Sofia **Dutta**

Senior Data Scientist
(443) 554-4170 | sofia.dutta17@gmail.com | GitHub | LinkedIn | Personal homepage | YouTube

WORK EXPERIENCE

Senior Data Scientist @ NewWave Telecom & Technologies, Inc. **Jan 2021 – Present**
Windsor Mill, MD, USA

- Working on the Medicaid Data Quality Analytics (MDQA) project. Performing data quality analysis on millions of healthcare records from Centers for Medicare & Medicaid.
- Was able to successfully improve computation speed by 10-fold by deploying data analysis workflow in Google Cloud Platform (GCP) clusters and using Apache Spark for quality metrics computations.
- Utilizing Google Cloud Storage and BigQuery for storage and faster processing of large quantities of healthcare data.
- Training machine learning models using thousands of rules for quality computation metrics.
- Leading data-processing efforts, guiding new employees to quickly ramp up on data analytics and achieve project goals.

Data Scientist Intern @ NewWave Telecom & Technologies, Inc. **May 2020 – Dec 2020**
Windsor Mill, MD, USA

- Carried out data visualization tasks using LookML, Matplotlib, and Seaborn to present data quality outcomes from various quality computation metrics.
- Created an end-to-end architecture design and database schema design for the MDQA data quality analytics platform.
- Collaborated with the Product Owner and other engineers in creating mechanisms for generating fake training data using Python programming to test out the efficacy of machine learning algorithms used in the project.
- Carried out necessary DevOps tasks for setting up Big Data Analytics environment by configuring GCP environment to execute Python programs and connected the cloud infrastructure with Looker's dashboards for delivering computed results to be presented to customers.

Software Engineer @ Tata Consultancy Services **Nov 2010 – Feb 2018**
Kolkata, India

- Led the design, development, and delivery management of seven projects for clients of TCS.
- Created API interfaces using PL/SQL stored procedures for daily usage for clients of TCS.
- Carried out change based regression analysis and documented software functional specifications.
- Prepared test plans and executed system integration testing and user-acceptance testing.
- Ensured client systems were up in four hours after migration activities saving millions of dollars in potential revenue lost.
- Implemented scripts for data migration of a billion records while adhering to strict time SLA bounds.
- From 2013 - 2018, managed continuous integration and continuous deployment in production environments.

SKILLS

Coding languages	Python, Java, SQL, PL/SQL, T-SQL
Data Science tools	PyTorch, Sci-kit Learn, Apache Spark, Keras, Tensorflow, Hive, Hadoop, MLlib, Matplotlib, Seaborn library, Looker, LookML
Enterprise tools	Google Cloud Platform, Google Dataproc, Google Compute Engine, Google Cloud Storage, Google Cloud SQL, Google Big Query, Amazon Web Services S3, Azure Databricks, Oracle Global Human Resources Cloud, Oracle Talent Management Cloud, Oracle Financial Management
Back-end tools	Oracle Databases, PostgreSQL, Microsoft SQL Server, MongoDB, JSON
IDEs/Dev tools	Jupyter Notebook, Google Colab, PL/SQL Developer, Git

EDUCATION

University of Maryland, Baltimore County (UMBC), Baltimore, MD, USA
Master's in Data Science GPA: **4.0** **2019 – 2020**
Coursework: Practical Deep Learning, Machine Learning & Data Analysis, Big Data Processing, Databases: SQL, NoSQL

West Bengal University of Technology, Kolkata, India
Bachelor's in Computer Science GPA: **3.5** **2006 – 2010**
Coursework: Data Structures & Algorithms + Labs, Operating Systems + Labs, Artificial Intelligence + Labs

RESEARCH

Ebiquity Research Lab, UMBC | Graduate Student Researcher **Sep 2019 – May 2020**
Baltimore, MD, USA
Authored an Ontology for Smart Home Access Control, extending earlier research in Semantic Web. Developed an Android app for handling context-sensitive access control in a Smart Home Environment. Created YouTube videos for presentation to the National Institute of Standards and Technology. Published a paper at the IEEE Big Data Security 2020 conference.

