# SUMEET SANJAY VYAS

High Energy Data Scientist with 4 years' experience Targeting Assignments in **Data Science and AI**

**Mail ID:** vyassumit1994@gmail.com **Mobile:** +91 9421706901

## Profile Summary

- An IT professional with work experience in **Machine Learning, Deep Learning, Computer Vision, Natural Language Processing** in the Domains of **Retail, Insurance, Defence, Energy, Education, Healthcare** as a **Data Scientist / Machine Learning Engineer**
- Data Scientist role consists of gathering requirement, scoping of the goals to be achieved, data gathering, AI/ML product/offering demonstrations, aligning with the technology and working in close quarters with the Solution Architect to design, presenting the AI/ ML solution to the clients , performing **Exploratory Data Analysis on the Data, Visualizations, perform Modelling, validation and deploying the solution.**
- Conduct **Internal Trainings, Workshops and Lectures** for **Statistics, Machine Learning and Deep Learning,** and discuss Case Studies
- Received "On the Spot Award" at previous organisation, Faculty Award and qualified for **"Master Orator Championship" for Public Speaking and Communication**

## Core Competencies

| | | |
|---|---|---|
| Machine Learning | Computer Vision | Client Engagement |
| Deep Learning | Natural Language Processing | Strong Communication |
| Statistical Analysis | Traditional ML | Public Speaking |

## IT Skills

**Technologies Used:** Tensorflow, Sklearn**,** Numpy, Scipy, Pandas, Matplotlib, PyQt5, OpenCV, Tkinter, PIL, Tensorflow Detection API, Keras, PySpark, HandySpark, etc

## Work Experience

**Quantiphi Analytics Solutions Pvt. Ltd, Mumbai** **Dec 2018 - Present**

## Projects:

### 1. Build a Predictive model for detecting Shows/No-Shows using the Appointment Data and Sentiment Analysis of Call Recordings

**Client Sector:** Medical Industry

**Description:** A systematic AI driven framework consisting of **Predictive model to detect the "Show" and "No Show"** of a patient after the appointment has been booked. An Automation pipeline with Deployed model **(XGBoost)**, had **accuracy 91%**, with prediction in **batch mode**. **Transcription of the call recordings with Google Speech-to-Text Service**, cascaded with **Google NLP service for Sentiment analysis** extracted features that were used in modelling. Other algorithms tested – **CatBoost, Random Forests, Gradient Boosted Trees**. Data Size = **20 million samples**. Other Technologies Tested: **HandySpark, Google DataStudio, Google Dataproc.**

### 2. Customer Segmentation and analysis of Undercuts/Discounts

**Client Sector:** HealthCare and Ecommerce Industry

**Description:** The goal was to **find patterns** in the **Sales Data and Insights** related to the same. The model was deployed in **Batch Mode**, primarily trained using the **K-Prototypes algorithm**. Worked with the client starting from Data Collection phase to getting the Business Logic. Other algorithms tested – **KModes, KMeans Clustering**. The exercise involved **Extensive EDA, Outlier Removal**, and study of

feasibility of Data reduction Techniques and Feature Transformation. Encoding for categorical variables included **One-Hot Encoding, Count-Frequency Encoding** and other preprocessing techniques like **Normalization, Cyclic Transformation.**

## Case Studies:

### a. Marketing Mix Modelling for Insurance domain:

**Description:** A case study of an existing deployed model was done for an **Insurance domain Client**, to understand the deployed model, it's shortcomings, and suggest New Models that would provide better performance compared to the existing Models. Models Tested - **Shapely Value Regression, Log-Linear Regression, Response / S-Curves.** Suggested improvements – Additional preprocessing, addressing the **Multi-collinearity** in existing Model, etc

### b. Model Interpretability modular system:

**Description:** A modular system to interpret the existing Models was developed using the **Shap and Anchor Models** Interpretability Techniques. The goal of the system is to provide with interactive Visualizations, showing the relative **feature importance** and provide metrics to help **feature Engineering.**

### c. Integrated Tool for Graphical Image to Data Conversion:

**Description:** An Image processing solution was designed to convert the existing **Survival Graphs** in bulk, to CSV tables. The following Image processing Techniques were used - **Harris Corner Detection, Image Gradients, Orientation Filters, OCR** to detect values on the respective axes.

