# Shaunak Sen

**Data Scientist**

**Email**: shaunak1105@gmail.com | GitHub | Personal Blog | LinkedIn | Stackoverflow

## 💼 Work Experience

**Data Scientist | Sparkbox.ai, UK | October 2019 - present**

**Model building and optimizations**

1. Built **machine learning and deep learning models** and optimized them to increase performance by over 10 percent on certain product lines
2. Built a framework using **Bayesian Optimization** that enables **efficient hyperparameter search** over a large space of tunable parameters - this contributed to improving performance and reducing run-time and cost of production. Built an **ensemble of models** and a system to **search the best possible combination of models** and their respective weights that maximize performance
3. Built a model using **spacy** for **Named Entity Recognition** to extract relevant keywords from product descriptions and classify them into **Style, Material, Color** to better compare across different brands. Achieved an average F1 score of 0.83 on all brands
4. Implemented unsupervised learning algorithms to **cluster sales trends** of various department-category combinations and used the cluster information as a feature for the machine learning models

**New tools and applications**

1. Modeled a **constraint satisfaction problem** using Bayesian Optimization to help clients find the best price that satisfies a set of specified constraints - this help us improve margin by over 8% from the previous method
2. **Model Analysis Tool**: Built a **web application** using **streamlit** that helps developers and users track the performance of the ML models using a variety of **interactive charts** (like treemap, bubble chart, line plots, etc.) and get statistical reports. This helped **automate** a tedious and error-prone process using excel sheets and saved time for data and business teams,
3. **Explainable AI Toolkit**: Research into **Explainable AI** and developed a **web application** that allows users to **gain insights and trust** on each prediction and understand the marginal contributions of each feature towards the prediction
4. **Feature Analysis Toolkit**: Built a **web application** that automates the process of searching and evaluating the **best combination of features and end-to-end testing** of model performance using new feature sets. This helped us reach a reduced feature set that improved training time by 11% and reduced mean squared error by 7%
5. Developed a system to solve the common problem of **over-predicting sales** on certain products by using **Quantile Regression** to provide an estimate of the upper limit of sales. This helped reduce mean squared error by 25% in the over-predicting products

**Trainee Decision Scientist | Mu Sigma, India | July 2017 - August 2018**

1. Responsible for building various data engineering pipelines and architectures using Python and Microsoft SQL Server
2. Built an automation system using Python and SQL that helped us in end-to-end testing of a metric-driven dashboard, that helped reduce errors and saved 30 mins of manual testing time for the team on a daily basis
3. Underwent 8 weeks of training and mock projects at Mu Sigma University, where I learned the foundation skills for analytics, data science, machine learning, and statistics

## 🏫 Education

### MSc in Data Science

*University of Southampton, UK | 2018 - 2019*

- Graduated with First Class with Distinction (77%)
- Achieved the highest score of 84% on my Master's Thesis in our batch

### B.Tech in Information Technology

*National Institute of Technology, Durgapur, India | 2013 - 2017*

- Graduated with First Class with Distinction (C.G.P.A of 8.64 out of 10)

## 🎨 Skills

*Programming and problem solving*: Python, SQL, R, JavaScript

*Machine Learning, Deep Learning*: scikit-learn, PyTorch, keras, tensorflow

*Natural language Processing*: spaCy, nltk

*Data Analysis and Manipulation*: numpy, pandas, tidyverse

*Data Visualization*: plotly, Tableau, matplotlib, seaborn, ggplot

*Database*: SQL, MongoDB

*Software development*: Front and back end development using Streamlit, Flask, unit testing, API design and development

*Other skills*: Mentorship (mentor@Python Programmer discord), Writing (personal bog), teamwork, and communication

## ⚒️ Personal projects

**AI for Web Accessibility (Masters Dissertation) | University of Southampton, UK**

Explored, analyzed, and built systems for improving web accessibility:

1. Built an **Image Captioning System** that would **automatically caption images on a website**, without a suitable "alt" tag. Used CNN, LSTM, and Transfer Learning.
   a. Built a **text-to-speech web app** running this model to read out image captions from web pages
   b. Extended the system to evaluate **text-to-image similarity** to detect if the image present on a web page is related to the text surrounding it
2. Built a **Contextual Hyperlink Detection** system using NLP that would help a user detect if the hyperlink text is in context with the source and target URLs
   a. Used web scraping to build a custom text corpus of source hyperlinks, target hyperlinks, and their associated texts
   b. Built and trained a word2vec model using the created text corpus and improved results by 5% compared to pre-trained word embeddings

**Driven Data competition (Pump It Up) | University of Southampton, UK**

Collaborated in a team of 7 to solve this multi-class classification challenge to predict the operation of water pumps in Tanzania, Africa.

Implemented every stage of the Data Science pipeline - Data Collection, Data Cleaning, Exploratory Data Analysis, Feature Engineering and Selection, Machine Learning, and Visualization.

Extended the dataset by engineering features related to:

- Water level data at desired locations captured using open satellite data (GRACE by JPL)
- Demographic data related to health status, criminal activity in the location collected from Open Africa
- Weather data like temperature, humidity, wind, and pressure conditions

This helped us improve our benchmark performance by 2.79%

**Sequence Recognition System | University of Southampton, UK**

In this reproducibility challenge, I collaborated in a team of 3 to build a deep learning model based on Convolution Neural Network and trained it for multi-digit recognition from street view imagery, reproducing from scratch the work of Ian J. Goodfellow and others in the paper: http://bit.ly/2YUhX5W

We achieved an accuracy of over 95% at a coverage of 93%, which was comparable to the results the authors obtained

## 🎯 Accomplishments

### Tech Nation Exceptional Promise in AI Visa Award | Tech Nation | 2020

I was awarded the Tech Nation Visa in AI, which took into consideration previous work experiences, achievements, projects, and extracurricular activities as well as future goals and plans.

### Spot Award | Mu Sigma | 2019

I was awarded a certificate of appreciation for my work at Mu Sigma on automation and improving the architecture and runtime of my projects.

### Data Analysis in Azure | Microsoft Student Partners | 2020

Attended the machine learning workshop organized by Microsoft Student Partners and learned about various features of Azure Machine Learning Studio (certificate). Achieved the highest model accuracy in the workshop competition

### Food Waste Reduction | National Institute of Technology, Durgapur, India | 2015

Helped reduce food waste and mess bills by building an application for mess management of our hostel in National Institute of Technology, Durgapur, which was greatly appreciated by students and staff