Daniel Gil

Data Scientist and AI/ML Engineer

Location: Melbourne, VIC, Australia

Email: dfgilto@gmail.com - **Phone**: 0490928836 - **Github**: danielgil1 - **Website**: https://danielgil1.github.io/ **LinkedIn**:danielgil1

Summary

Data Scientist and AI/Machine Learning Engineer with over 12 years of experience leading IT data-driven projects across various industries in the US, Europe, and Australia. I hold a Master's degree in Data Science from the University of Melbourne. In my recent roles, I have implemented projects in Generative AI, developed Large Language Model (LLM) AI Agents, and created models for information extraction using NLP and Computer Vision, as well as implemented recommendation systems and predictive models.

Skills: GenAI/LLM (Agents, RAG - Retrieval Augmented Generation, Anthropic, OpenAI, LLama, Langchain), MLOps (AWS SageMaker, Databricks MLFlow), Exploratory Data Analysis (Jupyter Notebooks, Streamlit, Grafana, BI), Predictive Modeling (Recsys, Classification, Regression, Time Series Forecasting, Clustering), Deep Learning/ML Frameworks (PyTorch, Keras, Scikit-learn, Dart Time Series). Data Pipelines (Airflow, Dagster, DBT, PySpark, AWS Glue, Python, SQL), Frameworks (Nvidia Merlin, NVTabular), Infra as Code (Terraform, Ansible), Platforms (AWS, GCP, Databricks)

Career History

Senior Data Scientist at Homes.com.au

Jan 2021 - Feb 2024

My role supports the user engagement through artificial intelligence and data capabilities to improve the search experience and personalisation in the form of data and machine learning end-to-end products, from acquiring the raw data, build data pipelines to ingest into our data lake and build AWS SageMaker pipelines for ML Model process, train and deployment for inference using CICD and IaC terraform to meet Engineering best practices.

I am also a contributor to build the data science practice, mentor junior data scientists and collaborate with product and engineering stakeholders with the following achievements:

- Implemented a RAG (Retrieval Augmented Generation) solution across thousands of daily real estate listings increasing listings suburb coverage by 25% using LLM Anthropic Claude through AWS Bedrock in a event-driven architecture and guardrails deployment.
- Implemented prompt compression technique that reduced LLM costs to 30% compared to baseline.
- Implemented multi-modal content-based recommendation system based on users' shortlisted real estate listings achieving a 0.78(precision@5) utilizing PyTorch and NVIDIA frameworks for GPU optimisations.
- Implemented multi-modal algorithms to extract topics and concepts from real estate listings text, images and spatial data to be used in downstream tasks like search. Deploying models as APIs (ECS or Sagemaker Endpoints) and AWS Batch inference jobs.
- Implemented algorithms to rank real estate images for personalized user experience achieving 0.89(precision@1) using ViT and contrastive learning.
- Implemented LLM Agents POC using the ReAct framework
- Productionized AWS SageMaker ML models to be consumed as APIs or batch inference aligned to data and engineering practices like Infra as code, containerization, model governance, event-driven architectures, big data processing, online and offline inference.
- Built vector embeddings representations that support multimodal search and

recommendations representing items and user interaction profiles while communicating business value to internal stakeholders through clustering and visualization techniques.
- Designed and implemented data pipelines with Airflow, Dagster and AWS Data Lake storage to generate clean and curated datasets for analytics and machine learning Implemented big data transformations and data catalog with pyspark, SQL and DBT models.
- Delivered data visualizations for business stakeholders using tabular, geospatial and time series data.
- Demonstrated AI value with proof of concepts prototyping ML apps using a combination of streamlit, jupyter notebooks and metrics evaluations/visualizations.

Data Scientist / Advanced App Engineering Specialist at Accenture

Sep 2019 - Dec 2020

Developed machine learning model experiments for projects' early cost prediction. Developed machine learning pipeline framework using Databricks MLFlow to track models and deploy to docker containers / generate artifacts for tensorflow js.

- Implemented Natural Language Processing Deep Learning model to extract entities from user feedback.
- Implemented data pipelines with Apache Airflow.
- Implemented data visualization dashboards with Apache Superset. Developed data model and python flask REST API backend for *Estimator app* to predict project cost.
- Elaborated a proof of concept to integrate myWizard Automatic Ticket Resolver (ATR) with Digital Desk chatbot.

Skills: Predictive Model, Transformers, MLOps, Machine Learning Pipelines, Data Pipelines, Pytorch, NLP, Nvidia, Airflow, Dagster, Data Lake, PySpark, DBT, SQL, Streamlit, Terraform

Data Scientist - Capstone Project at SEEK Ltd - University of Melbourne

Mar 2019 - Nov 2019

- Implemented Natural Language Processing Deep Learning model to extract entities from Job Ads (Named Entity Recognition)
- Trained and evaluated a Language Model for contextual word embeddings in 500K jobs ads.
- Evaluated pre-trained word embeddings performance like BERT and GPT-2 Performed Exploratory Data Analysis for 500K+ job ads
- Developed sampling data strategy and performed data cleaning and annotation tasks

Research Community Coordinator, Natural Language Processing (NLP & Data Science) at University of Melbourne

May 2018 - Jul 2019

- Led a community at Research Computing Services with 150+ graduate researchers helping to analyze large text datasets using python and natural language processing. Conducted workshops, trainings and meetups to discuss natural language processing
- Developed multimedia content to engage the research community.

