Sofia Dutta

Tech Lead, Data Scientist, Data Engineer

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

WORK EXPERIENCE

Senior Data Scientist, Tech Lead @ NewWave Telecom & Technologies, Inc. Jan 2021 – Present

Windsor Mill, MD, USA

- Worked on the Imersis project, creating the end-to-end system architecture design, database schema design, cloud infrastructure setup and Looker data visualizations dashboards for the Imersis data quality analytics platform.
- Handled hundreds of millions of healthcare records from Centers for Medicare & Medicaid.
- Worked with unstructured customer data, and reduced data pre-processing time by a factor of ten.
- Leading data-processing efforts, guiding new employees to quickly ramp up on data analytics and achieve project goals.
- Built Apache Airflow data orchestrators for automated data analysis workflow for Imersis project.
- Built data analysis pipelines using Databricks data lake to further speed up analysis work.
- Built proof-of-concept solutions for customer demonstration to win project contracts on behalf of NewWave.

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.

Graduate Student 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, Tech Lead @ 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.

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	Deep Learning, Machine Learning, Big Data Analytics

EDUCATION

Udacity Nanodegree: Deep Learning — 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 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_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