|  |  |
| --- | --- |
| HE Yingxu | *Singapore Permanent Resident* | 17 Jalan Masjid, Singapore  +65 91752741 | yingxu.he1998@gmail.com  [Linkedin](https://www.linkedin.com/in/yingxu-he-855088140/) | [Personal Website](https://yingxuh.github.io/index.html) |
| Experienced machine learning researcher with a proven track record of utilizing advanced prompting techniques on ChatGPT, developing efficient machine learning pipelines at Dyson, and building task-oriented dialogue systems and information retrieval systems at NUS labs. Seeking opportunities in the fields of ML and NLP. | |

# Working Experience

## Machine Learning Researcher / Freelance. Jan 2023 – Present

* Devised and implemented a confident learning method to tackle label inaccuracies in object detection models. On an oil storage detection dataset, adjusted over 60 training images to achieve a notable increase in the YOLOv8's recall (from 56% to 64%) and MAP@50 (from 62 to 65) over 2k+ test images.
* Employed advanced Chain-of-Thought and Maieutic prompting techniques on ChatGPT to create 10k+ accurate captions from annotations in the xView dataset. Fine-tuned a CLIP model and significantly improved the recall@5 of image-to-caption classification from 52% to 82.3% over 2k samples.

## Machine Learning Research Engineer / Dyson Ltd. Sept 2021 – Dec 2022

* Implemented an object localization model in a few-shot context by semi-supervised training. The model achieved comparable results to professional software with improved adaptability and robustness.
* Designed and implemented an air quality estimation model, using LGBM, Bayesian Regression, etc., with geographical and meteorological features. Demonstrated its advantages over spatial interpolated methods and deployed the pipeline with AWS Step Function services.

## Machine Learning Research Assistant / NUS-Singtel Cyber Security Lab Sept 2020 – July 2021

* Identified anomalies from system logs leveraging DBSCAN and hierarchical clustering for model training.
* Developed an information retrieval method for web-attack strategy identification from system and firewall logs. The recall@3 rate achieved 80% on 100+ hand-labelled samples.

## Data Analyst Intern / GIC Pte. Ltd. Dec 2018 – July 2019

* Deployed an R application that forecasts the mid-term returns of portfolio with visualization using R shiny.
* Optimized the coefficients of a mean reversion forecasting model using the Genetic Algorithm.

## Data Analyst Intern / PropertyGuru May 2018 – Aug 2018

* Developed dashboards in Tableau to analyze the user behaviors and listings’ performance to better match user demand to agents’ recommendations.
* Implemented a POC to calculate and geographically visualize the liveability score for properties.

# Education

## MComp in Artificial Intelligence / NUS Aug 2020 – Sept 2021

* **School of Computing**: CAP 4.42/5.0
* **Teaching Assistant**: Advanced Analytics and Machine Learning  **Jan 2021 – May 2021**

## BSc. (Hons) in Business Analytics / NUS Aug 2016 – June 2020

* **School of Computing**: CAP 4.15/5.0, Dean’s List in Semester 3 AY 2018/2019
* **Distinction**: [Analytics Techniques Knowledge Area](https://github.com/YingxuH/YingxuH.github.io/blob/master/cert/Certificate%20of%20Distinction%20in%20Analytics%20Techniques%20Knowledge%20Area.pdf) **Dec 2020**
* **Teaching Assistant**: Programming Methodology in python **Aug 2017 – June 2018**

# Academic Projects

[**Dialogue Response Generation (Master Thesis)**](https://www.academia.edu/63387276/Towards_Enriching_Responses_with_Crowd_sourced_Knowledge_for_Task_oriented_Dialogue) **/** **NUS NExT++ Lab Nov 2020 – Aug 2021**

* Built an enriched task-oriented response generation by implementing copy-mechanism on GPT-2 using Pytorch. The proposed model is capable of naturally incorporating external tips/user reviews about venues into responses. The generated response outperforms many state-of-the-art models on user satisfaction.

**Property Resale Price Prediction Jan 2021 – May 2021**

* Fitted CatBoost, LGBM, XGBoost on 43k pieces of property sales data. Selected features by correlation and information gain. Engineered new features describing properties’ livability. Reduced data dimensionality with WOE encoding. The final ensemble methods’ accuracy achieved 5th/64 place.

# Skills

|  |  |
| --- | --- |
| * **Python (Pytorch, Tensorflow), R:** Machine Learning, Deep Learning, Data processing * **SQL, Spark:** Data query and big data | * **Tableau, PowerBI:** Visualization development * **Java, Git, Scala, JavaScript, HTML, CSS:** Software Development |