Ernie Maldonado

☎ 412-304-4020 ✉ ernie.maldonado.data@gmail.com Webpage LinkedIn Github

Work Experience

Consumer Protection Big Data and Machine Learning Consultant — Oct 2023 – Dec 2023

Bank of America, Tampa FL — *Hybrid*

- Evaluated the entire fraud Hadoop ecosystem. The resulting document bench-marked current practices against best practices through data driven analysis, aiming to improve the performance, reliability and quality of the data across the entire cluster.
- Stack used: **Impala & Hive** for data manipulation,**Python** for data analysis.

Sanctions Machine Learning Engineer — May 2022 – Oct 2023

Citigroup, Tampa FL — *Hybrid*

- **ML development:** Automated human manual review of around 500,000 monthly alerts by developing multiple machine learning pipelines (Random Forest, Gradient Boosting, XGBOOST) to determine if customers flagged as a legal risk are in fact the person reported on sanctioned lists. This process included pipelines design, experiment design, data extraction, transformation and creation of over 50 different NLP features, machine learning model development, validation and documentation.
- **Leadership:** Supervised the development, implementation and monitoring of multiple machine learning models.
- **ML documentation:** Created documentation under Citi's the Model Risk Management framework to validate the implementation and monitoring of multiple machine learning models in accordance with financial regulations such as OCC 2011-12 and FR 11-7, fair lending etc.
- **Process Improvement:** Reduced the number of transactions to be dispositioned in half by discovering through data analysis that Citi's internal transactions were screened by two systems instead of one.
- **System Architecture:** Created an analysis showing that switching to a big data platform the transaction processing time would be reduced by 90% and the sophistication of the screening algorithms could be increased by utilizing higher level programming languages available in Hadoop.
- **Stack used:** **SQL** for data extraction, **Pandas** for data transformation, **Python** for model development, **Scikit-Learn, Torch, XGBOOST** for model creation, **FastAPI** for model deployment

Data Scientist Fraud Detection Machine Learning — May 2021 – Apr 2022

Certegy Payments, Tampa Florida — *Hybrid*

- **Machine Learning:** Developed machine learning models using both Python's PyTorch, Scikit Learn, TensorFlow and SAS Viya doubling the check fraud detection rate by developing supervised machine learning models to predict check fraud in Walmart stores. I used random forest, gradient boosting and logistic regression models that have improved the KS statistic from between 0.05 – 0.3 to 0.6 – 0.8.
- **Data engineering:** Developed a new machine learning pipeline for the organization by building machine learning features in Oracle SQL, made the models available through an API in AWS to receive the calls, routing the calls to the correct machine Learning model, scoring the transaction with the appropriate model using the features that had been developed in Oracle. Features were both calculated in real time and historical stored in tables.
- **Stack used:** SAS Viya, Oracle SQL, Oracle OML, Python, PyTorch, Scikit Learn, TensorFlow.

Sr. Data Scientist — Feb 2021 – Apr 2021

Nielsen, Tampa FL — *remote*

- **Software development:** Maintained and develop Java code to accurately adjudicate viewership to TV programming and advertisement. Developed new features for the code base to accurately adjudicate TV program viewership.
- **Stack used:** CodeHub, Jupiter Lab, Git Lab, git, S3, Java.

