

# Chun Hei Siu (Charles)

me@chunheisiu.com | WeChat chunheisiu | Kowloon, Hong Kong SAR

## EDUCATION

### **Master of Science in Data Science**

**University of San Francisco**

San Francisco, CA, USA

Jul 2019 – Jun 2020

- Scholarship recipient, Fall 2019, Spring 2020.

### **Bachelor of Science in Data Science, Minor in Design**

**University of San Francisco**

San Francisco, CA, USA

*cum laude*, Aug 2014 – May 2018

*Concentration in Computational Analytics*

- USF Provost's Merit Award recipient, 2014 – 2018.
- Dean's Honor Roll, Fall 2014 – Spring 2018.

## WORK EXPERIENCE

### **IBM**

*Junior Data Scientist*

Lantau, Hong Kong

Nov 2020 – Current

- Designed and implemented data parsing and transformation scripts in Python for Apache Airflow workflows, including XML and JSON to CSV conversion scripts, which are used for data ingestion in cloud databases.
- Developed and integrated Informatica ETL workflows and mappings for business layers in the cloud data warehouse, which simplify OLAP queries for analytics and optimizations.

### **Hong Kong Applied Science and Technology Research Institute (ASTRI)**

*Engineer, Software Development*

Shatin, Hong Kong

Aug 2018 – Jun 2019

- Designed and developed REST APIs in the backend module for the Intelligent Ad Management Platform.
- Implemented DevOps such as CI/CD with Jenkins and GitLab CI in the development environment. Improved productivity by eliminating the need for manual image building and deployment.
- Integrated service mesh management with Istio on Kubernetes and presented the topic to team members.
- Designed Kibana dashboards for metric aggregation and distributed tracing in the development environment.
- Maintained workstations and server equipment for software development and client demonstration. Coordinated with the project manager for equipment procurement.

## INTERSHIP EXPERIENCE

### **Valimail**

*Data Science Intern*

San Francisco, CA, USA

Oct 2019 – Jun 2020

- Tackled the problem of classifying a backlog of 100K+ unknown internet domains generated by Valimail Defend.
- Developed an end-to-end machine learning pipeline that classifies domains by detecting whether they belong to low-risk categories such as real estate. The Gradient Boosting Machine (GBM) model achieved a 95%+ precision rate with test data when classifying real estate domains using Natural Language Processing (NLP) for web content analysis. Going into production, the pipeline greatly reduced the reliance on manual data labeling.
- Designed and implemented RPC APIs using Flask in Dockerized modules in the pipeline.
- Built web scrapers using BeautifulSoup to gather multiple external data sources for ML model training.

### **Macroview Telecom Limited**

*Summer Intern, Software Development*

Quarry Bay, Hong Kong

Jun 2016 – Aug 2016

- Performed quality control tests for in-house software applications integrated with Cisco UC Solutions.
- Developed and presented a demo application integrated with Cisco Jabber Web SDK.

## PROJECTS

- **HireReady:** This project was a web application that provided interview preparation for data science-related job seekers by using machine learning algorithms to generate mock interview questions that are tailored to a specific job posting. It was selected as one of the best projects and received positive feedback during the presentation to a panel of venture capitalists.
- **NBA Player of the Week & Salary Prediction:** This project used various data sources related to NBA players to predict how likely a player is to receive the Player of the Week award using logistic regression, as well as to predict his salary using linear regression, for a given NBA season.
- **USF MEDA Practicum:** This project, in collaboration with the Mission Economic Development Agency, analyzed and visualized various distance and school placement metrics of elementary school students who are clients of MEDA, using publicly available school quality and demographic information combined with proprietary data.

## TECHNICAL SKILLS

- Programming Languages: Java, Python, R, JavaScript, SQL, HTML, CSS
- Operating Systems: Windows, macOS, Ubuntu, CentOS
- Tools & Platforms: Jupyter Notebook, Apache Spark, Docker, Kubernetes, Jenkins, Elasticsearch, Kibana, Redis, StreamSets Data Collector, Apache Kafka, Apache Airflow, Informatica, RStudio