Joseph Lee

(206) 458-9073 aimlopsengineer@gmail.com linkedin.com/in/lee-sangwoo/ github.com/josephswlee **U.S. Citizen**

Experience

Capital One, Machine Learning Engineer **Aug 2024 - Current**

- Develop a software development kit (SDK) providing a unified interface for managing, storing, and retrieving machine learning features across diverse data formats (e.g., Pandas, Apache Spark, Polars), ensuring consistency, scalability, and ease of use for Applied Researchers, Data Scientists, Machine Learning Engineers, and Data Analysts throughout the machine learning lifecycle
- Deploy and maintain the codebase within a CI/CD pipeline using Jenkins, automating artifact deployment
- Engineer tools to prevent the exposure of enterprise secrets by developers and scientists, utilizing Logistic Regression, Random Forest, Boosting and Support Vector Machines to detect potential instances of secret exposure within code repositories, resulting in a 100% reduction in secret leaks (from 4 bi-monthly instances to 0) over the past two months

US-Ignite, Data Scientist Intern **May 2024 - Aug 2024**

- Led a team of three graduate students in designing and implementing a robust data pipeline on Azure Databricks using Apache Spark, integrating over one million row data records from disparate sources to enable downstream analytics
- Automated the ETL process for over a million data records, achieving 99% uptime and reducing monthly manual workload by 40%, significantly improving data processing efficiency
- Supervised and mentored a team in developing machine learning models to predict building energy consumption, applying Random Forest, XGBoost, RNN, and LSTM to building geo location data, occupancy, and energy consumption data

Wharton Analytics Fellow - PETCO, Senior Analyst **Feb 2024 - Apr 2024**

- Built machine learning models to predict creative features that lead to a higher click-through rate (CTR) for email advertisements
- Led a team of 5 Senior and Junior Analyst to perform EDA and extract text and image features from the advertisement emails from PETCO
- Utilized **GPT-4 API** and transformer model (**BERT**) to extract relevant text features and evaluating features using **linear models**, **random forest**, and **XGBoost** against Click-through Rate (CTR)
- Created a customized email generator for individual customers using **RAG** and the **GPT-4 API**

ConcertAI, Data Scientist Intern **May 2023 - Dec 2023**

- Developed five advanced reporting solutions for Janssen Pharmaceuticals (Johnson & Johnson) and the analytics team, utilizing **Azure Databricks**, advanced **SQL** techniques, and **PySpark** to provide reproducible and customizable reports for assigned Multiple Myeloma treatment medication
- Enhanced data curation by assembling comprehensive datasets from diverse sources, including pharmaceutical and medical claims data, strategically implemented business rules tailored to each dataset to optimize individual reporting solutions
- Conducted meticulous and insightful **ad hoc analyses** of EMR data to tackle critical business challenges, such as enhancing persistency rates and marketing strategy for the client

Education

University of Pennsylvania **May 2024**
Master's degree in Data Science, GPA: 3.97/4.0 Philadelphia, PA
Coursework: Machine Learning, Artificial Intelligence, NLP, Big Data Analytics (**TA**)

University of Virginia **May 2022**
Bachelor's degree in Computer Science, GPA: 3.97/4.0 Charlottesville, VA

Selected Projects

Personalized Course Recommendation for Education with LLMs | *GPT-3.5 API, LangChain, HuggingFace API, Selenium, Python*

- Developed end-to-end machine learning solution for personalized course recommendations for students
- Wrote Python script using the Selenium package to extract and combine data from the University of Pennsylvania's course catalog website, generating a structured dataset for analysis
- Developed and implemented automated chatbot solutions using appropriate NLP methodologies such as Retrieval-Augmented Generation (RAG) and SQL Chain querying a vector database to mitigate LLM hallucinations

AI Professional Profile Picture Generator | *Full-Stack, Generative AI, HuggingFace API, Flask, React.js, HTML/CSS*

- Developed a full-stack machine learning application for generating AI-driven professional headshots
- Applied the HuggingFace diffusers package to construct a Stable Diffusion model pipeline, enabling text-to-image generation
- Applied machine learning algorithms such as Stable Diffusion with customized low-rank adaptations (LoRA) and trained the model with user images to effectively capture essential facial features for generating personalized headshots

Skills

Programming Language: Python, R, SQL, Java, C++, JavaScript
Technologies/Frameworks: Pandas, Numpy, PySpark, PyTorch, Scikit-Learn, Hadoop, Huggingface, Linux, Databricks, MongoDB, Google Cloud Platform, Amazon EC2, Amazon S3, SQL server, Git, React.js, HTML/CSS, Flask