Sofia Dutta

Technical Leader, Software Engineer

(443) 554-4170 | sofia.dutta17@gmail.com | https://sofiadutta.github.io

WORK EXPERIENCE

Technical Leader @ NewWave Telecom & Technologies, Inc.   **Jan 2021 – Present**

Windsor Mill, MD, USA

- Designed the complete system architecture, database schema, and data workflow for our data quality analytics platform.
- Built the project's cloud infrastructure setup from scratch.
- Built scripts to process hundreds of millions of healthcare records from the Centers for Medicare & Medicaid.
- Created pre-processing scripts for consuming large batches of unstructured customer data.
- Developed Apache PySpark code to compute data quality metrics for customer data.
- Created Looker dashboard visualizations with drill-down options that "explains" why the data quality came out to a certain value.
- Due to the lack of large quantities of data, re-designed the machine learning goal of the project into a data quality "Explainable AI" system.
- Built a data quality system for State governments that allows them to understand where to improve their upstream data ingestion processes and helped them observe how that improves their data quality over time.
- Reduced costs at four levels of the project:
  - Pre-processing: Built scripts that brought down the data pre-processing time from ten days to a couple of hours.
  - Uptime of cluster: Analyzed causes of high cloud expenditure and deployed Apache Airflow workflow management platform scripts to automate resource uptime only during hours of usage. Reduced cost from thousands of dollars to a couple hundred.
  - Storage v/s data transfer: Performed the cost advantage analysis of using a Google compute engine with large storage versus using more network data transfer.
  - Partially working in local servers: Built a system that handled pre-processing in our local data farm to reduce project costs from over ten thousand dollars to a few hundred in a month.
- Quickly learned new technologies like Apache Airflow, Databricks, and Google Cloud Platform and guided team members in their technology ramp-up for the past 2 years and helped them in setting up data workflows in the cloud.
- Provided educational expertise and mentoring to junior team members.
- Built product feature lists with stakeholders, conducted system design sessions with other architects on the team and led code review meetings.
- Investigated root causes for customer-found defects. Carried out several customer demonstrations to help sell the product and handled rapid prototyping and solution building for ad-hoc requirements and last-minute feature requests from the customer.
- Advised management, business, and technical staff on the usage of specific technologies like Apache Airflow and Google Cloud Platform.

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 quality measures based on chosen computation metrics.
- Successfully improved computation speed by 10-fold by deploying data analysis workflow in Google Cloud Platform (GCP) clusters and using Apache Spark for quality metrics computations.
- 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.

Researcher @ Ebiquity Research Lab, UMBC   **Sep 2019 – May 2020**

Baltimore, MD, USA

- Built Ontology for Smart Home Access Control and 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 and published paper: "Context Sensitive Access Control in Smart Home Environments" at IEEE Big Data Security 2020 conference.

Software Engineer, Technical Leader @ Tata Consultancy Services   **Nov 2010 – Feb 2018**

Kolkata, India

- Led the design, development, and delivery of API interfaces using PL/SQL stored procedures for several projects of TCS.
- Carried out change based regression impact analysis, created software functional specifications, prepared test plans for several projects of TCS.
- Performed system integration, user-acceptance and performance testing and ensured client systems had very high uptime even when carrying out data migration activities. Saved millions of dollars in potential revenue lost to the client and was awarded for said effort by clients.

SKILLS

Coding languages	Python, Java, SQL, PL/SQL, T-SQL
Data Science tools	PyTorch, Sci-kit Learn, Apache Spark, Keras, Tensorflow, Hive, Hadoop, Looker, LookML, OpenCV
Enterprise tools	Databricks, Google Cloud Platform, Apache Airflow, Google Dataproc, Google Compute Engine, Google Cloud Storage, Google Cloud SQL, Google Big Query, AWS S3
Back-end tools	Google BigQuery Table, Oracle Databases, PostgreSQL, Microsoft SQL Server, MongoDB, JSON
IDEs/Dev tools	Jupyter Notebook, Google Colab, Git
Domain knowledge	Big Data Analytics, Deep Learning, Machine Learning

EDUCATION

Udacity Nanodegree: Deep Learning		October 2021

University of Maryland, Baltimore County (UMBC), Baltimore, MD, USA		
Master's in Data Science	GPA: 4.0	2019 – 2020

West Bengal University of Technology, Kolkata, India		
Bachelor's in Computer Science	GPA: 3.5	2006 – 2010

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 AutoRegressive 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/home/previous-semesters/fall-20/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_Amazon_Laptop_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, Transformer, Generative Adversarial Networks
- 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