# SHINOJ PHILIP JOHN

Melbourne, VIC | +61 456 995 237 | shinoj4005@gmail.com | linkedin.com/in/shinoj-philip-john | shinojphilipjohn.github.io

## SKILLS

**Data Science & ML:** Supervised ML (Logistic Regression, Random Forest), EDA, Feature Engineering, Model Evaluation, Predictive Analytics, Statistical Modelling, Experiment Design
**Programming & Tools:** Python, Scikit-learn, NumPy, Pandas, Matplotlib, SQL, Jupyter, Git, R (basic)
**Cloud & Data:** AWS (S3, Lambda, EMR, Redshift), Snowflake, Data Pipelines, ETL/ELT Development, Data Wrangling, Data Cleaning
**Visualization & Storytelling:** Power BI, Tableau, Data Narratives, Communication for Non-technical Stakeholders
**Soft Skills:** Analytical thinking, structured problem-solving, collaboration, consulting communication

## WORK EXPERIENCE

**Data Team Intern, IBISWorld | 2025**
• Delivered Snowflake ELT workflows using AWS S3 external tables, improving ingestion efficiency and reducing unnecessary storage duplication by **~20–25%** across recurring loads.
• Parsed and flattened complex multi-level JSON using **LATERAL FLATTEN**, increasing usable analytical fields by **~35%** and cutting manual preprocessing time by **~30%** through a structured multi-layer data flow (staging → parsing → transformation).
• Enhanced pipeline traceability with structured logging and clear transformation checkpoints, reducing debugging time by **~25%** and supporting faster onboarding through high-level architecture documentation.

**Data Engineer, nbn Australia (via Infosys) | 2021–2023**
• Built and maintained AWS S3 → EMR → Redshift ETL pipelines supporting financial and reporting teams, achieving **~99% successful daily run stability** across batch workloads.
• Optimised SQL transformations and Redshift table structures (sort/distribution keys), improving end-to-end pipeline runtime by **~25–30%** on recurring data loads.
• Developed and managed **50+ SQL scripts** including delta-load logic, DDLs, and control-table updates in MySQL, ensuring alignment with E-R designs and JIRA user stories.
• Reduced manual intervention during executions by **~40%** through parameterising ETL behaviour using YAML/JSON files and automating jobstreams in IBM Workload Scheduler.
• Enhanced data quality and reduced SIT→UAT defect leakage by **~30%** by validating missing data, identifying upstream design gaps, and coordinating fixes with analysts and operations teams.

## PROJECTS

**Customer Churn Prediction (Python, Scikit-learn)**
• Cleaned and preprocessed **7,000+** customer records, handling missing values, encoding categorical fields, and scaling numerical features to prepare data for modelling.
• Conducted EDA to uncover trends in tenure, contract types, and billing patterns, enabling more informed feature selection.
• Improved model accuracy from a baseline of ~**70%** to **~82%** by applying feature engineering, hyperparameter tuning, and cross-validation.

**End-to-End Payroll ETL Framework (Python, Pandas, MySQL)**
• Built a command-line–driven ETL engine that automated extract → DDL creation → transform → load workflows, reducing manual data preparation from **hours to minutes** by integrating CSV and SQL sources into a single, reproducible pipeline.

## EDUCATION

**Master of Analytics (STEM), RMIT University | 2023–2025**
**B.Tech– Electrical & Electronics Engineering, APJ Abdul Kalam Technological University | 2016–2020**