Anant Tripathi

Data Scientist
+1 7067140717 | tripathianant1203@gmail.com | Github | Linkedin

EXPERIENCE

Safekeep, NYC, USA 09/2020 - 11/2020
Role: Data Engineer/Data Scientist

- Performed Data Collection, Data Cleaning and Data Visualization using **SQL queries**, **Pandas, Python** and extracted key statistical findings to develop business strategies.
- Created data pipeline using **AWS Glue ETL** and Spark. Source as **S3 Data Lake** and target being **Amazon RDS.**
- Developed interactive **dashboard** using **AWS Quicksight.**

Liberty Mutual Insurance, GA, USA 08/2019 - 09/2020
Role: Data Scientist

- Performed Data Collection, Data Cleaning and Data Visualization using **SQL queries**, **Python Libraries, Deep Feature Synthesis** and extracted key statistical findings to develop business strategies.
- Analysed and grouped products into different **clusters** based on product description, purchase and historic data using techniques such as **k-means clustering**.
- Employed **auto-classification** of products based on customer databases by drawing inferences from products ordered together. Created **ETL program** for supporting Data Extraction, transformations and loading using **SQL**.
- Used **Multiple Linear Regression, Decision Tree Regression and ensemble learning like Random Forests & Gradient Boosting Machine** to train the model & the models were **optimized** using **Grid Search** and the predictions were made on the test set using each trained model.
- Tackled a **highly imbalanced** Fraud Dataset using **under-sampling, over-sampling** with **SMOTE** and cost sensitive algorithms with **Python**.

University of Georgia, Athens, USA 08/2018-07/2019
Role: Graduate Research/Teaching Assistant

- Created different types of deep learning models on EPS dataset to predict EPS of Financial years. **Using Python, Tensorflow, Pytorch, Pandas.** Focused on different types of **forecasting methods** to get reduced **Loss(MSE)** in comparison to analysts.
- Developed and deployed dashboard to visualize the dataset and for **ANN** and **IANN** model creation and testing. **Using Tensorflow, Plotly Dash, Flask, Heroku, AWS.**
- Taught Systems Programming Lab with 30+ students 3+ hours a week by helping students with better understanding of C++/Unix concepts. Developed grading scripts in Bash for coding assignments of students to generate final grade.

Tata Consultancy Services Pvt Ltd, India 03/2016–12/2017
Role : Data Scientist

- Responsible for all stages in the modeling process, from **collecting, verifying, & cleaning data to visualizing model results, presenting results, and making client recommendations.**
- Implemented **market basket algorithms** from transactional data, which helped identify ads clicked together frequently. Discovering frequent ad sets helped unearth Cross sell and Upselling opportunities and led to better pricing, bundling and promotion strategies for sales and marketing team.
- Used **SQL** to create Statistical algorithms involving **Multivariate Regression, Linear Regression, Logistic Regression, Random forest models, Decision trees** for estimating the risks.
- Developed **Python code** for data analysis (also using NumPy and SciPy), Curve-fitting. Worked on text mining, string manipulation, API user interface. Created various types of **data visualizations using Python and Tableau.**

TECHNICAL SKILLS

Programming Languages: Python, R, Apache Spark, GitLab, GitHub.
Machine Learning: OpenCV, Numpy, Scikit-Learn, Pandas, Keras, Tensorflow, NLTK, Spacy, Gensim.
Cloud Technology: Docker, Google cloud platform (GCP), Amazon web service (AWS), AWS Sagemaker, Auto ML, Data Proc, DataFlow, Pub/Sub, Google Data Studio, AI Platform, AWS S3 Bucket, Google Bucket, AWS Glue.
Data Analyst: AWS QuickSight, Power BI, MatplotLib, Seaborne
Databases: MySql, Oracle 11g, MS SQL, MongoDB.

EDUCATION

Master's in Computer Science (12/2019) University of Georgia, Athens, GA.
Bachelor's in Information Technology (06/2015) Uttar Pradesh Technical University, Noida, India.

COMPUTER VISION PROJECTS

Applying Deep Learning for Large-scale Quantification of Urban Tree Cover:

- **Pre-processed** image data by **resizing**, converting it into black and white and **removal of noise** using **opencv and numpy.** Created the **mask** using a json file by keeping the vegetation cover and removing the other object.
- Performed **multi-modal analysis** over cityscape dataset making use of deep learning models such as **Unet, Mobilnet_v2, Xception net.** Implemented transfer learning technique on the pre-trained model usings weights of image net. Python, keras, Google cloud platform were used to develop the code.

Cilia Segmentation (Research project under Dr Shannon Quinn University of Georgia)

- Given the frames of video of the cell body the problem was to predict the cilia in the frames.
- Analysed the moving behaviour of cilia using **optical flow, beat frequency, fluctuation variance** which were used as a feature for the model analysis. **Using Python**
- Analysed the static cilia using image processing technique such as high and low pass filter
- Trained the model using Keras on a **U Net convolutional neural network** to make predictions.

OTHER PERSONAL PROJECT

Microsoft Malware Classification (Dataset - 500 GB data)

- Created a **Large scale document classifier** to classify malware families based on **feature extraction** from **byte code and op-codes** under 9 categories.Employed **logistic regression** and **random forest algorithms** to develop classifiers with **Spark-ML library and GCP.** Performed **NLP techniques** such as **stop word removal** and **Word Embeddings** using **Spark-ML library**, **NLTK and Python.**

NLP with Disaster Tweets (Kaggle competition)

- Analysed and visualised data distribution for classes such as word length, number of characters and average word length in tweets. Matplotlib and seaborn on python.
- Developed python script to pre processed data by removing stop words, regular expression, emoji and hyperlinks.
- Created bag of words and calculated tf-idf. Employed Xg boost and Bert with TFhub to develop the classifier. NLTK, scikit-learn, pandas and keras on python were used.

Credit-Card-Fraud-Detection

- Analysed data to find fraud patterns and anomalies in the data.
- Created supervised and unsupervised ML/DL solution to predict fraud transaction
- Tuned Gradient boosting tree classifiers that performed best on the data to improve accuracy even further and decrease false positives. Tried unsupervised and semi supervised approaches like autoencoder and clustering.
- Used Scalable frameworks such as Apache Spark and Tensorflow on python

PUBLICATION

Y-net: Biomedical Image Segmentation and Clustering : Pathan,S. Tripathi,A. Paper Link

CERTIFICATION

- Google Cloud Platform Big Data and Machine Learning Fundamentals, by Google Cloud and offered through Coursera.
- Machine Learning with TensorFlow on Google Cloud Platform Specialization, by Google Cloud and offered through Coursera.