Yu(Francis) Tao

585 Bloor St E, Toronto, ON M4W0B3 | P: 778-681-6828 | E: ty199695@gmail.com | LinkedIn | GitHub

Professional Summary

Master students with major in DA/ML and 3 years of solid experience in data analysis, reporting, automation, and machine learning/DL across healthcare, waste management, and medical databases industries, who's eager to apply data expertise in an internship to support business decision-making and further develop ML competencies.

Skills

Data Analytics/Science: Pandas, Numpy, Sklearn, Power Query, DAX
Data Visualization: Tableau, Power BI, Matplotlib, Seaborn
ML/DL: PyTorch
Web Scape and Analytics: Selenium, Google Analytics
Others: Excel, Access, Outlook, Slack, HubSpot, SharePoint, Git(Basic), Splunk
Statistical Programming: Python, SAS, RStudio
Database: PostgreSql, MySQL, T-SQL, SSMS
Cloud: GCP BigQuery, Azure, Databrick
ERP/CRM: Microsoft NAV, Dynamics 365
Agile Project Management/SaaS: Jira, Azure DevOps

Work History

Policy Reporter Vancouver, BC
Data Quality Coordinator II *09/2023- 06/2024*

- Develop the pipeline in **Python** with **Postgresql** connection and **SharePy/Google API** integrated to automate the daily refresh of over 1000 total tracking errors in company's database/portal, daily backlog report, index specialization sheet, and automatically output the data with adjusted format to Excel spreadsheets in OneDrive by **openpyxl** and **ExcelWriter** which helps the team to find the associated specialists, unclaimed tasks, coverage/non-coverage documents, missing/rebranded payers, and fix the issues.
- Leverage **pandas** library and **CTEs** in **Python** & **Postgresql** stack to address the discrepancy and difference over 700 portal users and 37 access before and after each user bulk upload on a weekly basis and reduce the inconsistence locating time by 90%.
- Script in **pgAdmin(Postgresql)** to locate the client & user emails which are/aren't category alerted/subscribed: CTEs, Unnest, Join, Case when, Union, Leading, Trailing, Trim, etc used.
- Leverage **pypdf**, **PdfReader**, **IO**, **Beautifulsoup**, **Levenshtein, SequenceMatcher** and **tqdm** to carry out the duplicate textual analysis and eliminate 1800+ duplicates from 8900+ suspicious document pairs; Automate the Medicare, Medicaid and Government Plan document type detection by extracting content from over 41k PDF & html documents with API calling to company's portal and monitor the progress.
- Define recursive function and **pandas merge** in **Python** to find the depth of product category levels, apply hierarchical joins and create the ultimate category mapping of product indictions.
- Utilize the **ThreadPoolExecutor** to boost the code computing speed by 4.8x faster for most automation projects.
- Generate the automated weekly dashboard to present the weekly stats and visuals of tracking errors, backlog, index processed and influx by **matplotlib**, **seaborn, datetime** and **html** to showcase team's productivity to management stakeholders.
- Resolve various ad hoc requests from internal and external team by providing technical solutions to Customer Support, Data Quality Assurance and Data Analysis department.
- Host the **SQL** training sessions for junior da and co-op biweekly, which improves the overall data skill level in Data Quality Dept.

RecycleSmart Richmond, BC
Data Coordinator *11/2021 – 12/2022*

- Pull the data from **Microsoft SQL Server**, generate 250+/3300+ customized/standard waste diversion reports monthly in **Excel** and present insights of recycling data through **Power Query** and **Power BI** dashboard.
- Clean and transform over 8300 rows of customers/vendors' addresses by building **Python** validation tools with **Pandas**, **CanadaPost/BingMap API**, computing textual similarity using **theFuzz/Levenshtein Distance** with 75.1% accuracy, converting kmz file to json and creating region/neighborhood codes based on **Polygon** geospatial analysis(coordinates).
- Write **SQL** query and utilize **Power Query/Power BI** to transform the data and create visualizations which help a&f team to address 17.51% of uneven vendor lines and 106.7k invoice discrepancy for a single agreement type from Jun 2021 to Jun 2022.
- Develop a data audit tool with stacks including **SQL Server**, **Regex**, **Pandas**, **dateutil.parser**, **datetime**, **calendar** and **self def-func** in **Jupyter Python** which automates the checking process of 3500+ customer sites in terms of their waste diversion numbers, reducing the workload by 66.2% in average over past 12 months and saves 6 hrs monthly.
- Connect with IoT team and build a dashboard in **Power BI** to visualize bin's fill levels for applicable customer sites and optimize hauler's garbage pick up strategy, which increases the efficiency and cost control.
- Automate the data quality fixation and migration process of generating over 15k vendor purchase contract lines' unique key, unit price, region code and product id through the entire ETL in **SQL server(CTE, Subquery, Join, Concat, Dense_Rank, Case When)** and **Python(lambda, pandas)** for new ERP(**Dynamics 365**) system and UAT.

