Hernan Maldonado ● 412-304-4020 ● ernie.maldonado@tutanota.com

VP Consumer Protection (Machine Learning), Bank of America, Tampa FL *Oct 2023 – Dec 2023*

Data & System Architecture: 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. I made recommendations to change the ingestion monitoring metrics by developing processes and dashboards to monitor, detect and alert on data differences between the system of record (SOR) and the target Hadoop landing area. Additionally, I recommended to increase frequency on file compression processes to avoid the existence of small files as they reduce query efficiency. I recommended to synch the data partitioning to match the most used fields in queries with fields that were good for partitioning, such as dates. I also recommended to make increased efforts in correctly implementing data typing in tables such that memory was better allocated and queries would avoid data casting at run time. The analysis was done through impala queries, along pythons scripts and python analytics (pandas, matplotlib, etc).

Tools: Hadoop Impala.

Systems Engineer (Machine Learning), Citigroup, Tampa FL *Apr 2022 – Oct 2023*

Machine Learning: 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. Additionally I have supervised the development, implementation and monitoring of multiple machine learning models. This work was done using a standard python data science stack including PyTorch, Scikit Learn, TensorFlow, pandas, numpy, etc.

System Architecture: I recommended to switch the sanctions platform from an Oracle database environment to a Hadoop based environment. This recommendation was based on three core considerations: first, the sanctions system is a batch processing system and not a real time system, making it easier and faster to process the large datasets of transactions and customers using Hadoop. Second, the Oracle databases were not effective to store historic large transactions data which were often needed for regulatory reporting, analytics and internal reporting. Data was kept to a maximum of one year due to database size increase when five years were needed. Third, the Hadoop ecosystem is an inexpensive system to store historical transactions and customer data because it is a shared platform at Citi, however databases were still recommended to be used as back-end for human dispositioning through user interfaces (UIs). Additionally, transitioning to Hadoop was smooth as the entire dispositioning process is Java based which runs natively in Hadoop. Finally, Hadoop's parallel processing offered an increased computational capacity to perform complex string comparisons during the screening system, such as multiple name spellings, addresses, etc.

Process improvement: The transaction dispositioning logic was redundantly evaluating transactions that were between Citi clients. This logic accounted for most of the transactions being screened, Citi's customers are already screened against sanctions and non-sanctions lists. I recommended to waive this process to avoid screening redundancy resulting in a reductions in about half of the transaction screened.

Tools: Python, PyTorch, Scikit-learn, TensorFlow, SQL

Data Scientist, Certegy, Tampa FL ***May 2021 – Apr 2022***

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.

Tools: SAS Viya, Oracle SQL, Oracle OML, Python, PyTorch, Scikit Learn, TensorFlow.

Sr. Data Scientist, Nielsen, Tampa FL ***Feb 2021 – Apr 2021***

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.

Tools: *CodeHub, Jupiter Lab, Git Lab, git, S3, Java.*

Big Data Engineer, Citigroup, Tampa FL ***Apr 2018 – Feb 2021***

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.

Tools: PySpark, PyTorch, Scikit Learn, TensorFlow, Hive, Oozie, Python, Linux, SQL, Jira, BitBucket, Hadoop.

Quality Measures Analyst, Health Services Advisory Group (HSAG), Tampa FL ***Jan 2018 - Apr 2018***

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.

Tools: SAS, Tableau, SQL

Data Scientist, WellCare Health, Tampa FL ***Jun 2016 – Aug 2017***

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).

Machine Learning: Reduced hospital readmissions 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 readmissions 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.

Tools: PySpark, Hive, Python, R, Linux, SQL.

Team Manager, Sr. Decision Support Analyst, Gateway Health, Pittsburgh PA ***Jan 2015 – May 2016***

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.

Tools: Python, R, SAS, SQL.

Database Analyst, UPMC Health Plan, Pittsburgh PA ***Aug 2012 – Oct 2014***

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.

Tools: SAS, Java, SQL, Python.

Project Manager, University of Pittsburgh, Pittsburgh PA ***Jan 2011 – Aug 2012***

Management: Managed all aspects of two grant 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, supervise study participants recruitment, data collection and performance reporting.

Education & Training

M.Sc. Computational Mathematics, Duquesne University	2011
Masters in Public Policy, KDI School, Seoul, S. Korea	2005
M.A. in Economics, Universidad de los Andes, Bogotá, Colombia	2003
B.A. in Economics, Universidad de los Andes, Bogotá, Colombia	2001

Certificates

AWS Certified Cloud Practitioner	2022
Tensor Flow Specialization	2020
Deep Learning Specialization	2018
Machine Learning Specialization	2017
Machine Learning Certificate	2015