**Tata Consultancy Services Ltd, Pune** **Oct 2015 – Dec 2018**

### 3. Development of algorithms to localize and classify Rust Levels

**Client Sector:** Energy and Oil Industry

**Description:** A **Video Analytics** based solution that consumes Images and output the image with boundaries of the possible Rust regions, and its **Classification** into different severity levels. The complete application with GUI was delivered to Client with the **accuracy figures of 80% for segmentation and 82% for classification**. **Features** from **HSV, YCbCr colorspaces, GLCM, HOG** were used to train the following **Models - Support Vector Classifier, K-Nearest Neighbor**. Features were optimized through Recursive **Elimination Techniques**, **Cross-validation** for testing.

### 4. Development of algorithms to detect underwater pipes and detect anodes

**Client Sector:** Energy and Oil Industry

**Description:** A **Video Analytics** based solution to Automate the **detection process of an Underwater Pipe** attribute called as "Anode" and provide a **probabilistic score for its Segmentation**. A **Fully Convolutional Neural Network** was trained with **Accuracy Figure of 91% (pixel-wise) accuracy**. Other Models that were experimented with, from **Tensorflow Object Detection API - Single Shot Multi-Detector**. An Image processing solution approach was also designed which involved the following preprocessing - **Histogram Equalization, Filters, Orientation filters, Hough Transform, Morphological operations, HOG, LBP, Clustering** and the **Machine Learning** classifier that was tested was a **Support Vector Classifier** with the accuracy of **76%.**

### 5. Automation of Control Systems Testing for an Underwater Vehicle

**Client Sector:** Defence Industry

**Description:** An **Automation solution** was designed to speed up the process of Testing of a Control System. The Automation involved getting feed from the Control System GUI, **recognizing the GUI Objects, designing an OCR**, analyze the image with few Image Processing techniques and validate the results. The **Machine learning** algorithms used were - **CNNs for OCR, KMeans Clustering for Dominant Color Detection**. Preprocessing Techniques - **Thresholding, Canny, Color filter, Contours, KeyPoint matching (SIFT, ORB)**

## Udacity Nanodegree Projects:

### 1. Machine Translation from English to French

**Description:** Build a deep neural network that functions as part of a **Machine Translation** pipeline. The pipeline accepts **English text** as input and **returns the French translation**. The goal is to achieve the highest translation accuracy possible. Accuracy achieved: **96%**

**ML Algorithms: Recurrent Neural networks, LSTMs. GRUs**, **Data Preprocessing:** Tokenization, Normalization

### 2. Spam Classifier

**Description:** Using the **Naive Bayes algorithm** to create a Classifier dataset SMS messages as **spam or not spam**.

**ML Algorithms: Bag of Words, Bayes Theorem implementation, Naïve Bayes classifier**

### 3. Part of Speech Tagging

**Description: Pomegranate library** is used to build a hidden Markov model for part of speech tagging with a universal tagset. **Hidden Markov Models** were able to achieve **>96%** tag accuracy with larger tagsets on the given text corpora.

**ML Algorithms: Part of Speech Tagging, Hidden Markov Models**

## Research Paper

**Title**: "Design of **Signal Conditioning Circuit** for **Biomedical Sensors** and **Battery Monitoring Circuit** for a **Wearable Communication System**" published under **American Journal of Engineering and Research**

## Certification

- **Programming, Data Structures and Algorithms in Python** from **NPTEL**
- **Convolutional Neural Networks** from deeplearning.ai **Coursera**
- **Natural Language Processing Nanodegree** from **Udacity**
- **Statistical Learning** from **Stanford Online**
- **Applied Machine learning in BigQuery** from ACloudGuru
- IoT Foundation Program (**IoT, Architecture & Applications**) from Nexiot
- **Andrew Ng's Machine Learning** on Coursera

## Education

- **BE (Electronics and Telecommunications)** from PVG's College of Engineering and Technology, Pune in 2015 with 63%
- **12th** from Deogiri College, Aurangabad (Maharashtra Board) in 2011 with 74.5%
- **10th** from Deogiri College, Aurangabad (Maharashtra Board) in 2009 with 88%

## Personal Details

- Date of Birth: 11/01/1994
- Languages Known: English, Hindi, Marathi, Japanese(basic)
- Address: Shalom, Room No 3, Kharodi, Malwani church, Malad West , Mumbai