ehsAI Vancouver, BC
Data Labelling Analyst *02/2021 – 10/2021*

- Process EHS documents and conduct language analysis including citations, language labelling, splitting, and classifications to feed the NLP model.
- Develop internal QA survey, evaluate the employees' work performance by tracking workflow in **Jira** and analysing assignments including trend prediction based on visualization through **Power BI/Excel Query**, **Pivot**, Charts, **Python Pandas**, Matplotlib, **NumPy**, **Seaborn** and improve the task accuracy by 10% within 6mo.

Providence Health Care Vancouver, BC
Data Quality Clerk *01/2020 – 06/2020*

- Assist end-users with registration-related processes, liaise with various departments and escalate problems.
- Extract, sort and aggregate data using **SQL** and verify them from client registry/clinical information reports to maintain data integrity and to ensure compliance with high accuracy of above 85%.

ECCO Shoes Canada Richmond, BC
Brand Ambassador *11/2019 – 10/2021*

- Greet customers, help to find the wanted shoes, fix the displays and resolve the return and exchange issues.
- Generate more than $1500 in weekly sales with successful marketing, sales and customer relations approaches.

Education

University of Toronto, St. George Toronto, ON
Master of Engineering, MIE, Emphasis in Data Analytics and Machine Learning, CGPA:3.93/4.0 *Expected 09/2024 - 05/2026*
Courses:

- ECE 1513 (Fall): Introduction to Machine Learning (A+)
- ECE 1508 (Fall): Applied Deep Learning (A-)
- APS 1050 (Fall): Blockchain Technologies and Cryptocurrencies (A)
- MIE 1628 (Fall): Cloud-Based Data Analytics (A)
- APS 1070 (Winter): Foundations of Data Analytics and Machine Learning
- MIE 1517 (Winter): Introduction to Deep Learning
- CHE 1148 (Winter): Process Data Analytics
- APS 1080 (Summer): Introduction to Reinforcement Learning

The University of British Columbia Vancouver, BC
Bachelor of Science, Statistics *05/2019*
Dean's List 2016 – 2017; Member of Golden Key Society: Top 15% in faculty; Graduated with 82%

Personal Projects

Network Intrusion Detection – Developed a intrusion detection system based on 1.3 millions of records by leveraging **Logistic Regression**, **SVM**, **Random Forest** models and also building **Artificial Neural Network** with dropout layers, learning rate scheduler and regularization, achieving the best test accuracy of 99.8889% with the **ANN** model.

Face Mask Detection – Designed a real-time detection system using **YOLOv5** and **Faster R-CNN** to classify mask compliance into three categories. Data preprocessing and modeling using torchvision, albumentations, matplotlib, etc. Achieved mAP@0.5: 0.949 with **YOLOv5** and mAP@0.5-0.95: 0.668 with **Faster R-CNN**.

Bike Share Demand Prediction (Link) – 0.4075 RMSLE(301/3243 on Kaggle, top 9%); Data manipulation with **Pandas**, **NumPy** and datetime; Histogram, scatterplot and bars visualization with **matplotlib**; Modelling and training with **Train_Test_Split**, **LinearRegression**, **RandomForestRegressor** and **XgBoost**.

Heart Disease Prediction (Link) – **GridSearchCV** Logistic Regressoin with 88.5% accuracy; 303 rows of data; Barplot, scatterplot and heatmap by **matplotlib** and **seaborn**; Modelling and training with **LogisticRegression**, **KNeighborsClassifier**. **RandomForestClassifier** and **DecisionTreeClassifier**; **Hypertuning** using **RandomizeSearchCV** and **GridSearchCV**

Certification

(1). Data Analyst with Python
(2). MCSA: SQL 2016 Database Development
(3). Tableau Desktop Specialist
(4). SAS Certified Specialist: Base Programming
(5). Data Science Professional Certificate