to assess performance of core operations, enhancing visibility into **data quality, pipeline health**, and execution reliability by 25%.
- Migrated legacy SQL workflows into PySpark to support a **transition from traditional warehouses** toward a flexible, **lakehouse-oriented analytics environment** used for scalable ML experimentation and advanced financial reporting.
- **Collaborated with cross-functional teams** including data engineers, analysts, and product stakeholders to gather requirements and deliver clean, version-controlled datasets used for modeling, reporting, and AI solution development.

Wilfrid Laurier University **September 2024 – April 2025**
Teaching Assistant *Waterloo, ON*

- Supported 200+ students across DATA100 and MA103 by **delivering tailored walkthroughs & debugging code**.
- Provided individualized assistance in lab sessions, helping students **build functions, troubleshoot statistical models, and interpret regression outputs** in real time, while reinforcing problem-solving strategies and core programming concepts.
- **Assessed and graded over 200 assignments** spanning code scripts, math problem sets, and written analysis, ensuring fairness, consistency, and detailed, personalized feedback under tight academic timelines.

RBC **May 2024 – August 2024**
Operations Officer Co-Op *Toronto, ON*

- Processed daily mortgage disbursements **exceeding $2.5M with 97% accuracy** under regulatory timelines.
- Built Excel-based tracking tools to **reduce escalation time** for corrupted records by 20%.
- **Supported reconciliation teams** by flagging data inconsistencies and ensuring audit-ready records.

Projects

GenAI Finance Assistant | *Python, OpenAI API, LangChain, Streamlit, FAISS*

- Developed a **GenAI-powered chatbot** to assist financial analysts in querying internal finance data and terminology, reducing lookup time for documentation and standardized reporting logic.
- Used **LangChain with OpenAI's GPT model** to implement **retrieval-augmented generation (RAG)**, grounding LLM responses in structured CSV-based finance datasets through vector similarity search.
- Built a front-end interface in **Streamlit** for real-time testing and stakeholder feedback, supporting prompt tuning and assistant behavior evaluation across common finance-related queries.
- Engineered prompt templates, applied output filtering techniques, and monitored for hallucinations to ensure model responses remained concise, relevant, and auditable in a finance context.

Credit Card Fraud Detection System | *Python, PyTorch, Pandas, NumPy*

- **Developed a machine learning model** to **identify potentially fraudulent credit card transactions** by analyzing features such as credit score, transaction history, and spending behavior.
- Applied weighted probability distributions to **calculate risk scores** to identify fraudulent transactions.
- Performed data preprocessing, feature engineering, and model evaluation to **improve prediction accuracy.**
- Gained **hands-on experience in fraud analytics**, supervised learning, and real-world financial data handling.

Optimal Workout Case Study | *PySpark, Pandas, NumPy, Matplotlib*

- Engineered a data-driven analytical framework to determine optimal workout splits for achieving a lean, muscular physique, using variables such as height, weight, body fat percentage, and gender.
- Ingested and cleaned raw fitness datasets using **PySpark and pandas**, performing large-scale ETL processes to ensure schema consistency and analytical readiness.
- Conducted exploratory data analysis with **NumPy and Matplotlib** to uncover correlations between workout regimens and body composition outcomes.
- Developed a modular pipeline for reproducible experimentation and future extensions into personalized workout recommendation systems.