Joseph Lim

647-929-6726 | j67lim@uwaterloo.ca | jhlim0921 | josephlim0921 | josephhyunjinlim.com

TECHNICAL SKILLS

Programming Languages: Python (Pandas, NumPy, PySpark Matplotlib, Seaborn), SQL (MySQL, PostgreSQL, MS SQL)
Machine/Deep Learning: Scikit-Learn, XGBoost, PyTorch, TensorFlow, Keras, Imblearn, Hugging Face, Unsloth
Data Engineering/Analytics: AWS, Azure Databricks, Git, MLflow, Tableau, Power BI, Looker, Docker, Airflow

EXPERIENCE

Data Scientist | **Gore Mutual Insurance Company** May 2024 – Aug 2024 | *Toronto, ON*

- Analyzed a Random Forest classifier that predicts large losses in personal properties using SHAP values, reporting 5 critical risk factors to underwriters that commonly contribute to increased propensity
- Revamped the large loss detection model by resolving dataset imbalance with SMOTE and testing multiple classifiers, boosting recall to approximately 70% and improving lift curve performance for early risk detection
- Developed a pipeline for a commercial auto loss cost project that trains an XGBoost Tweedie Regressor and generates custom partial dependence plots, enhancing model explainability and providing deeper insights into 20+ key features
- Consolidated data across 5+ sources using SQL and PySpark to enable downstream predictive modeling for commercial auto projects, improving data reliability through rigorous validation checks

Data Scientist | **PepsiCo** Sept 2023 – Dec 2023 | *Mississauga, ON*

- Spearheaded a national store segmentation project for Quaker, employing PCA and K-Means Clustering on demographics data to cluster 3000+ Canadian stores, uncovering optimization opportunities in retail operations across 7 product categories
- Commercialized a ML project with senior data scientists by building 10+ interactive Power BI dashboards linked to model outputs in Delta Lake, providing real-time shopper insights to business stakeholders
- Analyzed over 1B rows of POS sales and demographics data using SQL, Pandas, and PySpark, driving strategic execution recommendations for the field team in preparation for a new Frito-Lay product launch
- Developed Ridge Regression models to forecast the sales performance of non-existing store-product combinations across 4 competitor product lines, identifying 1,000+ high-potential stores to target for competitive market entry

Associate Producer | **Zynga** Jan 2023 – Apr 2023 | *Toronto, ON*

- Developed SQL queries in MS SQL and used Pandas to streamline data collection and analysis on team KPIs, increasing the efficiency of processes by more than 80%
- Built interactive reports and dashboards in Looker to equip 10+ cross-functional agile teams with valuable insights for data-driven improvements to their sprint performances
- Analyzed project data using SQL by generating relevant statistics on resource availabilities and project durations to create project roadmaps, resulting in a 50% increase in project/OKR tracking efficiency for teams

Junior Product Manager | **Front Rush, NCSA, Zcruit** May 2022 – Aug 2022 | *Chicago, IL*

- Utilized Heap to analyze customer data across 3 products by defining KPIs and usage metrics that generated insights on over 10,000 daily users, ultimately guiding future product decisions and feature enhancements
- Conducted a customer retention analysis for the Zcruit portal, identifying critical improvement areas that informed the development of a strategic product plan to enhance user satisfaction

PROJECTS

ARC Prize Challenge | *Pandas, PyTorch, Unsloth* Dec 2024 | *Waterloo, ON*

- Tackled the ARC-AGI Challenge by experimenting with LLMs, VLMs and fine-tuning them using LoRA, exploring techniques to solve novel tasks
- Engineered specialized models by fine-tuning LLaMA language models on augmented clusters of similar puzzle tasks using Unsloth, achieving notable performance gains in early evaluations

MLB Pitch Classification | *Pandas, Scikit-Learn, XGBoost, TensorFlow, Keras* Dec 2024 | *Waterloo, ON*

- Developed a pipeline consisting of data preparation, hyperparameter tuning, and model evaluation, enabling experimentation with multiple machine and deep learning algorithms, ultimately achieving a 79% test accuracy in classifying pitch types

NBA Game Winner Predictor | *Pandas, Scikit-Learn, XGBoost* Mar 2024 | *Toronto, ON*

- Led a comprehensive machine learning project to predict NBA game outcomes by overseeing and executing all stages from data collection to model evaluation, achieving a test accuracy of 70%

EDUCATION

University of Waterloo | **BASc. Systems Design Engineering** Sept 2020 – Present | *Waterloo, ON*

- Courses: Data Structures and Algorithms, Probability and Statistics, Applied Linear Algebra, Foundations of AI, Deep Learning, Intro to Pattern Recognition, Optimization and Numerical Methods, Advanced Machine Learning, Computational Simulations of Environmental and Societal Systems
- Cumulative GPA: 86.9%