PAUL JOE SEBASTIAN

Melbourne | +61 469302801 | LinkedIn | pauljoe.ps38@gmail.com | Full Work Rights In AU

OBJECTIVE

Dedicated and detail-oriented Data Scientist with a strong foundation in statistical analysis, data mining, and advanced analytical techniques. Proven ability to transform complex datasets into actionable insights to support business decision-making processes, enhance operational efficiency, and drive data-driven strategies. Adept at leveraging machine learning, data visualization, and programming skills to solve complex problems and optimize performance.

PROFILE SUMMARY

- **AI Research Intern at ERETS Space:** Spearheaded the development of an application supporting satellite tracking, debris analysis, and collision prediction. Translated complex astrodynamics and AI model outputs into an intuitive user interface and robust backend system.
- **Data Engineer (Part-Time) at Meshynix:** Developed medallion architecture on Azure Databricks for Takachar, a climate tech company, creating data ingestion pipelines using PySpark and building an interactive R Shiny dashboard for data collection and metric reporting.
- **Data Scientist intern at Blackcoffer:** Focused on data engineering, data visualization and building scalable data pipelines. Achieved up to 32% improvement in prediction accuracy in machine tasks by implementing modern machine learning algorithms and techniques.
- **Data Generation and Anomaly Detection:** Developed an advanced data generation and anomaly detection system for industrial applications using Heavy-Tailed Generative Adversarial Neural Networks (GANs), resulting in a 23% improvement in performance over standard approaches. This work earned the Best paper Award at the ERASMUS+ conference.
- **AI-Powered Image Retrieval:** Designed an AI-powered image retrieval system leveraging Approximate Nearest Neighbour (ANN) algorithms and Neural Networks, achieving a 37% reduction in image retrieval times while also improving scalability significantly.
- **Improved real estate pricing predictions:** Applied cutting edge clustering algorithms to segment the dataset, followed by individualized model processing for each cluster, resulting in a 32% improvement in predictions compared to previous implementations.
- **AI Nutritionist Project:** Engineered an AI-driven system using computer vision and object detection algorithms to analyze refrigerator contents and generate recipe suggestions using image processing and cutting edge neural networks.
- **Advanced Data Science Skills:** Proficient in statistical analysis, data mining, predictive modeling, with deep understanding of machine learning frameworks.
- **Technical Proficiency:** Skilled in SQL, data storage solutions and cloud platforms such as GCP and AWS with expertise in BigQuery, Bigtable and AWS EMR for handling large-scale data processing tasks.
- **Programming & Tools:** Strong understanding of programming languages and tools including Python, SQL, R, Tensorflow, Keras and Scikit-learn.
- **Big Data & Pipeline Development:** Expertise in architecturing and implementing scalable data pipelines, data processing frameworks, and big data technologies.
- **Analytical and Problem-Solving abilities:** Possess strong analytical thinking and problem-solving skills, with a focus on optimization computational performance and system efficiency.

KEY SKILLS

✓ Machine Learning / AI	✓ Data Visualization	✓ Python	✓ Tensorflow
✓ Statistical Analysis	✓ Data Modelling	✓ SQL	✓ NLP
✓ Data Processing	✓ Data Mining	✓ Big Data	✓ Pandas

EDUCATION

Master of Data Science, RMIT University, Melbourne	**2023-2024**
Bachelor of Data Science and Analytics, Jain University	**2019-2022**
Databricks Certified Data Engineer Associate	**2024**

PROJECTS

1. Blood Demand Forecasting for Australian Red Cross Lifeblood — **Databricks, Pandas, Scikit-learn, LightGBM, Chronos**

- Implemented **SCINet** and **automated feature** engineering with tsfresh model temporal dependencies and enhance blood demand forecasting accuracy.
- Developed an **end-to-end ML pipeline** using **Chronos** for multi-horizon predictions, capturing seasonal patterns and adapting to new data sources.
- Optimized model performance through hyperparameter tuning and ensemble techniques with **LightGBM**, improving forecast reliability for inventory management.

2. Data Generation Using Heavy Tailed GANs — **TensorFlow, Numpy, Matplotlib, Seaborn, Scikit-Learn**

- **Developed a Generative Adversarial Network (GAN)** specifically designed to model heavy-tailed data distributions.
- Utilized a Pareto GAN approach to handle data with extreme values more effectively than traditional GANs.
- Achieved a **23% improvement in overall performance** compared to standard GAN models.
- Addressed challenges in data generation and anomaly detection by focusing on heavy-tailed distributions.
- Presented findings and methodology in a paper that won the **Best Paper Award at the ERASMUS+ conference**.
- Demonstrated the ability to generate high-quality synthetic data that mirrors real-world heavy-tailed datasets.
- Enhanced the robustness and accuracy of anomaly detection in datasets with extreme outliers.

3. AI-Powered Image Retrieval — **Keras, Streamlit, Annoy, TensorFlow, Scikit-learn**

- Developed an **AI-powered image retrieval system using Artificial Neural Networks (ANN)** and Approximate Nearest Neighbour Algorithms to enhance search capabilities.
- Achieved a **37% reduction in search time** compared to traditional K-Nearest Neighbors (KNN) methods.
- Implemented advanced feature extraction and matching to improve the accuracy and scalability of image searches.
- Designed the system to handle large-scale image databases efficiently, providing rapid and precise search results.
- Demonstrated high scalability and robustness in various search scenarios and image types.

4. Real Estate Price Prediction Model — **GCP, Scikit-learn, Pandas, NumPy, Matplotlib, Seaborn**

- Developed a real estate price prediction model during an internship for a client, addressing issues with a previously implemented SVM regressor.
- Conducted in-depth data analysis, identifying the need for data segmentation to improve prediction accuracy.
- **Implemented clustering techniques to segment data**, allowing for the use of specialized models to each cluster.
- **Improved model performance by 32%** by the application of individual models to clustered data.
- Enhanced prediction accuracy, providing the client with more reliable real estate price predictions.

5. AI Nutritionist — **PyTorch, Flask, OpenCV, YOLO**

- Developed an AI nutritionist that suggests recipes based on photos of fridge contents.
- Utilized **OpenCV** and the **YOLO** algorithm for **advanced image processing and object detection**, enhancing ingredient recognition from images.
- Implemented a recommendation algorithm to generate recipes based on identified ingredients.
- Designed a user-friendly interface for capturing photos and receiving recipe suggestions.
- Enhanced user experience with personalized meal recommendations, reducing food waste.
- Integrated computer vision techniques to accurately interpret and analyse food items in the fridge.

6. Data Visualization Dashboard — **SQL, GCP, Big Table, Big Query, Data Studio**

- Developed a comprehensive dashboard for a client to aggregate and visualize data from multiple sources.
- Leveraged **Google Cloud Platform (GCP), including Bigtable and BigQuery**, for efficient data storage and querying.
- Utilized SQL for extensive data wrangling, transforming raw data into actionable insights.
- Designed and implemented the dashboard using Data Studio, providing an interface for **real-time data visualization**.
- Enhanced decision-making capabilities by delivering a centralized platform for data analysis and reporting.