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

Education

University of Maryland, Baltimore County, Baltimore, MD *Spring 2019 – Fall 2020*
Master's Professional Studies, Data Science, **GPA: 4.0**
West Bengal University of Technology, Kolkata, India *Fall 2006 – Spring 2010*
Bachelor of Technology, 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, MLlib, Keras, Tensorflow, LookML
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 machine learning models and carrying out data analysis tasks on Centers for Medicare & Medicaid Services (CMS) healthcare claims data. Building data exploration platform using cloud-based tools like LookML, a software from Looker.
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 a paper at IEEE Big Data Security 2020 conference.
Tata Consultancy Services (TCS), Kolkata, India *Nov 2010 – Feb 2018*
Software Developer: Led a team of developers in designing, developing and testing PL/SQL stored procedures. Built API interfaces for PL/SQL stored procedures. Prepared functional specification, requirement and change based regression documents, and test plans for performing system integration testing and user-acceptance testing. Completed client data migration from legacy Oracle Apps.

Graduate studies projects

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: horse to zebra & orange to apple.
Big Data Twitter Stream Sentiment Analysis @ UMBC *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 and geographic distributions on trending cryptocurrency topics to find out if "humans on Twitter" are feeling positive or negative about Bitcoin's future. Created my own sentiment classifier by training on popular 1.6 million tweet data set.
Sentiment Analysis on user review datasets from Amazon and IMDb @ UMBC *Spring 2019*
Compared performance of traditional machine learning algorithms like support vector machines, logistic regression, versus neural networks using Keras CNN, Keras Bidirectional LSTM to empirically prove neural networks are better at sentiment classification
Data characterization projects using Python Sci-Kit Learn @ UMBC *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