Ibrahim Mohammad

Full-Stack Data Scientist | IBM Certified imohammad@uh.edu | +1 713 366 2544 GitHub LinkedIn

Programming	Data Science	Big Data & Cloud	Full Stack
C++, Java, Python, Haskell, Go, MATLAB JavaScript	Scikit-learn, Tensorflow, Keras, scikit-image, Pandas, SparkML, SystemML, Matplotlib, NLTK, Numpy, OpenCV, Scrapy, Selenium, Seaborn	Hadoop, Spark, Pig, Hive, Avro, Parquet, Snappy, pydoop, pyspark IBM Cloud, Amazon AWS	HTML, CSS, PHP, Angular, MySQL, MongoDB, PostgreSQL, SQL Server

EDUCATION

University of Houston – Master of Computer Science | GPA – 3.5 Aug 2018-May 2020

Coursework: Advanced Computer Vision, Advanced Numerical Analysis, Software Design, Machine Learning, Computer Networks, Big Data Analytics, Data Structures and Algorithms, Database Management Systems, Digital Image Processing, Data Warehouse & Data Mining

Research: Prediction of unscheduled rehospitalization in patients with Multiple Chronic Conditions within 30 days, using Deep learning principles Variational Auto Encoders and Generative Adversarial Networks.

PROFESSIONAL EXPERIENCE (3 YEARS)

Graduate Research Data Scientist - *University of Houston* Nov 2018-Present

Exposure: Python, Tensoflow, Scrapy, Selenium, MonogDB, Statistics, Linear Regression, Visualization, PowerBI **Houston, USA**

- Developed automated python scripts to scrap Yelp and Google Maps and collected more than 10,000 restaurants data into MongoDB
- Analysed, Pre-processed and wrangled unstructured data for successful data management in PostgreSQL
- Performed Exploratory data analysis and implemented **Future Sale Prediction** model on a 250GB of web scrapped data with **89% accuracy**
- Experience in performing descriptive and Predictive analysis on large data sets and developing **story telling dashboards** in PowerBI
- Scrapped and Trained 10000 captcha images to build a Captcha Decoder using object detection in deep learning with an **accuracy of 90%**
- Collaborated with a PhD student in implementing Decision Trees, Random Forest, Bagging and Boosting for a recommendation system.

Software Engineer - *Cognizant Technology Solutions* Sep 2016-Jul 2018

Exposure: Informatica, SQL, TeraData, Java, Jira, Data Modeling **Hyderabad, India**

- Developed and tested Extraction, Transformation & Load (ETL) processes using Informatica in **Jira agile environment**
- Experience in designing Star schema & Snowflake schema using fact & dimension tables.
- **Modelled Data Marts** by understanding the existing Oracle Data Models, and create tables, views to feed data for reports.
- Created scheduled jobs to automate the execution of ETL process and refresh materialized views to generate reports.
- Wrote complex SQL scripts & stored procedures to avoid informatica lookups to improve performance as the volume of data was heavy
- Involved in performance tuning, fixing production issues and responsible for migrating data from **Oracle to Teradata**.
- Implemented **Test Driven Development** to complete the unit testing and support system testing to reduce the bugs by at least 20%.

PROJECTS

Blindness Detection – Kaggle: (Python, Deep Neural Networks, Keras, Image processing, scikit-image, OpenCV) kaggle

- Developed a Deep Neural Network model to detect diabetic retinopathy to stop blindness before it's too late using advanced image processing techniques including, intensity slicing, histogram matching, Otsu's image thresholding and auto cropping

Gender Classification: (Supervised Learning, POS tagging, Seaborn, NLP, Scikit-learn, Pandas, Seaborn) GitHub

- Developed a Supervised Machine Learning model to classify gender based on text from collection of articles written by men and women
- Performed Feature engineering on writing patterns in men & women including POS tagging & Chi square test & Visualized t-SNE correlation
- Implemented Stochastic Gradient Descent on SVM which learns & identifies best features from the base model and classifies the target class

Neighbourhood Clustering-Houston:(Python, IBM cloud Folium, Pandas, k-Means, geopy) GitHub

- Implemented the objective of finding which neighbourhood of the Houston city is a good choice for a new restaurant business to open
- Clustered 88 neighbourhoods of Houston based on similarities between restaurants present in each neighbourhood using the **foursquare** location data. Performed Exploratory Data Analysis & Implemented k-Means clustering, deployed on **IBM Watson Studio**

Pairwise Similarity Measure: (Pydoop, Pyspark, Avro, Parquet, Snappy) GitHub

- Implemented pairwise document similarity on 2 million documents by creating inverted index on most popular words using Map Reduce and calculated similarity matrix for each pair of docs. Used Avro & Parquet file formats to optimize R/W access and Snappy for data compression

Content Delivery Network: (Python, HTTP, Amazon EC2, Google Cloud) GitHub

- Built a network of proxies and multithreaded CDN servers distributed worldwide to deliver content relatively at high performance rate.
- Implemented **Distance Vector Routing Algorithm** to find the shortest path among servers and two types of caching mechanism to optimize the content delivery. Deployed CDN servers in different geographical regions using **Amazon Web Services and Google Cloud**

CERTIFICATIONS: **Cognizant Certified ETL Developer | IBM Certified Professional Data Scientist | Tensorflow in Practice Specialization**