Senior Business Analyst and Product Manager at Globant

Jan 2016 Jan 2018 (2 years)

- Led product definition with a distributed development team across Latin America and Spain for Mortgages and Loans at OpenBank, the digital bank of Santander Group.

- Maintained a roadmap and backlog with Client stakeholders.
- Supported Jhonson & Jhonson Salesforce.com new internal features for Europe and Asia - Pacific (APAC).
- Led the offshore Business Analyst team for Southwest.com with development teams across the United States and Latin America. Analyzed different data sources to maintain a defect backlog to prioritize the work of 20+ developers

Lead Software Architect at Innobile Software
Jan 2013 Jan 2016 (3 years)

- Designed and developed the first on-cloud RFID enabled Warehouse Management System and inventory tracking built in Colombia introducing Software as a Service model.
- Implemented components in Amazon Web Services (AWS) with JavaEE, Android and .NET Mobile architecture.
- Led more than 20 projects for data capture automation for medium-large size companies and government agencies (Department of Defense).
- Developed a partnership with Zebra Technologies, local vendors and delivered company investors pitch, San Francisco, USA.

Education

Master of Data Science from University of Melbourne, Australia
Graduated Feb 2020

Graduate Diploma of Information Systems Management from Universidad Católica del Oriente, Colombia
Graduated Sept 2009

Bachelor of Science (Computing) from Universidad Católica del Oriente, Colombia
Graduated Sept 2006

Projects

Data Science Specialist at Daytaset Inc.
Implemented consultancy work in Latin America, US and Europe for projects involving:

- Implemented data pipelines in GCP and deployed time series data based ML Model for sales orders predictions and recommendations achieving MAPE below 10%.
- Led data lake and data governance consultancy projects.
- Integrated ML on audio data to implement voice-based interactions for in-store inventory management.
- Implemented a Grafana OSS observability dashboard integrated with InfluxDB to monitor energy operations.

SEEK Ltd, Named Entity Recognition and extraction in Job Ads.
Named Entity Recognition (NER) in Job Ads to extract company names, skills and requirements using contextual character-based word embeddings and BiLSTM-CRF Neural Networks.

Security Analytics, Machine learning for anomaly detection.

https://github.com/danielgil1/security_analytics

Machine Learning for Anomaly Detection in DDoS cybersecurity attacks and Security Analytics with Splunk.

Social Media Analytics & Cloud Computing, Scalable Machine Learning

https://github.com/danielgil1/social_media_analytics

Social Media Analytics application using Ansible automated scripts to deploy and scale in the cloud automatically a cluster of NoSql databases (CouchDB), queue services (RabbitMQ) and bots that crawl twitter feeds to perform sentiment analysis and political views summarized through map reduce plotted in a geospatial environment.

Ticket Intent Extraction, Accenture Project X Bootcamp.

Helpdesk ticket extraction using Unsupervised Learning (LDA Topic Model and K-Means), exploratory analysis and pre-processing using NLTK and visualizing results with pyLDAvis. A Jupyter Notebook is available to run models, a Flask API and Web App can be run separately or hosted on Splunk. Docker image is also available.

Language Models and conversational interfaces, Presented at Research Bazar 2019, University of Melbourne.

https://github.com/danielgil1/hangman-resbaz

Voice interaction for a hangman game AI with Amazon Alexa device using Language Models and AWS Lambda Functions.

Statistical Machine Learning, Missing Links prediction in Social network & Multi-armed bandits for Ads Optimization

https://danielgil1.github.io/machine-learning/

Supervised learning, semi-supervised and active learning, unsupervised learning, kernel methods, probabilistic graphical models, classifier combination, neural networks. Reference

Mathematical Statistics

https://danielgil1.github.io/mathematical-statistics/

Basic statistical concepts including maximum likelihood, sufficiency, unbiased estimation, confidence intervals, hypothesis testing and significance levels. Estimation of model parameters, hypothesis testing using analysis of variance, model selection, diagnostics on model assumptions, and prediction considering computational techniques, including the EM algorithm, Bayes methods and Monte-Carlo methods.

Artificial Intelligence, Reinforcement Learning Agent for Pacman

https://danielgil1.github.io/artificial-intelligence/

Reinforcement Learning; Game Theory; Search algorithms and heuristic functions; Classical (AI), probabilistic and non-deterministic planning, Monte-carlo Tree Search.

Natural Language Processing and Search, Question - Answering System

https://danielgil1.github.io/web-search-text-analysis

Text classification algorithms such as logistic regression; vector space models for natural language semantics; structured prediction, Hidden Markov models; N-gram language modelling, including statistical estimation;alignment of parallel corpora, Term indexing, term weighting for information retrieval; query expansion and relevance feedback.