(343) 989-4743
Toronto, Ontario
sudhandar@gmail.com

Sudhandar Balakrishnan

Data Scientist

Portfolio: sudhandar.com
github.com/sudhandar
linkedin.com/in/sudhandar

Seasoned data science professional with 2+ years of industry experience in leveraging machine learning, natural language processing, data mining and statistics to solve cross functional business problems for clients. Skilled in generating insights and recommendations by analyzing data and developing visuals to be used across the organization to make better decisions.

Professional Experience

Data Scientist **May 2019 - Aug 2021**

ZoomRx Inc. (Pharmaceutical data science and management consultancy)

- Performed quantitative analysis (impression, reach, user demographics) and qualitative analysis (**sentiment analysis, biological named entity recognition**) by streaming **real-time tweets** using the **Twitter API** to capture twitter buzz in the Oncology space for the largest pharmaceutical company in the world, increasing the team's revenue by **$50,000 per conference**
- Developed and tuned a **Random Forest model** using python (**scikit-learn**, GridSearchCV) to predict the persistence of drug usage among patients (classification) with an **F1 score of 83%** and later **monitored the model's performance**. This helped the clients to realign their marketing strategies and save **$100k per marketing campaign**
- Streamlined and automated **ETL pipelines** by scraping data from 6 websites, parsing the **JSON and XML data**, and storing it in **Google Cloud's SQL database**. These pipelines **saved 20+ hours of manual work per month** and improved the quality of the search results of our core search engine (worth $2M).
- Initiated and owned the designing and implementation of a multipage python **Dash (dashboard) data app** to automatically update records in SQL server based on user inputs and deployed the app using Google Cloud's Compute Engine. The app was used by **10+ users daily saving 8+ hours of manual work per week**.
- Built highly interactive **Tableau dashboards** with segment drill-down capabilities to monitor **KPI metrics** and used the dashboards to provide **strategic recommendations** to clients and senior-level management

Skills

Languages	Python, SQL (MySQL, PostgreSQL, HIVE), R (basics)
Frameworks & Libraries	Pandas, NLTK, SpaCy, PySpark, Hadoop, Pytorch, Tensorflow, Keras, Plotly, Seaborn, Flask, FastAPI
Tools	Microsoft Excel, GIT, JIRA,Tableau, Jupyter Notebook, Google Cloud, AWS, Azure
Methodologies	**Machine Learning (Logistic Regression, Gradient Boosting, K-Means Clustering, Trees)**, Natural Language Processing, Statistics (Hypothesis Testing), Time Series Analysis, MLOps

Education

Masters in Electrical and Computer Engineering (Specialization in Artificial Intelligence) | GPA: 4.25/4.30 **Sep 2022**

Queen's University, Canada

Bachelor of Engineering in ECE | GPA: 8.16/10 | *Anna University, India* **Apr 2019**

Certifications: Deep Learning Specialization, Python and Machine-Learning for Asset Management with Alternative Data Sets, Introduction to Machine Learning in Production, Machine Learning Data Lifecycle in Production, Credit Risk Modeling in Python

Projects

End-to-End Pyspark pipeline for US Medicare Data **Oct 2022**

- Ingested US Medicare provider data into an **HDFS** (Hadoop) single-node cluster and used **PySpark** to preprocess and transform the data to get distinct physicians in each US city categorized based on transactions and specialty. Extracted the transformed data, stored it in **PostgreSQL and HIVE** databases and transferred the data to **Azure Blob Storage** and **Amazon S3 bucket** to make data accessible for further analysis in the future **saving 30+ hours**.

Prefix Tuning Language Models on Noisy Financial Data (Accepted at the EMNLP Financial NLP 2022 workshop) **Sep 2022**

- Evaluated the robustness of **prefix tuning and fine-tuning** using language models like BERT and RoBERTa to find the best training method to build a financial chatbot for a major Canadian bank to **reduce technical debt**. Corrupted the clean **financial phrase bank dataset** with varying noise levels by simulating real-world noise in textual data and performed sentiment analysis to show **fine-tuning is more robust to noise** and found that prefix tuning has a high variance in F1 scores.

Predicting Household Energy Consumption **July 2022**

- Analyzed and cleaned the noisy US household survey data to predict household energy consumption and stored the data in a **SQLite database**. Discovered the non linearity in the data by verifying the assumptions of **linear regression** using the residuals plot. Developed an **XGBoost Regressor model** with an **R-square value of 91.6%** and served the model using a **Flask REST API** on AWS Elastic Beanstalk to get instant predictions based on the given input features.

Predicting Stock Movements using BERT **Dec 2021**

- Analyzed and preprocessed the **stock prices and tweets** of 88 companies and filtered the data based on **relative stock movement percentages**. Implemented and fine-tuned the **BERT** model by **reinitializing the top 3 layers** and used grouped layer-wise learning rate decay to **improve the baseline accuracy by 7%**.