Cybersecurity Big Data and Machine Learning Engineer — Apr 2018 – Feb 2021

Citigroup, Tampa FL — *Hybrid*

- **Machine Learning:** Developed from scratch a Machine learning pipelines in production that supervised the online footprint of 300,000 employees and over 5TB of data on a weekly basis. The model used multiple metrics to measure employee cyber activity in relation to employees in similar roles, such as employees reporting to the same manager or employees on the same organizational role. This analysis was conducted by taking metrics such as data sent out of the corporation and system access and normalizing it in relation to the employee's peers. Other metrics such as number of connections from outside of the country and outside of the typical location where also included. Finally, the metrics were incorporated in a k-means cluster to determine clusters of typical activity and cluster of unusual activity. The clusters of unusual activity were reported for follow up by teams in charge of cyber monitoring. The data ingestion was developed in Hadoop scoop to extract data from databases, Python for data parsing, Hive for data transformation and Spark for data analytics and machine learning modeling. The entire system was run on a Hadoop Oozie scheduler. Features were created with Hive and the model was developed and run on Spark. Smaller models were also developed using PyTorch, Scikit Learn, TensorFlow.
- **Data Engineering:** Developed data ingestion pipelines for over a dozen data sets into Hadoop. The pipeline involved creating data imports using either Apache Scoop or Flume, python parsing, and hive data transformations. I also developed a monitoring system of the proper flow of all data feeds to the Cyber Security Fusion Center (270TB of data, Hadoop). Data pipelines were built using Oozie as the scheduler for Flume or Sqoop jobs to move data into Hadoop, python parsers and Hive transformations and Spark analytics to create the final tables that are accessed by users.
- **Stack used:** PySpark, PyTorch, Scikit Learn, TensorFlow, Hive, Oozie, Python, Linux, Unix shell scripts, SQL, Jira, BitBucket, Hadoop.

Quality Measures Analyst — Jan 2018 – Apr 2018

Health Services Advisory Group HSAG, Tampa FL — *On-Site*

- **Statistical analysis:** Used SAS to implement a statistical analysis developed by a team of statisticians to measure population disparities across multiple health metrics. Developed statistical analysis based on research papers.
- **Data Engineering:** Ingested multiple data sets from databases and flat files turning them into SAS files for faster accessibility and creation of statistical analysis
- **Stack used:** SAS, Tableau, SQL

Data Scientist Machine Learning — May 2016 – Oct 2017

Wellcare Health Plans (Acquired by Centene), Tampa FL — *On-Site*

- **Machine Learning:** Reduced hospital re-admissions by 60% by implementing a machine learning model using Scikit learn and XGBOOST in Python to flag people that were likely to be readmitted to the hospital. Hospital re-admissions are a significant problem for health insurance companies as they get penalized by Medicare when this happens.
- **Statistical Analysis:** Created a multilevel model to compare two different medical practice compensation strategies. Coefficients measuring fixed effects were associated with the different payment regimes and coefficients measuring random effects were those associated with the different medical practices.
- **Project management:** Earned company-wide recognition for being a key contributor in the implementation of a $1M per year capital project involving multiple external vendors, a big four consulting company and multiple business units. My role was to articulate teams that worked on different platforms to streamline a unified process (Informatica, Hadoop, SAS, Python and Hive).
- **Stack used:** PySpark, Hive, Python, R, Linux, SQL.

Sr. Decision Support Analyst — Jan 2015 – Apr 2016

Gateway HealthPlan (Acquired by Highmark), Pittsburgh PA — *On-Site*

- **Data Analytics & Reporting:** Directly worked with the CFO to implement data solutions of high visibility and financial impact that overcame long standing analytics and reporting problems. The reporting improvements were done by reviewing the entire accounting data reporting to identify reconciliation gaps between internal systems and government data. I recommended and implemented improvements to the ETL process that allowed for proper reconciliations between insurance claims data and accounting data.
- **Data Engineering:** Developed and implemented data pipelines from various sources (vendors, databases, flat files, etc) and combine them by using SAS to create data layers that are useful to different teams. Extracted and exchanged (in-out) data with vendors for analytical purposes.
- **Statistical Analysis:** Developed a ARIMA time series forecasting model using R to predict the cash flows of the Medicare programs. This used previous data as well as input from predicted revenue from data analytic models.
- **Management:** Managed and trained a team of three computer scientist to implement data solutions across the organization. The work of this team covered process improvement, automation, vendor performance evaluation, data layering, data retrieval, reporting and statistical analysis. Supervised an on premise SAS administrator and a team of SAS remote administrators (external vendor) to keep the SAS platform operational for the entire organization.
- **Stack used:** Python, R, SAS, SQL.

