Vipul Rai

Linkedin - https://www.linkedin.com/in/vipulrai/
Blog - https://vipulrai.me

vipulrai8891@gmail.com
+91-8892598819
Bengaluru, Karnataka

7+ years of experience in handling petabytes of data, implementing engineering pipelines, predictive modeling, machine learning, and deep learning. Python expert, passionate about using analytics to drive strategic business decisions.

Skills/Algorithms/Platforms

- TensorFlow, Pytorch, Keras, Fast.ai
- Decision Trees, Random Forest, Logistic Reg
- Supervised, Unsupervised & Transfer Learning
- AWS (S3, EC2, Lambda, API Gateway)
- CNN, RNN, LSTM, GAN, Transformers, BERT, DETR
- KNN, K-means, SVM, XGBoost
- Python, Django, MySQL, Hive, HDFS, Spark, py-spark

Professional Experience

Senior Data Scientist

Jun 2017 – May 2020

SmartNomad

Bengaluru

- Designed and implemented end to end pipeline to create personalized itinerary generation engine (Flights, Hotels, Restaurants, POIs) for travelers
- Object Detection to select similar images and discard images containing unqualified objects using YOLOv3 Topic analysis for tagging categories of each point of interest (POI) using Latent Dirichlet Allocation in sci-kit learn
- Multi-class Sentiment Analysis for POI reviews using BERT in PyTorch and HuggingFace Transformers
- Developed Video Creator app using OpenCV and Moviepy
- Integrated multiple Google APIs (distance matrix, maps, places) and other partner APIs Implemented a serverless microservices architecture on AWS to support recommendation modules
- Led and mentored a team of 7 people (Data scientists and Data Engineers)

Consultant

Oct 2015 – May 2017

Affine Analytics

Bengaluru, Seattle & Los Angeles

- **Gaming Analytics** - Fortune 100 client, the biggest game publisher in the world
  - Developed the end to end pipeline for data orchestration to consumption using PySpark and Kafka
  - Raw data was being generated in json / feeds on Hive
  - Converted data to parquet format for faster retrieval and less storage.
  - Creating automated QA module on Spark, which did basic QC of data such as calculating min, max , avg etc for each day's fresh feed thereby removing the necessity of manual intervention
  - Wrote an automated Pyspark program which calculated user stats on daily, weekly and monthly level
  - Created input data set for cross sell model using stats such as available PS Plus points, weekly spend on PS points
  - Created model to detect fraud and cheat during gameplay

- **Retail Store Revenue Optimization** - Fortune 100 Retail giant
  - The pricing team was tasked to identify optimal promo/clearance sale prices which would result in maximum sales.

    - Identified multiple feature which impact unit sales during promotions such as Price Elasticity, seasonality, % sales in each day of the week, the effect of circulars/promotions, breakage due to unavailability of different sized SKUs, the impact of public holidays, etc. using Spark-Scala

- **Travel Portal Optimization** - Fortune 100 Online Travel brand
    - Gave an optimized solution to reduce the number of unqualified searches it is getting from meta channels, to improve the business KPIs like ROI/profitability/efficiency
    - Wrote the Optimization Program using LPsolve from Pulp package using PySpark and logistic regression

- **Travel Portal - Email Marketing and Campaign Analysis** - Fortune 100 Online Travel brand
    - Data was being generated at about 2GB per hour in json format.
    - Converted data to parquet format for faster retrieval and less storage.
    - Used client API to map the data and convert into standardized format.
    - Used Qubole for accessing the spark cluster and running the queries.
    - Wrote program in Pyspark and Scala Spark to select different campaigns for various categories of customers based on Business Rules
    - Hands on experience working on AWS and S3

Big Data Developer

Feb 2015 – Oct 2015

AIG (American International Group)

Bengaluru

- **Log Analyzer** - Internal Project
    - To parse data incoming from 8 different type of servers to be consumed by Power BI
    - Logs from 8 different types of devices were being generated simultaneously.
    - Used Kafka and Pyspark to generate insights such as - which user was accessing the websites which were blocked etc.
    - Designed the end to end architecture and developed the same

Software Engineer

July 2013 – Jan 2015

Blue Star Infotech

Bengaluru

- Learned and worked around Big Data tools such as Hadoop, Hive, etc.
- Set up 10 node apache hadoop cluster for in-house development
- Migrated data from Oracle to Hive, modified table normalizations and increased the reporting speed by 3X from original

Kaggle - https://www.kaggle.com/vipulrai

pandas-dev/pandas - Active contributor to pandas-dev, a participant in discussions, bug fixes and documentation, test, generic

Certifications/Courses

- Databricks Developer Certification for Apache Spark
- Deeplearning.ai - Coursera Deep Learning Specialization

Education

Computer Science and Engineering Jawaharlal Nehru Technological University 2008 – 2012