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
Data Engineering/Analytics: Azure Databricks, Git, MLflow, Tableau, Power BI, Looker, BigQuery, Docker, Airflow

EXPERIENCE

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

- Utilized SHAP values to analyze an existing Random Forest classifier that predicts large losses in personal properties, identifying and reporting 5 key features to underwriters that commonly contribute to increased propensity
- Led the revamp initiative to enhance the large loss model by implementing SMOTE for dataset imbalance and experimenting with various classifiers, achieving a strong lift curve performance and a testing recall score of approximately 70%
- 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
- Synthesized datasets from 5+ sources using SQL and PySpark for downstream predictive modeling in commercial auto, ensuring data quality and consistency 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 effectively cluster 3000+ Canadian stores, identifying opportunities to optimize 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
- Conducted comprehensive data analysis on 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, thereby generating a prioritized list of 1000+ 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 Python libraries (Pandas, NumPy) 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

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

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

K-pop Song Recommender | *Pandas, NumPy, Scikit-Learn, Spotipy* — Dec 2023 | *Toronto, ON*

- Constructed a data pipeline using a Spotify API to extract and transform features of songs from multiple K-pop artists, thereby creating a well-structured dataset for efficient downstream analysis
- Developed a content-based recommendation system for K-pop songs, utilizing cosine similarity to calculate similarity scores and suggest top song recommendations to users

EDUCATION

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

- Courses: Data Structures and Algorithms, Probability and Statistics, Machine Learning, Applied Linear Algebra, Foundations of AI, Deep Learning, Intro to Pattern Recognition, Optimization and Numerical Methods
- Cumulative GPA: 85.8%