Biraj Parikh

Last Updated on September 19th 2020

Google Cloud Certified Professional Data Engineer

Bloomington, Indiana | 812-650-2716 | birajparikh16@gmail.com | GitHub | LinkedIn | Portfolio

EDUCATION

Indiana University, Bloomington, USA — **August 2019** - **May 2021**
Degree: Master of Science in Data Science. — **GPA: 3.9**

Dwarkadas J. Sanghvi College of Engineering, University of Mumbai, India — **August 2013** - **May 2017**
Degree: Bachelor of Engineering in Mechanical Engineering. — **CGPA: 3.7**

PROFESSIONAL EXPERIENCE

Apothecary.ai | Massachusetts, USA — 4 months | **May 2020 - present**
Machine Learning Engineer Intern

- Developed and automated a script to scrape over 200,000+ product reviews and ratings using Python and BeautifulSoup.
- Devised and championed a Collaborative Filtering for giving personalized product recommendations based on user-given information and achieved an accuracy of ~ 90%.
- Orchestrated and automated ETL pipeline workflow using Apache Airflow and deployed the containerized model API on AWS Elastic Beanstalk.

Advanced tech skills & tools: Python, PostgreSQL, Apache Airflow, Docker, AWS RDS, FASTAPI, Tableau.

Reliance Jio Inc. | Mumbai, India — 1 yr 2 mos | **May 2018 - June 2019**
Data Engineer

- Worked with cross-functional stakeholders to streamline real-time and scalable data pipelines resulting in 20% redundancy reduction.
- Optimized Real-Time and Robust spark jobs by 15% to transform the streaming data into the optimal format and making fault-tolerant code on the On-Prem (production) cloud data lake environment.
- Independently analyzed and enhanced the performance of complex Hive scripts resulting in 5% improvement in the performance.
- Reported the analyses by developing powerful detailed Dashboards on ZoomData to make data-driven decisions.

Advanced tech skills & tools: Apache Spark, Spark Streaming, Scala, Hadoop, Apache Kafka, Hive, Nifi, Apache Airflow, ZoomData.

Piramal Corporate Service Ltd | Mumbai, India — 7 mos | **October 2017 - April 2018**
Data Science Intern

- Implemented a predictive and prescriptive model for a Fraud Detection use case to predict the feasibility of debtor loan repayment, utilizing past loan history and customer behavior metrics to make a tangible business impact.
- Innovated web-scraping framework as another measure to validate user information accounting for 5% of the overall business decision.

Advanced tech skills & tools: Python, R, OOPs, Selenium, Microsoft Excel, PowerPoint, Word.

SKILLS

- **Machine Learning:** Classification, Regression, Clustering, Neural Networks, Ensemble Learning, Forecasting, Statistical techniques, CNN, Time Series Analysis, Dimension Reduction (PCA, SVD), Natural Language Processing.
- **Languages and Scripts:** Python, R, Scala, PySpark, SQL, shell-scripting, HTML.
- **Libraries:** Pandas, Numpy, Scikit-learn, NLTK, Requests, Beautiful Soup, Plotly, Matplotlib, tidyverse, ggplot2, dplyr, Tensorflow, Keras.
- **Framework/Tools:** PyTorch, Flask, FASTAPI, Hadoop, Apache Spark, Spark Streaming, HBase, Hive, Nifi, Apache Kafka, Apache Airflow, AWS, Google Cloud Platform (GCP), Tableau, Looker, Git, GitHub, Bash, MS Office, IntelliJ, Jupyter Notebooks.
- **Databases:** MySQL, MongoDB (NoSQL), PostgreSQL, Cassandra, Amazon Redshift.
- **DevOps Tools:** Docker, Kubernetes.

PROJECTS

Music Data Analysis on AWS (Spark, Apache Airflow, AWS Redshift, AWS EMR, S3, Star Schema model) — **August 2020**

- Developed an ETL pipeline which extracts data from the data lake hosted on S3, stages them in Redshift, and transforms into a set of dimensional tables using Spark application deployed on AWS EMR cluster.
- Orchestrated the ETL pipeline using Apache Airflow to schedule and routinely monitor the workflow.

Human Protein Multi-Label Image Classification (PyTorch, Convolutional Neural Networks, Transfer Learning) — **June 2020**

- Implemented ResNet34 model architecture, to identify and classify (multilabel classification) mixed patterns of proteins in microscopic images to accelerate biomedical image analysis.
- Optimized the model performance using regularization and state-of-the-art techniques like Transfer Learning, learning rate finder, augmenting, batch normalization, gradient clipping which improved the accuracy up to 87%.

Glassdoor Analytics (Python, Machine learning, Deployment) [App Link] — **February 2020**

- Deployed an ML model on Heroku with 85% accuracy to predict whether the employers will renew the job slot product subscription based on the past subscription contract behavior with the Glassdoor.

Cancer Diagnosis Multi-Class Classification (Statistical analysis, Confusion Matrix, Probabilities, ML models) — **January 2020**

- Efficiently identified important features using Univariate data analysis on the input feature.
- Stacked models like Logistic Regression, Support vector machines, and Naive Bayes to classify the genetic mutation based on text-based clinical literature and achieved a Log loss of 1.15.