

WILLIAM GEMBA

DATA SCIENTIST | DATA ANALYST | MACHINE LEARNING

Miami – Fort Lauderdale Area Cell: 917.588.1702

Email: gemba.william@gmail.com

LinkedIn: [William Gemba](#)

GitHub: [wgemba](#)

Web Portfolio: <https://wgemba.github.io/webportfolio/>

Summary

As an **experienced data analyst** with a **Master's degree in Data Science** and 2 years of experience in analyzing and remediating **ETL pipelines**, I bring a proven track record of **team leadership**, **complex problem solving**, and **effective solution implementation**. I am passionate about **machine learning** and am eager to leverage data to drive value for organizations. I am seeking a data analyst/scientist role that offers exposure to machine learning **model development**.

Skills

Python ■ SQL ■ Excel ■ Machine Learning ■ Data Mining ■ Predictive Analysis ■ Clustering Analysis ■ Forecasting ■ Regression ■ Classification ■ Artificial Neural Networks ■ Natural Language Processing (NLP) ■ Jupyter Notebook ■ Data Wrangling ■ Data Visualization ■ Feature Engineering ■ Exploratory Data Analysis ■ Pandas ■ NumPy ■ Scikit-Learn ■ Matplotlib ■ Seaborn ■ TensorFlow 2.0 ■ Natural Language Toolkit (NLTK) ■ Extract, Transform, Load (ETL)

Relevant Work Experience

Ernst & Young (EY) | Data Analyst, Consultant

2021 - 2023

- Spearheaded **data quality remediation** efforts for a major U.S. bank, ensuring compliance with the latest federal liquidity risk management standards.
- **Led** workgroups of **4-7 people**, using **agile methodology** and a **Jira** ticketing system to coordinate effective business rule implementation.
- Analyzed and remediated **complex ETL pipelines** by utilizing **Oracle SQL** to query and test data migration across multiple stages. This effort resulted in the detection and rectification of **millions of dollars** in imbalances stemming from incorrect ETL processes across **multiple lines of business** and **hundreds of attributes**.
- Collaborated with client stakeholders, including subject matter experts and data engineers, to address data lineage problems and institute new data quality controls.

Projects

Wine Recommendation System Using Unsupervised Content-Based Filtering

2024

- Performed exploratory data analysis of and created visualizations for a dataset of 130,000+ wine taster reviews.
- Utilized natural language processing (NLP) to parse through reviews and engineer a taste profile for each wine reviewed.
- Implemented a complex neural network architecture using TensorFlow 2.0 to predict the rating each user gave to each wine, achieving a final training loss of 0.003, and validation and test loss of 0.007.
- *Tools used: Python, Jupyter Notebook, Matplotlib, Seaborn, Scikit-Learn, Natural Language Toolkit, TensorFlow 2.0.*

U.S. SBA Loan Default Classification

2023

- Analyzed a dataset from the SBA consisting of 900,000 instances and 27 features, performing heavy feature engineering.
- Created a function to test 64 permutations of hidden layer width, selecting the model with the lowest training loss and validation loss. Top performing model was able to achieve an accuracy of 92%.
- *Tools used: Python, Jupyter Notebook, Matplotlib, Seaborn, Plotly, Scikit-Learn, TensorFlow 2.0.*

Education

Fordham University, MS Data Science

2021

Additional Accreditation: Advanced Certificate in Financial Econometrics and Data Analysis

Fordham University, BS Global Finance and Business Economics

2020

Certifications

Azure Data Scientist Associate (DP-100) | Microsoft

In Progress

Machine Learning Specialization | Stanford University / DeepLearning.AI (Coursera)

2023