Sofia Dutta

Data Scientist
github.com/sofiadutta • linkedin.com/in/sofiadutta • sofiadutta.github.io
(443) 554-4170 • sofiad1@umbc.edu

Education

University of Maryland, Baltimore County, Baltimore, MD *Spring 2019 – Fall 2020*
Master's in Data Science, **GPA: 4.0**
West Bengal University of Technology, Kolkata, India *Fall 2006 – Spring 2010*
Bachelor's in Computer Science, **GPA: 3.5**

Technical Skills

Programming Languages: Python, SQL, PL/SQL, T-SQL, Java
Data Science Tools: PyTorch, Sci-kit Learn, Apache Spark, Apache Hive, Apache Hadoop, MLlib, Keras, Tensorflow, LookML, Microsoft Azure AI
Development Tools: Docker, Jupyter Notebook, Google Colab, PL/SQL Developer, Git
Enterprise Tools: Google Cloud Platform, Amazon Web Services S3, Oracle Applications
Backend Tools: Oracle Databases, PostgreSQL, Microsoft SQL Server, MongoDB, JSON

Work Experience

NewWave Telecom & Technologies, Inc., Woodlawn, MD *May 2020 – Present*
Data Scientist Intern: Working on Data Science project building rules-based machine learning error-detection models to carry out data quality analysis tasks on Centers for Medicare & Medicaid Services (CMS) healthcare claims data. Instrumental in building a data exploration platform using cloud-based tools like LookML, software from Looker. Created visualizations of customer data to provide the best actions when choosing a data quality improvement algorithm.

Ebiquity Research Group, UMBC, Baltimore, MD *Sep 2019 – May 2020*
Student Researcher: Performed research in Semantic Web, Context-based Access Control in Smart Homes and published the following paper at IEEE Big Data Security 2020 conference:
Sofia Dutta et. al., "Context Sensitive Access Control in Smart Home Environments", InProceedings, *6th IEEE International Conference on Big Data Security on Cloud (BigDataSecurity 2020),* May 26, 2020, Baltimore, MD, USA. doi: 10.1109/BigDataSecurity-HPSC-IDS49724.2020.00018

Tata Consultancy Services (TCS), Kolkata, India *Nov 2010 – Feb 2018*
Software Developer, IT Analyst:

- Led a team of developers in preparing PL/SQL stored procedures.
- Designed, developed, and tested API interfaces for PL/SQL stored procedures.
- Prepared functional specification documents.
- Performed requirement and change based regression analysis.
- Prepared test plans and performed system integration testing and user-acceptance testing.
- Worked on Oracle Fusion HCM (Core HR) functionalities for clients.
- Wrote scripts for managing customer data migration task of over a billion records.
- Completed client data migration from legacy Oracle Apps (11i) to Oracle ERP Suite (R12).
- Managed continuous integration and continuous deployment to production environments.

Graduate School projects

Capstone project: Retraining a BERT-based NLP model for Chatbot using PyTorch *Fall 2020*
Image to image translation using CycleGAN *Spring 2020*

- Used CycleGAN to train an unsupervised image translation model via the Generative Adversarial Network (GAN) architecture using unpaired collections of images from two different domains.

- Performed object transfiguration on couple of datasets:
  - Image transformation from horse to zebra & reverse translation from zebra to horse.
  - Image transformation from orange to apple & reverse translation from apple to orange.

Big Data Twitter Stream Sentiment Analysis *Fall 2019*

- Learned to use Twitter data APIs. Collected tweets, then cleaned and pre-processed using Python's libraries.
- Used Apache Spark streaming API to collate data and applied Map-Reduce operations to track trending cryptocurrency topics.
- Visualized sentiment movements on trending cryptocurrency topics to find out if "humans on Twitter" are feeling positive about Bitcoin's future or are they feeling negative.
- Visualized geographic distribution of tweets for trending cryptocurrency topics
- Created my own sentiment classifier by training on popular 1.6 million tweet data set
- Compared my classifier and well-known social media classifier. Found we agreed 7/10 times.

Sentiment Analysis on user review datasets from Amazon and IMDb *Spring 2019*

- Compared performance of traditional machine learning algorithms like support vector machines, logistic regression, versus neural networks created using Keras CNN, Keras Bidirectional LSTM to empirically prove neural networks are better at sentiment classification

Data characterization projects using Python Sci-Kit Learn *Spring 2019*

- Analyzed Baltimore City Employee Salary data to prove there is no income inequality in Baltimore City Government
- Studied New York City Film Permits data to figure out top filming locations for popular movies
- Combined two different datasets from the New York City Fire Department and showed that it is possible to use data analysis techniques to find high impact incidents

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.

Introduction to Data Science *Spring 2019*

- Performed data analysis projects using supervised and unsupervised machine learning packages.
- Worked on data collection, storage, transformation, cleaning, analysis, and visualization.