Sofia **Dutta**

Data Scientist

[sofia.dutta@newwave.io](mailto:sofia.dutta@newwave.io) | (443) 554-4170

[GitHub](https://github.com/sofiadutta) | [LinkedIn](https://linkedin.com/in/sofiadutta) | [Personal homepage](https://sofiadutta.github.io/) | [YouTube](https://www.youtube.com/watch?v=DZq-oZ_Cv1g&list=UUEB2cqBZTym9WxWvGu2UA0w) | [Scholar](https://scholar.google.com/citations?user=xdlmrWYAAAAJ) | [Twitter](https://twitter.com/dutta_sofia)

WORK EXPERIENCE

Senior Data Scientist @ NewWave Telecom & Technologies, Inc.

Windsor Mill, MD, USA Jan 2021 – Present

Working as a Senior Data Scientist

o Working on the Medicaid Data Quality Analytics (MDQA) project, a next generation data science project focused on creating platforms for analyzing and improving the data quality of US healthcare systems. Implementing an end-to-end Data Science workflow from data acquisition, data processing, data integration, model creation, model validation, machine learning prediction to visual representation. Automating the collection of data using tools like Google Cloud Fusion to quickly aggregate data from various sources. Pre-processing of data via data cleansing, transformation, formatting to allow proper consumption by downstream analytics systems. Building scalable, distributed, fault tolerant, load balanced systems for analyzing huge quantities of data, typically referred to as Big-Data analytics, using Google Cloud Fusion tool, Google Cloud Storage systems and Apache Spark technologies. Creation of predictive models using machine learning, deep neural networks, and ensemble methods. Using knowledge of and skills with various cloud-based platforms and technologies like Google Dataproc, Google Compute Engine, Google Cloud Storage, Google Cloud SDK, Google Cloud SQL, Google Kubernetes Engine, Google Big Query, Google Cloud CDN, Google Dataflow, Google Cloud Run to analyze data quality of patients and presenting results using visualization platforms like Looker, Jupyter Notebook and Azure Databricks. Resolving business challenges in an efficient and timely manner. Effectively communicating, presenting the most relevant information, and collaborating with team leads and product owners to ensure a positive impact on the US healthcare system and US citizens' healthcare data. Thinking outside the box and researching ways of improving data security and enabling faster data lookups.

Data Scientist Intern @ NewWave Telecom & Technologies, Inc.

Windsor Mill, MD, USA May 2020 – Dec 2020

Worked as a Data Scientist Intern

As an intern on the Medicaid Data Quality Analytics (MDQA) project, applied data visualization skills like Matplotlib, Seaborn to present data quality outcomes from various quality computation metrics. Learned to use the Looker tool from Google as team was heavily using it. Created an end-to-end design and architecture of the data quality analytics platform. Created the database schema design for the MDQA project. Collaborated with product owner and other engineers on building the first version of the data quality platform for the MDQA project. Created mechanisms for generating fake training data using Python programming to test out efficacy of machine learning algorithms used in the project. Trained machine learning models using thousands of rules for quality computation metrics. Worked on improving computation efficiency by several orders of magnitude by implementing data analysis workflows in Google Cloud Platform (GCP) clusters and using Apache Spark for quality metrics computations. Exported results to Looker for creating intuitive dashboards visualizations. Worked with huge quantities of US healthcare data from Centers for Medicare & Medicaid. Created clusters in Google Cloud Platform and installed all required Big-Data analytics packages and software. Configured 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

Kolkata, India Nov 2010 – Feb 2018

Worked at Tata Consultancy Services (TCS) as a software developer.

o Worked for seven different projects for clients of TCS. Led the design and development of software requirements that came from the customer. Built API interfaces using PL/SQL stored procedures for daily usage for clients of TCS. Led meetings to capture requirements from multi-national clients like DHL UK, Staples USA, Hyatt USA, Kaiser-Permanente USA. Carried out change based regression analysis and documented software functional specifications for those changes. Prepared test plans and executed system integration tests and user-acceptance tests. Worked on data migration projects and ensured migration activities were completed in a time-sensitive manner millions to reduce the risk of revenue loss for the clients. Implemented scripts for data migration of over a billion records while adhering to strict time SLA bounds. Completed client data migration from legacy Oracle Apps (11i) to Oracle ERP Suite (R12). From 2013 - 2018, managed continuous integration and continuous deployment in production environments. Additionally, worked as a team lead ensuring engineers joining the team newly were brought up to speed quickly. As a requirement from the clients and due to needs of the projects, completed certifications for Oracle Apps technologies including, Oracle Global Human Resources Cloud 2017 Implementation Essentials, Oracle Talent Management Cloud 2017 Implementation Essentials, Oracle Global Human Resources Cloud 2016 Implementation Essentials, Oracle Advanced PL/SQL Developer Certified Professional, Oracle E-Business Suite 12 Financial Management Implementation Specialist: Oracle Receivables, Oracle PL/SQL Developer Certified Associate and Oracle SQL Developer Certified Associate.

SKILLS

Programming Python, Java, SQL, PL/SQL, T-SQL

Data Science tools PyTorch, Sci-kit Learn, PySpark, Keras, Tensorflow, Hive, Hadoop, MLlib, LookML, Azure AI

Enterprise tools Google Cloud Platform, Amazon Web Services S3, Oracle Applications

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

Baltimore, MD, USA Sep 2019 – May 2020

o Authored an Ontology for Smart Home Access Control by extending earlier research in Semantic Web.

o Developed an Android app for handling context-sensitive access control in a Smart Home Environment.

o Created YouTube videos for presentation to the National Institute of Standards and Technology.

o Published a paper at the IEEE Big Data Security 2020 conference.

PROJECTS

Master’s degree capstone project: QABot: A Chatbot for Open Question Answering Using Neural Networks

Best deep learning project: Image-to-Image Translation Using CycleGAN

PUBLICATION

Sofia Dutta et. al., “Context Sensitive Access Control in Smart Home Environments”, InProceedings, 6th IEEE International Conference on Big Data Security on Cloud (Big Data Security 2020), May 26, 2020, Baltimore, MD, USA.