Theophilus Siameh

Data Scientist - AC NIELSEN

Tampa, FL - Email me on Indeed: indeed.com/r/Theophilus-Siameh/17775d78af10f39f

WORK EXPERIENCE

Data Scientist

AC NIELSEN - Saint Petersburg, FL - January 2017 to Present

Using best practices to understand the data and develop statistical, machine learning techniques to build models that address business needs.

• Writing and maintaining complex SQL queries on large databases including Netezza and SQL server databases.
• Utilizing tools such as Python, Tableau, Spotfire, SAS, etc. to perform complex data analysis.
• Creating data visualizations to provide management insight to complex business problems.
• Interacting directly with the client to understand its needs.
• Providing explanation of products and methodologies directly to the client in a format that meets their level of understanding.
• Using spark and databricks to analyze large volumes of data and push results to S3 for further processing.
• Used Spark, EC2 and S3 to interact with large datasets and automate python scripts.
• Automating Nielsen Reports using custom python scripts I developed.
• Weekly mentoring and training of Nielsen employees on how to use Apache Spark on Databricks.
• Moving flat files from one system to another using UNIX and run client enquiry analysis.

Data Scientist

FACORNE TECHNOLOGIES - Mountain Lakes, NJ - June 2016 to November 2016

Provided technical leadership in a team that designed and developed analysis systems to extract meaning from large-scale data.
• Developed pipelines to analyze large simulation datasets combining my own Python code.
• Analyzing large-scale data and analytics using advanced statistical and machine learning models.
• Developed machine learning models relying on decision trees, random forest, logistic regression, and other machine learning algorithms.
• Writing Scoop Jobs to Import/Export data to-and-from Hadoop and MySQL.
• Design and develop HDFS-based solutions in Hive and Pig
• Programming using Hive QL to generate reports
• Programming using Scala, Spark and Map Reduce framework on top of Hadoop.

Graduate/Research Assistant

DEPT. OF MATHEMATICS & STATISTICS - EAST, TENNESSEE, US - August 2014 to May 2016

Built a strong portfolio of clients that leveraged department statistical consulting services by manipulating and analyzing large volumes of data sets using PROC SQL, SAS Enterprise Miner,
Excel, SAS and JMP to conduct statistical analysis.
§• Developed techniques of visualizing high dimensional data sets using Manifold learning techniques as well as dimensionality reduction using PCA.
§• Empowered a business client to better project the potential of mortgage default by using logistic regressions.

§• Led two weekly lab sections for 70 students including creating lecture slides, preparing weekly presentations of material, grading, answering questions about homework and other course material, and instructing students in the use of SPSS for data analysis

Data Scientist Intern

RADICAL SHIPPING & LOGISTICS - May 2012 to August 2012

Work closely with various teams across the company to identify and solve business challenges utilizing large structured, semi-structured, and unstructured data in a distributed processing environment.
§• Retrieved stored data in MySQL, analyzed the data using the Python library scikit-learn
§• Utilize analytical applications like SAS to identify trends and relationships between different pieces of data, draw appropriate conclusions and translate analytical findings into risk management and marketing strategies
§• Built a logistic regression model end-to-end to predict members' likelihood to convert to a more profitable plan in order to prioritize marketing efforts.

Data Analyst

UNIVERSITY OF GHANA - August 2009 to December 2011

Exemplified excellence in performing various statistical analysis and other analysis tools to examine structured, unstructured and semi-structured data.
§• Effectively managed big data environments using SQL.
§• Correctly performed statistical analysis methods.
§• Customized new modeling techniques and other innovations as needed to complete projects.

DATA SCIENCE SIDE PROJECTS
Project #1: Analyze Loan Dataset

• Built a predictive model to predict interest rate on loan using machine-learning algorithm like Random Forest,

Ridge Regression.
• Data cleaning for missing and variable selection techniques were applied.

Project #2: Analyze Movie Ratings Using Apache Spark

• Get the user who has rated the most number of movies, Get the user who has rated the least number of movies

• Get the count of total number of movies rated by user belonging to a specific occupation;

• Get the number of under age users.

• Get the movies with the highest ratings.

Project # 3: Analyze San Francisco Crime Data

• Predict the category of crime committed using machine-learning algorithms like Random Forest, Gradient

Boosting, Naïve Bayes, KNeighbors, Decision Trees and Extra Tree Classifiers.

• Feature engineering on the dataset to determine the number of variables important for model building.

• Visualization of the dataset using manifold learning techniques like: PCA, Spectral Embedding and Isomap.

• Apply feature selection techniques to data; compute test accuracy and using ROC curve to compare each

machine learning algorithms under different model complexities.

Project # 4: Sentiment Analysis of Twitter Data using Apache Spark

• Streaming live tweets to find popular insights, most popular hashtags using spark streaming API

• Performed text preprocessing using lemmatization, stemming, stopwords removal, TF-IDF transformation, and

word2vec

• Applied SVM, Naïve Bayes, and ridge logistic regression to predict the positive/negative sentiments of tweets.

Project # 5. Multivariate Analysis of Astronomical Data Using R

• Employed 4 visualization techniques, namely PCA, factor analysis, multidimensional scaling, and clustering to

explore data

• Applied SVM, ridge logistic regression, linear/quadratic discriminant analysis, random forest, and adaboost to

classify astronomical data into five categories and compared the results of different methods.

EDUCATION

Bachelor of Science in Computer Science and Mathematics

University of Ghana - Accra, GH
2009

Master of Science in Mathematics and Statistics

East Tennessee State University - Johnson City, TN

SKILLS

SQL (4 years), STATISTICAL ANALYSIS (4 years), MACHINE LEARNING (3 years), LOGISTIC REGRESSION (3 years), APACHE (2 years)

LINKS

https://www.linkedin.com/in/theophilus-siameh-793a8626

ADDITIONAL INFORMATION

QUALIFICATIONS PROFILE

Strategic Data scientist and Analyst with an impressive arsenal of technology, team building, and communication skills that have consistently proven valuable in private sector and higher education settings and the ability to mine hidden gems located within large sets of structured, semi-structured and unstructured data. Able to leverage a heavy dose of mathematics and applied statistics with visualization.

• Modeling: Proven success developing strong research-based regression, predictive, machine learning, deep learning and distributed data processing.

• Analysis: Exceptional big data and advanced analytics experience that has drastically improved the intelligence and profitability of numerous small business clients.

CORE TECHNOLOGIES:

Programming: Scala, Java, Python (IPython, Matplotlib, Numpy, Pandas, Scikit-learn), C#, HTML, MySQL, SQL, PostGreSQL, SAS, SAS Macro Programming, Minitab, MatLab, SAS Enterprise Miner, SPSS, Spot Fire, Tableau, R, Microsoft Office: Excel & Access, UNIX command line.

Hadoop AWS (S3, EMR, EC2, Redshift), Apache Hadoop & MapReduce, Apache Hive, Apache Stack: Sqoop, Apache Pig, Apache Flume, Apache Spark on Databricks, Apache Cassandra and Apache Kafka.

Machine Classification, Clustering, Decision Trees, KMeans, Artificial Neural Network, Random Learning: Forest, Logistic Regression, feature engineering, support vector machines, Graph Theory, Text Analytics and Deep Learning.

Predictive modeling, Applied Multivariate Statistical Analysis, Time Series Analysis,
Courses:
Machine Learning, Operations Research, Mathematical Statistics, Regression Analysis,
Big Data: Storage, Analytics and Visualization, Applied Statistics, Applied Stochastic modeling, Principles of data mining, Hadoop Fundamentals.