|  |  |
| --- | --- |
| **JAMES LIM** | |
| +6598062879 | e0950510@u.nus.edu | <http://www.linkedin.com/in/jameslim2604>| <https://github.com/astrozoid2604> | |
| **EDUCATION** | |
| **National University of Singapore,** MSc in Data Science and Machine Learning | **Aug 2022 – May 2024** |
| **Nanyang Technological University,** BEng in Electrical and Electronic Engineering (1st Class Honours) | **Aug 2013 – Jun 2017** |
| **TECHNICAL SKILLS** | |
| **Languages:** C++17, Python, C, Perl, MatLab, ObjectScript, Scala, HTML, CSS.  **Data Science:** A/B testing, ETL, Statistics, Time Series, Hypothesis testing, OOP, OOD, Web crawling, Tableau, Power BI, Scripting.  **Data Engineering:** End-to-end data ecosystem (sourcing, ingestion, curation, data mart, visualization dashboard), RESTful API, SFTP server.  **Big Data Technologies:**  Spark, Hadoop, MySQL, MongoDB, Python (e.g. numpy, pandas, matplotlib, seaborn).  **DevOps Technologies:** Git, Svn, Linux, Unix, Jira, Confluence, Jenkins CI/CD, Docker, AWS Cloud.  **MLOps Technologies**: Python (e.g. Scikit-Learn, TensorFlow-Keras, Torch, NLTK, Gensim, spaCy, Transformer, SpellChecker), OpenAI, TensorBoard. | |
| **WORK EXPERIENCE** | |
| **AI Research Intern (NLP – Large Language Models),** Hewlett Packard Enterprise, Singapore | **Jan 2024 – Present** |
| * Co-first authored 2 research papers accepted in 2024 ICLR ([**Paper LINK**](https://openreview.net/pdf?id=I8bsxPWLNF)) and 2024 ACM ([**Paper LINK**](https://dcai-workshop.github.io/assets/pdf/accepted_papers/106.pdf)), presenting data-efficient adaptive sampling strategy for evaluating LLM benchmark datasets while still preserving 90% Pearson correlation coefficient using only 1% of the benchmark dataset. * Innovated a novel Python pipeline integrating NLP libraries (e.g. Scikit-Learn, NLTK, Gensim, spaCy, Transformer, SpellChecker) to select top 50 LLMs from HuggingFace leaderboard, and adaptively choose best sampling technique for each benchmark, considering Pearson correlation coefficient and Wasserstein distance results evaluated on benchmark dataset attributes, namely, text quality, topic classification, and latent space distribution. * Orchestrated a Large Language Model (LLM) Benchmark Description Database using OpenAI’s GPT-4 API, meticulously extracting and structuring abstract information and question samples from extensive benchmark datasets, to improve similarity score by 10% on in-house recommender pipeline. | |
| **AI Research Intern (Generative AI – Time Series Predictive Maintenance),** Robert Bosch, Singapore | **Aug 2023 – Jan 2024** |
| * Spearheaded end-to-end development and distribution of DoppelGANger time-series Generative AI model for predictive maintenance, resulting in a significant reduction of 48-hour downtime and SGD 10K maintenance cost per machine. Leveraged PyTorch, DDP data parallelism API, and AWS cloud infrastructure to establish predictive maintenance schedules based on frequency-domain profiles and near-future latent failure characteristics. | |
| **Data Science Intern,** InterSystems, Singapore | **May 2023 – Jul 2023** |
| * Developed an efficient ETL pipeline to process 50TB unstructured and structured data from client's AWS platform. Applied a 5s-latency PDF metadata extraction algorithm in Python to ingest payment invoice data into SQL databases and SFTP servers for 28 Singapore banks. Utilized techniques such as key-value pair tokenization, tabular data detection, and Data-Frame transformation to convert metadata to JSON & XML formats. * Deployed an interactive Iris BI dashboard with real-time integration to health insurance claim database system, empowering business stakeholders to monitor KPIs for insurance policy making and fraudulent claim detection. Presented dashboard's strategy and value proposition to 11 country heads and senior leaders, resulting in better decision-making processes, and strengthened fraud detection capabilities. | |
| **AI Research Intern (Generative AI – Computer Vision),** Huawei, Singapore | **Jan 2023 – Apr 2023** |
