Ran Xiao

Address: U2 16 Clive Street, Springvale

Mobile: 0402579388

Email: ran.xiao@factil.io

Driving Licence • Own Car

**Personal statement**

During my career in the field of data science, I have had the privilege to be directly involved in the engineering of various projects involving Data Analytics, Business Intelligence and Machine Learning. I have a passion for data science and am always on the lookout for new technologies. My interest in data science is what drives me to get up every day - I enjoy learning. Following on from my university course, I have continued to actively teach myself the mathematics behind Machine Learning and popular data science frameworks and libraries. My self-learning sessions give me the extra edge when building applications and allows me to suggest enhancements that may not otherwise be possible.

At Australia Post, I worked as a Junior Software Engineer where I had the rare opportunities to work on both frontend and backend applications within the Identity Services division. I was actively involved in Driver Onboarding and National Police Check applications. My time at Australia post exposed me to tech stacks including AWS, TypeScript, Python and Jenkins CICD pipelines and lean/agile development methodologies.

Later at Factil, our team was tasked to design and implement Machine Learning Pipelines for matching child protection data as part of the government BRII initiatives involving knowledge not only of good software design and implementation, but also knowledge in statistics and Machine Learning. My open attitude and ability to quickly adapt helped to succeed in these roles.

In addition to being flexible and adaptable, I am also a fanatic for details – particularly when it comes to application development and design. For example, in my recent role, my attention to detail lead me to discover an anomaly in our system, once remediated it prevented wrong information showing up during up an important product demonstration.

Building a Machine Learning application for child link from start to finish gave me a ton of insights at each step. I learned techniques that tackles lack of labelled data, data wrangling and transformation with sparse data, feature engineering and refinement, and adaptive importance sampling to more accurately assess our classifier. I believe these experiences will allow me to succeed in the role of a data scientist.

Although hard skills are incredibly important, communicating the findings and insights to the stakeholders can be just as important. During child link project at DET, I worked very closely to SMEs (subject matter experts). these exchanges helped them develop a better understanding of our product and methodology and also allowed me to collect useful feedbacks from them.

Skill Matrix

|  |  |
| --- | --- |
| Development Languages | *Python, Golang, Typescript, Javascript, Haskell, Julia, R, C, C++* |
| Python science tool stack | *Pandas, Matplotlib, Numpy, Numba, Cython, Seaborn, Scipy, Tensorflow* |
| Operating Systems | *Centos, Ubuntu, Mac OSX, Windows* |
| Database Technologies | *MySQL, Postgres, MongoDB* |
| Version control | *Git, Gitlab, Github, Bitbucket* |
| Cloud Environments | *AWS, Microsoft Azure* |

**Education**

**Monash University**

*(2016 – 2018)*

*Bachelors in Computer Science with WAM (weighted average mark) of 85*

**Work Experience**

### Application developer and data scientist (March 2019 -)

Factil, Melbourne

Project: ChildLink Project at DET

Responsibilities:

*Application development*

* Design and implement of our data matching pipeline which includes multiple stages.
* Load labelled data from Postgres database
* Transform and cleanse the data using Python and Pandas
* Train classifiers using loaded labelled data and perform data matching or scoring within one single data source or across two data sources
* Perform clustering based on scored result.

*Software engineering*

* + - Create models with Tensorflow and scikit-learn
    - Allow different classifiers, feature sets and blocking predicates be swapped using functional programming principles and deploying modular designs.
    - Refactor key existing components to reduce tight coupling and increase readability and allow easy future extension.
    - Unit test key functionalities of the applications

*Performance tuning*

* + - Reduce the memory consumption of our application by 50% using efficient data structure and knowledge in Python
    - Increase our classifiers precision and recall in our application from 70% to above 95% by both reengineering our feature vectors and improving the labelled data quality
    - Reduce running time by profiling our applications and identified and resolve multiple IO-bound performance issues.

*Stakeholder engagement*

* + - Communicate and educate our business SMEs on our data matching platform.
    - Communicate with SMEs to resolve multiple data source issues that could impact matching accuracy.
    - Produce professional charts and graphs using matplotlib, seaborn that are included in the final submitted report.

*Tools used:*

Python, Postgres, Centos Linux, matplotlib, seaborn, pandas, numpy, scipy, jupyter notebook

### Summer Research project (Dec 2018 – Feb 2019)

Factil, Melbourne

Project: Business Research and Innovation Initiative (BRII)

Responsibilities:

* + - Apply Machine Learning approach for record linkage and deduplication for NSW, ACT children data.
    - Perform data analysis on datasets to assist feature extraction and feature selections.
    - Perform data cleansing, feature extraction, feature selections and use various models such as logistic regression and gradient boosting that infer the similarity between records to high degree of accuracy.
    - Build models that can handle poor data quality (sparse datasets with many missing entries)

Tools used:

Python, Postgres, Centos Linux, pandas, matplotlib, jupyter notebook

### IBL Internship (Jan 2018 - June 2018)

Australia Post, Melbourne

Responsibilities:

* Complete a Machine Learning spike project that demonstrates the mail redirection efficiency can be improved substantially, where neural nets were used for data matching.
* Worked backend on driver onboarding program with technologies such as AWS lambdas, cloud formation and step functions.
* Worked frontend on national police check service with technologies such react library.

Tools used:

Typescript, Python, dynamodb, AWS cloud, cloud formation and Jenkins

### Summer research project (Nov 2017 – Jan 2018)

Factil, Melbourne

Responsibilities:

* Develop a visualization system for a business modelling language using popular JavaScript library d3 and webcola which is developed by my professor Tim Dwyer.

Tools used:

Typescript, Javascript, D3 and webcola.

### Summer research project (Dec 2016 – Feb 2017)

Zendesk, Melbourne

Responsibilities:

* Conducted time series unsupervised learning on leaked Hilary email dataset and Arxiv dataset using Machine Learning algorithm such as LDA and its variants
* Visualized the output using JavaScript d3 library

Tool used:

Python, C++, nltk and genism, BeautifulSoup4, JavaScript and d3