SIDHARTHA SANKAR MONDAL

Brampton ON
+14374505938
Sidharthmandal007@gmail.com

https://www.linkedin.com/in/sidhartha-sankar-mondal-651b04106
https://github.com/sidharthamondal

Data Scientist with a demonstrated history (5+ years) of successfully leading business-critical data science projects for major organizations (Financial, Retail, hospitality domain). Expert in data science programming languages (R/Python/Spark), statistics and end-to-end machine learning techniques to build large-scale data pipelines, algorithms, and analytics applications. Ensures high quality with well-designed, cleanly coded, well-documented and operationally stable models. Proven success in utilizing expertise in large-scale computational architectures, deep learning framework, data visualization, and simulation to solve research challenges. Experienced with a demonstrated history of working in the higher education industry.

PROFESSIONAL SUMMARY

- Proficient in Docker and AWS Elastic Container Registry for scalable application deployment.
- Designed serverless ETL processes using AWS Lambda, REST APIs, and Redshift Database.
- Expertise in AWS services including State Machines, Event Bridge, Lambda, SageMaker, Kubernetes.
- Led advanced text classification using BERT, ROBERTA, XLM models.
- Managed big data platforms like Cloudera, Impala, Hive.
- Strong coding skills in Python, R, and GIT for codebase management.
- Utilized libraries: Scikit-learn, TensorFlow, PyTorch, OpenCV, HuggingFace.
- Implemented Time Series Forecasting (ARIMA, Prophet) for financial planning.
- Solid grasp of Linear Algebra, Statistics, Probability.
- Proficient in Power BI for data visualization.
- Applied classical models: KNN, EM Clustering, Random Forest.
- Delivered projects within Sprint Models meeting tight deadlines.
- Blend of image processing, object detection, and cloud deployment skills.

SKILL SET

- Programming Languages: PYTHON, R, DAX, SQL
- Machine Learning and Deep Learning Libraries: Scikit learn, NLTK, NumPy, Pandas, TensorFlow, keras, Scipy, Matplotlib, PCA, OpenCV, Ultralytics, SpaCy, NetworkX, M Pillowatplotlib, Seaborn, Plotly, RASA, Caret, forecast
- Image Processing: Open Cv, PIL,
- Databases and Version Control: MySQL, PostgreSQL, NoSQL (MongodB), GitHub, Elasticsearch
- Big Data: Apache Spark – (Pyspark, Hive, MLlib, Impala)
- Cloud: AWS Sagemaker, AWS Lambda, AWS rest api, AWS EC2, Elastic Kubernetes service, Elastic Container Registry, LightSail, CloudFront, event Bridge, State machines, Docker.
- Data Visualization: PowerBi , Tableau
- Data Science Platforms: Dataiku, alteryx

EDUCATION

- Graduate Certificate Advanced AI solution development George Brown College Toronto Canada – 2023 – Present
- PG Diploma Business Analytics Praxis Business School Kolkata India – 2018
- BTECH Computer Science APJ Abdul kalam technical university Uttar Pradesh Ghaziabad – 2016

PROFESSIONAL WORK EXPERIENCE

SENIOR DATA SCIENTIST
BLOOM HOTEL GROUP   **(Nov/2021 – Dec/2022)**
Roles and Responsibilities

- Built state of the art intent-based chat model using RASA framework, greatly helped the guests coming in the hotels
- Worked on multiple projects for Image classification: Room cleaning and quality assurance.
- Understanding Hospitality data and build state of the art ML based forecasting models for footfall, inventory and budgets for each of the properties.
- Built projects for smooth customer acquisition and engagement.
- Directly reported to investors and directors.

SENIOR ASSOCIATE DATA SCIENCE GROUP
AFFINE ANALYTICS   **(Sep/2020 – Oct/2021)**
Roles and Responsibilities

- Analyse business requirement and build POC for multiple clients
- Direct client interaction and process understanding of client businesses
- Formulating, suggesting, and managing data driven projects which are geared at furthering the business's interests.
- Selecting and employing advanced statistical procedures to obtain actionable insights.
- Suggesting ways in which insights obtained might be used to inform business strategies
- Building ML pipelines for data cleaning, Machine learning models and result validation

DATA SCIENTIST II
ALUMNUS SOFTWARE LIMITED   **(Nov/2018 – Aug/2020)**
Roles and Responsibilities
- Analysing business requirements and formulating analytics problem statement for ML implementation.
- Statistical Modelling and problem-solving using ML techniques - Classification, Regression, Optimization and Clustering.
- Data Exploration, analysis and Model Metrics and results summary to build visualizations using Power BI
- Directly reported to CEO and CTO.
- Handled multiple clients and worked under very strict sprint models.