| * Programmed a knowledge transfer capability on Huawei's mesh and texture database system using Get3D data modelling algorithm and leveraging DNN-Lib and PyTorch. Trained generative adversarial networks with Blender-rendered Shape-Net 3D objects to create style-adjustable, topologically complex, and high-fidelity-textured explicit 3D meshes, resulting in 5% optimization in mesh quality and 7% reduction in training time. * Delivered image parsing functionalities in Python using object-oriented programming techniques, allowing real-time detection of human face and body poses from streaming images. Used technologies such as OpenCV and PyTorch to accurately plot pixels and 3D regions of interest (ROIs) and landmarks, resulting in 9% higher in detection accuracy. | |
| **Senior Lead Algorithm Engineer,** Micron Technology, Singapore | **Jul 2017 – Jan 2023** |
| * Led a team of five engineers in developing C++ and Python product quality test programs in a Linux/Unix environment, focusing on attaining low-latency performance through asynchronous handshake with low-level technology FPGA. Elevated product quality and reduced latency by 10%, capitalizing on joint efforts and effective leadership under weekly time pressure. * Deployed ML models on a weekly basis, harnessing on-the-fly multi-threading between Perl-script-triggered XGBoost algorithms and C++-based test programs, attaining a 90% defect catch rate, leading to colossal cost savings of $13 million in capital expenditures. Demonstrated expertise in algorithm optimization and rapid release methodologies, contributing to efficiency and performance. * Built a Tableau dashboard to analyse and visualize statistical evaluation, including analysis of variance (ANOVA), on a vast dataset of 15TB device performance data. It enabled timely feedback to refine test conditions for subsequent production test program releases on a weekly basis, resulting in 10% higher in overall testing efficiency. * Realized Jenkins as CI/CD pipeline solution, resulting in 20% increase in productivity by saving 3-man hours per product per week, streamlining entire test program development cycle, including code build, Git commit, baseline testing of corner cases, and production floor release. | |
| **NOTEWORTHY PROJECTS** | |
| **Stock Trading Bot with Automated Scheduling** | **Aug 2023 – Mar 2024** |
| * Coded and backtested a Python-based trading bot with automated scheduling on Robinhood brokerage platform, earning a 7% return over a 7-month period. Introduced price reversal detection harnessing candlestick patterns and applied exponentially weighted moving average smoothing on Holt-Winter’s model, resulting in well-executed trading strategies and investment returns. | |
| **Question-Answering (QA) Big Data Recommender System (**[**Public GitHub Link**](https://github.com/astrozoid2604/GPT2_qa_recommender)**)** | **Jan 2024 – Feb 2024** |
| * Orchestrated a question-answering based recommender system for Apache Spark big data applications by fine-tuning of a 6GB LinkedIn job dataset on HF Transformer’s GPT-2 and DistilBERTLanguage Model (LLM), applying advanced natural language processing techniques (text wrangling, tokenization, lemmatization, vectorization, feature extraction and prompt engineering), resulting in 30 BLEU, 40 ROUGE, and 0.2 METEOR scores. | |
| **MayBank Customer Segment Upselling (**[**Public GitHub Link**](https://github.com/astrozoid2604/maybank_assessment)**)** | **Nov 2023 – Jan 2024** |
| * Developed a predictive model and proposed upselling strategy for hidden affluent MayBank customers by utilizing TensorFlow ML algorithms (XGBoost, GaussianNB, and RBF-Kernel SVC) with StratifiedKFold cross validation and GridSearchCV hyperparameter tuning which successfully identified 78 potential affluent customers, contributing to 0.8% increased revenue on affluent customer segment. | |
| **AirBnB Data Warehousing System Designs (**[**Public GitHub Link**](https://github.com/astrozoid2604/AirBnB_Data_Warehousing_System_Design)**)** | **Aug 2022 – Dec 2022** |
| * Designed a relational database in MySQL and NoSQL for a 10GB AirBnB sales unstructured dataset, implementing data processing techniques to examine, clean, and optimize dataset structure and semantics. Employed Boyce-Codd Normal Form schema refinement to ensure conceptual and logical database design, refining data management and analysis efficiency. | |