Vaibhav Sharma

vaibhav.sharma12@hotmail.com | +91-9752236647 | New Delhi, Vikaspuri 110018 | vaebhav.github.io

PROFILE

Dedicated and inquisitive professional with 5+ years experience in IT industry. Passionate about simplifying my work using script automations and keen to use my analytics solution-building exposure in enabling business transformation and processes. Learning to tell stories from data by leveraging my development experience to build solutions for leading clients. Having a good exposure with Python scripting and its data analytics suite, along with end to end delivery

Skills / Interests

Languages: Python, PySpark, C++ ,Shell Scripting, GoLang, SAS, R , SQL
Frameworks/Tools: BigQuery, DataFlow, Azure Data Factory, Dataiku, DataBricks, Power BI , REST API, Hive
Cloud: Azure,AWS, GCP

EDUCATION

BE (Elec & Comm)	**6.9**	**2010 - 2014**
HSC	**6.9**	**2009 - 2010**
SSC	**7.6**	**2008 - 2009**

Experience

Deloitte Consulting LLP | *Consultant l Applied AI* **June 2019 - Present**

Mars Incorporated

- *Developed base calibration utility on for performing model selection across a range of 16 models , utilising correlation , **Z score** and Percentage Changes in lieu with **MAPE**.*
- *Developed model implementation for **FBProphet** , **Auto Arima** based on historical data for **Demand Forecasting***
  *leveraging **DataBricks** on **Azure** using **PySpark***
- *Developed a **parallel multiprocessing** Base Calibration Segmentation utility to bucket and classify similarly performing **TimeSeries** models on **DataBricks** using **PySpark**.*

HP Inc

- *Implemented **P&L Forecasting** solution across*
  - *Uni Variate Forecasting - Leveraging 16 models to generate forecast based on historical trends*
  - *Multi Variate Forecasting - Additionally taking into account external financial & economic drivers as a feature set , in lieu with internal factors affecting P&L , leveraging **XGBoost**,**ARIMAX, ProphetX** regressors*
- *Developed a seminal **Interpolation** & **Extrapolation** wrapper for missing data points across drivers*

AmeriGas Partners LLP

- *Analysed and processed data from multiple sources for **data modelling** , using MSSQL , spanning across **billion** data points. Developing flexible **data transformations** and **pre-processing** pipeline for model ready input*
- *Developed a parallel **multiprocessing Customer Segmentation** utility to bucket and update customers based on their historical transactions.*
- *Developed and analysed multiple models namely , **Logistic Regression ,XGBoost and LightGBM** for customer churn propensity across multiple clusters.*
- *Developed a seminal execution framework for generating predictions as a module , increasing portability and reducing manual intervention.*

MPI Analytics LLP

- *Worked as a **Data Lead** & managed a team of 2 members; developed **data model** as assets & its supporting dashboards. **Architectured solution framework**, keeping each asset loosely coupled, enabling flexible, robust & independent **monolithic** management enabling code reusability by **30%***
- *Assisted the development of **Sales Forecasting** asset , comprising of **TimeSeries** and **LinearRegression** predictions*

Boehringer Ingelheim

- *Spearheaded data ingestion & analytical framework to help support the deliverable dashboard, and intents related to major KPI's dynamics.*
- *Developed a python module based on **bigquery-client** with a dynamic & customisable rejection check, utilising the sweet spot between **BigQuery** and **DataFlow**, thus reducing the ingestion time by **10x folds** & increasing raw data accuracy by **30% across 30 billion** data rows.*

Optum | *Data Scientist* **Jan 2019 - May 2019**

- *Automated training **Acoustic** Models for **CallTranscriptionEngine** based out of **HMM Kaldi** toolkit , targeted for various business segments and languages, achieving ~**18%** WER (Word Error Rate).*
- *Undertook seminal development of **CallTranscriptionEngine**, utilising **MultiProcessing** thus reduced overall execution time by **60%** & enabling multiple requests at once end to end.*

TCS | *Developer Analyst* **Jul 2015 - Dec 2018**

- *Worked across multiple Projects , primarily focused with **application enhancement and development**.*
- *Created multiple **automation** & **wrappers** to support core business operations enabling **25%** increase in throughput*
- ***Lead** and **mentored**, peers across multiple levels*

ACHIEVEMENTS

- *Received Outstanding Performer & on the spot awards for various automations and achievement in multiple projects*
- *GCP Certified **Professional Data Engineer -** here*
- ***Microsoft - DAT210x** (Python for Data Science) Certification from EDX*

CAPSTONE SOLUTIONS

Fuzzy Logic based Data DeDuplication -

*Developed a Python module from scratch leveraging **fuzzy logic** algorithms to cluster similar records, aimed at eliminating duplication. The model clusters records into a single record based on edit distance and **token set ratio vectors**, creating a **similarity matrix** across a common hash space across multiple fields. The solution is aimed at **Master Data Management**, identified **37% duplicated** records for a leading Pharma Client ranging across 330K records*

***Technique used-:** Leaders Clustering, Levenshtein & Edit Distance , Token Set Ratio*
***Tools and Technology Used-:** Python, Fuzzywuzzy, GCP*

Time Series based Anomaly Detection -

*Built a time based model to identify **anomalous** data point(s). Input dataset was trained on lag values to predict & evaluate if generated trend values lie within the upper and lower boundaries , divergence between actual and predicted values are further evaluated upon a combination of static & sliding window by analysing **z-scores** for **classification**. The model was evaluated across a benchmark **Yahoo S5 labelled dataset**.*

***Technique used-:** FBProphet, ARIMA, Auto ARIMA*
***Model Validation-:** Confusion matrix , Recall vs Precision , Z Score*
***Tools Used-:** Python (Pandas, Numpy, Sklearn,Scipy etc)*

PyModBus Concurrency Wrapper -

*Developed a concurrency python script based on **asyncio** & **Modbus** protocol, to retrieve **holding registers** across multiple units values which were further ingested to **MySQL** using **POST REST API**. The script was scheduled using crontab and deployed on **Rasberry PI***

***Tool Used -:** Python , PyModBus , Rasberry Pi, Linux*

Remaining UseFul Life -

*Developed a predictive **multi-label and regression** models to predict **RUL** based on various features present within the NASA Dataset. Performed interactive **EDA** across multiple sensors, Models were trained and validated across 4 datasets across 100 engines each*

***Technique used-:** BinaryRelevance, MultiLabel, Regression*

***Tool Used -:** Python , Jupyter Notebook, Scikit -Learn*