PROFESSION PROJECTS (DEVELOPED AND DEPLOYED)
Custom LLM ChatBot   **George Brown College Canada**
Tech Used: Python, Pinecone, NLP
Role: Machine Learning Engineer
Dev Tool: VS Code
Developed a mechanism to create a large language model from 100s of pdf. Current LLM faces issues with context loss and hallucinations. Hence, we devised a architecture using vector embeddings to solve it. The project in POC stage achieved superior performance and maintained the contexts.
 Responsibilities:
- Built state of the art custom chunker module which was used to create knowledge base of the system, based on this knowledge base question answers were created.
- Create API around pinecone library to create and update the vector database.
- Trained LLM with the knowledge base created in batches.
- Created post methods which can find the most relevant context from the vector database.

Road Sign Detection Real time Autonomous vehicle   **George Brown College Canada**
Tech Used: Python, openCv, YOLO v5,v8 , ultralytics
Dev Tool: VS Code , google collab
Developed a mechanism to find the signals on the side of the roads in real time, The project was aimed at understanding majorly german road signs and built a model around it.
 Responsibilities:
- Configuring YOLO v5 and v8 architectures for road sign detection.
- Conducting model training and fine-tuning.
- Evaluating model performance using metrics like mAP (mean Average Precision).
- Configuring YOLO v5 and v8 architectures for road sign detection.
- Conducting model training and fine-tuning.
- Evaluating model performance using metrics like mAP (mean Average Precision)

YellowPort Data Pipeline   **Bloom Hotel Group**
Tech Used: Python, PowerBi, Dax, SQL
Dev Tool: VS code
Deployment: AWS lambda, Aws State machine, Aws Event bridge
To build pipeline to collect data from a no SQL database (Mongo DB) and push the data to Amazon redshift, which was further used to make powerBi dashboard in real time to understand daily business, Average daily rate, customer life cycle and much more. The dashboard also included new marketplaces and potential business.
Responsibilities:
- Built full architecture of code to pull data from property management service and scrape data from multiple hospitality website for pricing and new listing in locations
- Used KNN for identifying potential market based on the rates and services offered by the hotels
- PowerBi Dashboard to show all the KPI with multiple drill down options, used measures and DAX codes to transform data.
- Deployed using serverless cloud (AWS lambda) with 0 cost for compute. Highly scalable and cost efficient

Preventive action type classification   **Affine Analytics**

Tech Used: Python, Cloudera, SQL, HuggingFace, Bert, Roberta, XLM
Dev Tool: VS code, cloudera Data Platform
Cargill have multiple factories across the globe, each factory/entity generates a request which is further tagged into a particular department which takes up the work and completes it. The task was to tag the description given in a request to appropriate department
Responsibilities:

- Data collection from CDP and building pipeline to pull data, pre-process and upload the data into the CDP.
- Implemented pretrained language models to translate languages. Opusmt and facbook's MBART Implemented BERT and multilingual Bert based models to finetune the model with the data
- The results achieved is close to 80% accuracy for a multi class problem. Implemented the job which runs automatically in CDSW and pushes the code in Impala.

OPEX Forecasting   **Alumnus Software**

Tech Used: Python, R, Forecast package, Caret, TS outlier
Dev Tool: Jupyter notebook, Dataiku, R
Deployment: Dataiku
A leading European client needed to forecast their operational expenditure for the next 12 months. These expenditures were on different categories Salaries, Travel Expense, Allocations and Consultancy cost. There were around 16000 timeseries which required to be predicted.
Responsibilities:

- Involved in analyzing the data and connecting with business for understanding the data.
- Formulated error calculation metric and defined the success criteria Created and presented the slide decks to the clients the details of the project across 8 different sprints/milestones Implemented the base model in
- Used SARIMA and for outlier detection used tsoutliers to check Additive and level shift outliers. Implemented constrained regression for categories like salaries using headcount as the independent variables Optimized the model for multi-threading in R Created R-markdown reports to send it directly to the stake holders.
- Results saved the business around 400 hours of manual planning and was a great success in the business

Internal Treasury forecasting   **Alumnus Software**

Tech Used: Python, R, Forecast package, Caret, TS outlier
Dev Tool: Jupyter notebook, Dataiku, R
Deployment: Dataiku
The client had internal treasury movement from one Entity to other across the globe. The problem was to give rolling prediction for next 12 months. It was an award winning project for the client in European associates of corporate treasures: https:/eact.eu/articles/eact-award-2020
Responsibilities:

- Responsible for data understanding and defining success criteria. Prepared slide decks and gave presentation to the client across 4 sprints/Milestones Defined success criteria and different KPIs.
- Did extensive data cleaning and analyzed the data. Built model using R. Used Outlier detection using TS outlier (Additive outlier) and used SARIMA model on the adjusted time series. Built another model using some internal data where ensemble regressor techniques were used (Random Forest).
- Reduced around 300-man hours. The project is now in production. Optimized the model for multi-threading in R Created Power BI dashboard to look and view the model performance across different KPIs

---

TEACHING EXPERIENCE (2 Years and 6 Months)
- Visiting Faculty
  PRAXIS BUSINESS CLASS – India (2019 – 2020)
  Subjects: Advance Analytics, Python, EDA on large scale Data
- Assistant Faculty
  ICON EDUCARE – India (2016 – 2018)