Database Analyst — Aug 2012 – Oct 2014

UPMC Health Plan, Pittsburgh PA — *On-Site*

- **Data engineering:** Developed ETL process that resulted in the creation of new reporting metrics within data layers that were consumed across the entire corporation for reporting. This processes was done in a combination of Oracle SQL and SAS.
- **Machine Learning:** Using Python and scikit learn I developed and implemented a machine learning model (Random Forest) to segregate members that caused recurrent high cost from members that caused occasional excessive cost. Earned the 2014 ACES award for excellence in service and process improvement. This award is only given to the top 1% of the employees at UPMC.
- **Software Development:** Developed and implemented a Java application to automate the creation of PDF documents by various teams.
- **Reporting:** Developed and maintained data analytic processes and reporting that went out to external partners (i.e. Doctors' offices). Developed analytic metrics and dashboards to measured internal and external clients doctors and hospitals' performance.
- **Stack used:** SAS, Java, SQL, Python.

Project Manager — Jan 2011 – Aug 2012

University of Pittsburgh, Pittsburgh PA — *On-Site*

- **Management:** Managed all aspects of two grant $1.5M funded research projects. Completed all data collection ahead of schedule and under budget. The projects were featured on local TV for their positive impact on the community and I was the interviewee for the program. Hiring personnel (about 20 employees), training, supervision, participants recruitment, data collection and performance reporting.
- **Stack used: SQL, R, SAS**

Projects

NameGen: Name generator using Generative AI — Source Code

- A fast and succinct name generator that can suggest previously unknown names based on a list of names currently used
- The model was created by creating a directed graphs that has as edges the probability of transitioning between any pair of letters at each position of a name. Then by using those probabilities the model samples letters at each position using a multinominal distribution.
- The project is showcased as CLI interactive mode both running the entire pipeline and using pickle to persist the graph used for faster results.
- The project is also showcased as a local fast API deployment and under a dockerized deployment of the fast API version to show case potential alternatives for deploying the model in a production environment

E-Commerce Fraud Detection — Source Code

- Three tree based machine learning models tuned to detect fraud on an E-Commerce dataset.
- This project showcases data analysis, feature engineering and model tuning for Random Forest, Isolation Forest and XGBoost applied to fraud detection.
- For XGBoost the hyperparameter tuning is done using Optuna.
- Finally, pipelines along joblib are used to save the models for deployment.

Education

Duquesne University, Pittsburgh PA USA — 2008 - 2011

Masters in Computational Mathematics — *3.9 GPA*

KDI School, Seoul (Relocated to Sejong City), S. Korea — 2004 - 2005

Masters in Public Policy — *3.9 GPA*

Universidad de los Andes, Bogota, Colombia — 2001 - 2003

Masters in Economics — *3.7 GPA*

Universidad de los Andes, Bogota, Colombia — 1996 - 2001

Bachelors in Economics — *3.7 GPA*

Certificates

AWS Certified Cloud Practitioner **2022**

Tensor Flow Specialization **2020**

Deep Learning Specialization **2018**

Machine Learning Specialization **2017**

Machine Learning Certificate **2015**

Technical Skills

Languages: Python, Java, SAS, Matlab
Machine Learning: Scikit-Learn, Torch, TensorFlow, SAS Viya, PySpark MLlib, PySpark ML
Big Data: Apache Hadoop, Apache Hive, Apache Oozie, Apache Impala, Apache Sqoop, PySpark
Frontend: HTML, CSS
Cloud & Databases: AWS, Oracle DB, MS SQLServer, MySQL, Hadoop
Developer Tools: Git, GitHub, Gitlab, BitBucket, Postman, Docker, FastAPI, Flask

Publications

Participatory Assessment of the Health of Latino Immigrant Men in a Community... **2013**
Journal of Immigrant and Minority Health

Bayesian Regression Inference Using a Normal Mixture Model **2012**
Masters Thesis

Globalization: through political networks **2007**
Revista Colombiana de Filosofía de la Ciencia

Anti-Drug Policies: On The Wrong Path To Peace **2006**
CEDE working papers series

Farc Terrorism in Colombia. A Clustering Analysis **2004**
CEDE working papers series

El Acertijo de la Reforma Politica en Colombia **2003**
Semana Magazine

Cuidado con el Umbral en la reforma política **2003**
Semana Magazine