MACHINE LEARNING PROJECTS

Master's degree capstone project using Natural Language Processing: ***QABot: A Chatbot for Open Question Answering Using Neural Networks*** - Built "QABot", a Chatbot using the sequence-to-sequence Deep Learning model that utilizes the Encoder Decoder Neural Network architecture combined with Attention Mechanism to answer user search queries. Created a model by training a Deep Neural using the PyTorch Deep Learning Framework. Used Recurrent Neural Network architecture that are better at dealing with text sequences. Used both Teacher Forcing and Auto-Regressive approaches for model training and Auto-Regressive approach for model evaluation. Used BERT (Bidirectional Encoder Representations from Transformers) for tokenization and combined Transformer and GPT-2 for model fine tuning.
Ref: https://sites.google.com/umbc.edu/data606/fall-2020/sofia-dutta

Best deep learning project exploring Image Processing: ***Image-to-Image Translation Using CycleGAN*** - Implemented CycleGAN for an image-to-image translation. Trained an unsupervised image translation model via the Generative Adversarial Network (GAN) architecture using unpaired collections of images from two different domains. CycleGAN has previously been demonstrated on a range of applications and I chose to perform object transfiguration with it. Transforming images of horses to zebras and then back from zebras to horses.
Ref: https://sofiadutta.github.io/datascience-ipynbs/pytorch/CycleGAN_Img_Translation_PyTorch_Horse2Zebra.html

Big-Data Analytics exploring Machine Learning Classification: ***Sentiment Classification with Twitter Stream Data*** - Worked on large real-time streaming data from Twitter. Performed analytics using PySpark and created visualizations. Created a MyClassifier sentiment classifier via Word2Vec model using Spark MLlib. Used PySpark Big Data tool and performed analytics driven by the 6Vs of Big Data.
Ref: https://sofiadutta.github.io/datascience-ipynbs/big-data-analytics/Using_MyClassifier_Twitter_Data_Sentiment_Classification_and_Big_Data_Analytics_on_Spark_Dataframe.html

Exploring Machine Learning with Keras: ***Comparison of Word2vec and Doc2Vec model driven Sentiment Analysis using SVM, LR, Keras CNN, Bidirectional LSTM with and without pre-trained Word and Document Embeddings*** - Worked on applying opinion mining or sentiment analysis via word embedding and document embedding models to carry out sentiment classification of user reviews. Performed classification on laptop product reviews from Amazon's website and movie reviews from IMDb's website.
Ref: https://sofiadutta.github.io/datascience-ipynbs/sentiment-analysis/Sentiment_Analysis_IMDB_Movie_Review.html

RELEVANT COURSEWORK

Practical Deep Learning **Spring 2020**
- Learned the PyTorch open-source library for machine learning and used Google Colab technology to work on machine learning problems like Image Classification, Sentiment Analysis, Object Detection, Transfer Learning, Natural Language Translation, Auto Regressive Text Generation
- Used Attention Mechanism for Natural Language Processing and Machine Translation
- Created a Denoising Autoencoder to reconstruct MNIST images of numbers
- Created a Wasserstein Generative Adversarial Network to generate MNIST images of numbers

Platforms for Big Data Processing **Fall 2019**
- Using Apache Spark performed Map-Reduce operations on streaming data
- Learned Big Data technologies like PySpark, Spark SQL, MLlib, Spark Streaming, Hive, Hadoop
- Worked on practical projects with large datasets
- Used NoSQL storage (MongoDB) to manage large datasets collected from Twitter Data APIs

Introduction to Data Analysis and Machine Learning **Spring 2019**

- Worked on practical machine learning and data analysis problems.
- Worked on end-to-end processing pipeline for extracting and identifying useful features that best represent data, applying machine algorithms, and evaluating their performance for modeling data.
- Learned machine learning APIs like Sci-kit Learn, Keras, Tensorflow.
- Learned machine learning algorithms like decision trees, logistic regression, support vector machines, convolutional neural networks, recurrent neural networks, bidirectional LSTM.