

# SATHVIK RAJU

✉ [sathuraju@gmail.com](mailto:sathuraju@gmail.com) ☎ +1 628 688 7828 📍 San Francisco, USA  <https://www.linkedin.com/in/sathvikraju>  <https://github.com/SATHVIKRAJU>

## PROFILE

Data Scientist with varied experience in research and development of machine learning algorithms, Experimentation, creating end to end data science frameworks with in-depth understanding of Deep learning, Statistics, Machine learning, Bayesian Inference, Data visualization and Data wrangling. (Skills: Python, SQL, TensorFlow, AWS)

## EDUCATION

### University of Illinois at Chicago

Aug 2015 - May 2017

#### *Master of Science in Electrical and Computer Engineering*

*Related courses:* Neural Networks (CS 559), Machine Learning (CS 491), Statistical Digital Signal Processing, Detection and Estimation theory, Information theory, Advanced Computer Networking (EC 533), Probability and Statistics.

### Visvesvaraya Technological University

Sep 2011 - June 2015

#### *Bachelor of Engineering in Electronics and Communication Engineering*

*Related Courses:* Image Processing, Computer Algorithms, Signal Processing, Advanced Mathematics

## EXPERIENCE

### Senior Data Scientist, OnPoint, Vancouver CA , San Francisco, USA

August 2020 - Present

- Led implementation of data science workflows from data extraction, **statistical inference, machine learning models**, dashboards to model deployment. A senior lead to the company's machine learning product Cortex.
- **Managed and mentored data science interns** and process engineers. Played a pivotal role in growing the data science team at the company and creating value for the business through data science.
- Collaborated with key stakeholders to influence the product roadmap and successfully got buy-in from prospective clients for machine learning proof of concepts (PoC) on Cortex.

#### *Abnormal behavior detection and Recommendation diagnostics engine:*

- **Full stack** lead on developing the entire data science framework with 10+ different burner heaters including data ingestion, exploratory analysis , model building, model deployment and model maintenance.
- Model building using Gaussian mixture models, Autoencoders and Copula based outlier detection.
- Model deployment using SageMaker studio and **Causal inference** for higher anomaly points.
- This is currently one of the **highest revenue** generating products in the company, with each anomaly detected successfully reducing the operating cost by \$50K each time and reducing maintenance turnaround time by 6 months.

#### *Vegetation Failure forecasting and inference:*

- Forecast external failures in multiple electrical plants saving more than **\$10M in operating costs** and possible wildfires since deployment. Utilizing raw signal processing and crucial feature engineering built multiple ML classification models with custom metrics for validation and loss function with Probabilistic Neural Networks for key inference.
- Data Engineering pipeline for weather data, imagery and tabular data were implemented using AWS Sagemaker and EC2 instances and built SQL queries for efficient data quality and model performance.

#### *Type of Failure Detection (Energy grid):*

- Built a robust multi-classification model to detect the type of failure, enabling maintenance crews to schedule operations in a timely manner effectively reducing unit down times.
- A **1-dimensional CNN** for waveforms and Xgboost on tabular data were combined to give a more balanced accuracy score.
- Additional clustering with DBSCAN and Feature Engineering with AutoEncoders were used to effectively improve both the model and the data exploration.

### Data Scientist, OnPoint, New York City, USA

November 2017 - August 2020

#### *Machine learning product (Cortex):*

- **Developed several machine learning algorithms** from scratch for the machine learning product Cortex including KNN, Random Forests, Neural Networks, Bayesian Networks, LSTM, Model Selection.
- Implemented and modified **Deep Auto-Encoding Gaussian Mixture Models** from scratch using Python and Tensorflow. Scaled and deployable the model to tackle an unsupervised learning problem for anomaly detection. This PoC made it into the Cortex product to be used by non-data science users.
- Architected the pipeline for automation of data ETL and data preparation workflows. Established a **monitoring framework** to effectively visualize model outputs on being deployed.
- Developed key customer and product metrics that helped scope machine learning projects and ensured better product experience for the end user using **A/B testing and statistics**.

### ***Smart Combustion (Solex):***

- Real time forecasting, control and recommendation of NOx (*Nitric oxides*) values for the first ever single NOx burner called Solex.
- Lead Data Scientist in developing the solution from data extraction to model development and scaling using SVR, Xgboost
- The control optimization was executed using ***Markov Decision Process and Model Predictive Control***.
- The model was deployed widely on various burners using Programmable Logic Controller edge computing and saved close to \$1M in costs for the end client in the initial year of adoption.

### ***Unsupervised event prediction (Invista):***

- Researched and implemented an ***unsupervised approach*** to detect and predict system failure events.
- The challenges were met by implementing a combination of deep feature engineering and custom loss functions on LSTM networks and clustering some known events using ***DBSCAN*** and Deep AutoEncoders with Gaussian Mixture Models.

### **Data scientist, Privacera, Fremont, CA, USA**

**August 2017 - November 2017**

- Developed Natural Language Processing based topic modeling for document classification using Latent Dirichlet allocation (***LDA***).
- Implemented a custom named entity recognition to recognize different entities in a document for a cybersecurity client.

### **Data Scientist Intern, True Medicines, San Francisco, CA, USA**

**July 2017 - Sept 2017**

- Implemented various machine learning methods, feature engineering techniques to find patterns in the clinical data using Python.
- Developed a model to select the top most suitable medicines based on chemical composition with very low negative side effects.

### **Graduate Assistant, Department of Psychiatry, University of Illinois at Chicago**

**June 2016 - May 2017**

- Assisted with statistical analysis and inference of clinical trial data to assess the impact of drugs on alleviating anxiety disorders.

### **Data Science Intern, Utthunga Technologies, Bangalore, India**

**May 2015 - Aug 2015**

- Worked as part of the research team of 3 members, to analyze the performance of sensors to be deployed in an IoT network.
- Implemented a clustering algorithm to group individual sensors based on parameters such as temperature, voltage output, effect of physical parameters on sensors etc. by modeling real-time time series sensor data using R and SQL
- Designed an anomaly detection model that helped identify outliers which represented faulty sensors using Density based detection.

## **ACADEMIC PROJECTS**

---

### **Time series prediction on meteorological data from Beijing Airport using Deep learning on Python and Tensorflow**

**June - Sept 2017**

- Designed a multivariate time series prediction model based on ARIMA, ARMA for short trends in the data.
- Implemented the times series prediction on multiple features using LSTM neural networks and Deep learning.

### **Natural Language Processing to identify the source of food-borne diseases and suggesting the next course of action**

**June - Sept 2017**

- Designed a Supervised Learning model that incorporated Natural Language Processing for classification using Python.
- The model can be used to detect the food sources responsible for causing certain symptoms, and suggest the next course of actions that can then be taken to alleviate the symptoms

### **European Language Detection using Natural Language Processing**

**May - July 2017**

- Designed a Machine learning model for classifying 21 European Union languages with an accuracy of 97%.
- Logistic regression and Xgboost classifier were used to train the 5 GB dataset using Python.

### **Stock Market Prediction using Natural Language Processing**

**Feb - May 2017**

- Designed a Machine learning model that makes accurate predictions in stock market trends with the help of NLP, scraping data from Yahoo Finance and Reddit.
- Implemented a Random Forest based model and a Neural Network architecture based on the Backpropagation algorithm written in Python using Scikit Learn and Pandas.

### **Big data? A View from Information Theory**

**Aug - Dec 2017**

- Conducted State of the art research on how principles of information theory can solve the problem that relate to Big data technologies. Information coupling problem for reducing higher dimension data to lower dimension data was achieved.