SANKET GODBHARLE

Associate Data Scientist

CONTACT

7977070308

sanketgodbharle12@gmail.com

Pune,Maharashtra

ABOUT ME

Immediate Joiner | Ready to Relocate|n Data Scientist with around 3 Years of Experience in Python, Statistical Analysis, Machine Learning, GEN AI, Lang Chain, Prompt Engineering, LLMs, Fine tuning LLMs, Synthetic Data Generation, Model building, Natural Language Processing and Deep Learning, Azure, AWS and Data Visualization, Pyspark, Predictive Modelling. Seeking challenging opportunities to leverage my expertise in driving data driven decision making and solving real world problem.

EXPERIENCE

Associate Data Scientist | Shruteekatech Pvt Ltd | 3.2 Years | November 2021 to PRESENT.

SKILLS SUMMARY

Data Science, Python & ML -

- **Strong Python Programming skills using OOPS, Functions**
- **Sound Knowledge of Python's Data Analysis and Machine Learning Libraries.**
- **Implementation of regularization techniques like Lasso and Ridge in regression.**
- **Data Modeling: Using algorithms- Linear Regression, Logistic Regression, Naive Bayes Classifier, Support Vector Machine, Random Forest, XGBoost, Adaboost, Gradient Descent, Principle Component Analysis.**
- **Data mining algorithm experience in the families of predictive algorithms (Regression, KNN, Decision Trees)**
- **Data extraction, Data Manipulation, Tesseract.**
- **Data Clustering : Clustering Algorithms (k-means clustering),Hierarchical clustering.**
- **Data Analytics on large data sets to solve business problems.**
- **Power BI : visualization tool.**
- **Ability to process Text Processing using NLTK library and other Natural Language**
- **Thorough understanding of Probability and Statistics, Bayesian method, Hypothesis Testing.**
- **Exposure to Large Language models (LLM) : GPT (OPEN AI), BERT.**
- **Experience in data management tools – Non-Relational and SQL databases.**
- **Ability to use Web Scraping Tools as Beautiful Soup.**
- **Source code management and Version Control system using Git and GitHub.**
- **OCR TOOLS: Kofax, Simple OCR, Regex, Tesseract, Data extraction.**
- **Cloud platform for deployment- AWS, AZURE CLOUD.**

Technologies

- Python/ML Packages: Scikit Learn, Pandas, Numpy, RegEx, Matplotlib, Plotly, Seaborn for visualization.
- Deep Learning: Neural Networks, ANN, CNN, DNN, Transfer Learning, Back Propagation, Deep Neural Network, Linear Algebra, Activation & loss functions, Tensorflow 2.x, Keras, PyTorch
- NLP: Text understanding, representation, classification & Text clustering skills.
- Libraries: nltk, spacy, gensim, textblob, langdetect, googletrans, unidecode.
- Techstat: BOW, TFIDF, word2vec, doc2vec, sent2vec, keypharse extraction.
- AWS: Elastic Compute 2 , Sage maker , Notebook instance, AWS container, Simple Storage Services S3, Multi-cloud environment, Deployment, AWS Redshift, AWS Lambda.
- Databases: MySQL, Command, Constrains, Clauses, CRUD operations, Subqueries, Window functions, Joins.
- Maths & Stats: Filter, Wrapper, Embedded Methods, P-Value, T-Test, Z-Test, ANNOVA test, Chi-Square Test, Info-Gain Test, Hypothesis Testing. Probability,s tatistics, linear algebra

Education

- B.E | 2016 | Mumbai University
- HSC | 2011 | Pune | First Class
- SSC | 2009 | Pune | Distinction

Projects Handled

Project-1 Handwritten to Text Data Extraction using Deep learning (OCR).

Domain :Banking

Description:- Handwritten checks, forms, and financial documents can be converted to digital text for automated processing, data extraction, and record-keeping.To develop a system that can convert handwritten text into digital text using deep learning. It allows the recognition and extraction of characters from scanned images or handwritten documents, enabling the conversion of analog content into searchable and editable digital formats.

Roles and Responsibilities

Work on Machine Learning/Deep Learning

- Write quality code while following best practices
- Analyze and preprocess the handwritten text data.
- Clean and enhance the images for better performance.
- Used libraries: Tensorflow, Keras, OpenCV, etc.

Project 2 : Medical Domain Chatbot using RAG and LLaMA2
Domain :- Healthcare

Description: Developed an AI-powered chatbot for the medical domain using Retrieval-Augmented Generation (RAG) and LLaMA2. The chatbot extracts relevant information from large medical documents and provides accurate responses based on user queries.

Roles & Responsibilities:

- **Extracted data from 1000+ pages of medical PDFs.**
- **Split data into text chunks and embedded them using Hugging Face word embeddings.**
- **Indexed and stored embeddings using Pinecone for efficient similarity retrieval.**
- **Implemented a retrieval system using Pinecone vector stores for context-aware search.**
- **Integrated LLaMA2 to generate meaningful responses based on retrieved data.**
- **Used libraries: LangChain, Pinecone, Hugging Face, Flask, LLaMA2**

Project 3 : Sentiment Analysis Project with NLTK and Transformers
Domain:- E-Commerce

Technologies: Python, NLTK, Transformers (Hugging Face), VADER, Bag of Words, RoBERTa

Description:

Developed a sentiment analysis model to classify customer reviews on eBay as positive, negative, or neutral. Implemented multiple techniques, including:

VADER (Valence Aware Dictionary and Sentiment Reasoner): Used for lexicon-based sentiment scoring.
RoBERTa Pretrained Model: Leveraged Hugging Face's pipeline for deep learning-based sentiment classification.

Roles & Responsibilities:

- **Collected and preprocessed eBay customer reviews for sentiment analysis.**
- **Implemented NLP techniques such as tokenization, stopword removal, and text vectorization.**
- **Developed sentiment classification models using rule-based (VADER) and deep learning approaches (RoBERTa).**
- **Provided insights from sentiment analysis to improve customer satisfaction strategies.**

Project-4 Vehicle Insurance Fraud Detection

Domain :Insurance

Description :-The goal is to develop a Machine Learning model that can effectively detect instances of fraud in the context of vehicle insurance. Insurance fraud is a significant issue, costing companies and customers billions of dollars each year. By leveraging machine learning techniques, we aim to build a system that can identify fraudulent claims and help insurance companies take appropriate action.

Roles and Responsibilities:

- Understand, analyze, and interpret large datasets.
- POC (Proof of Concept).
- Develop advanced programs to extract the data needed, prepare data for further analysis.
- Discover, design, and develop analytical methods to support novel approaches of data and information processing
- Perform analysis to assess the quality of the data, determine the meaning of the data, and provide data facts and insights.

LANGUAGES:- ENGLISH, HINDI, MARATHI

HOBBIES : WATCHING CRICKET, WATCHING MOVIES, TRAVELLING

MARITAL STATUS:- UNMARRIED