AYAN SENGUPTA

@ ayan.sengupta007@gmail.com ☎ +91 9706318732 ⚲ Noida, India ⚲ https://victor7246.github.io/
in https://www.linkedin.com/in/ayans2017/ ⚲ https://github.com/victor7246

PROFESSIONAL EXPERIENCE

UnitedHealth Group (Optum)

Senior Data Scientist

📅 August 2019 – Ongoing   ⚲ Noida, India

Working with Optum's Risk, Quality & Network Solutions (RQNS) Team

- Working on building NLP solutions to extract structured information from noisy semi-structured scanned clinical charts which enables Optum in $1B+ in risk adjustment and Data as a service (DaaS) business
- Handling additional research activities on extracting Named Entity Recognition from noisy texts and deep learning based text denoising

Data Scientist

📅 May 2017 – August 2019

- Built and deployed pipeline to detect suspect, screening and hospitalization information for 7M+ member population across US
- Handled R&D activities on Voice of Customer Analytics and NPS Analytics
- Worked and presented work on Clinical Fraud detection to National Health Authority (NHA). POC was done for *PMJAY's Ayushman Bharat*

Corporate accolades received -
- 2019 innovation and 2018 Q4 mastermind awards for R&D initiatives
- 3rd place award for NPS Analytics in annual data science championship 2018

RECENT PROJECTS

Structured Information Extraction from Scanned Medical Charts

📅 Nov 2019 - Ongoing

- Extracting member and provider demographic information from noisy OCR outputs with NER using Transformers and Sequential models
- Extracting structured information from scanned medical charts using multi-modal deep learning methods leveraging image and text data
- Exploring knowledge graph embeddings incorporated into NER models

Denoising Noisy User Generated Texts

📅 March 2020 - Ongoing

- Working on categorizing and correcting noisy OCR and speech-to-text data
- Exploring Seq2Seq, Pointer Generator and Denoising Transformers

Voice of Customer Analytics

📅 Mar 2019 – May 2020

- Developed generative probabilistic model for detecting and prioritizing pain points from customer feedback
- Developed model for online topic detection from streaming feedback data
- Filed 2 patents and drafted 1 patent
- Submitted 3 research papers to core A conferences and prestigious journal in AI and Machine Learning

ML based Medical Chart Scoring

📅 May 2018 - Aug 2018

- Engineered categorical embedding based features and used GBM based models to classify medical charts into different scoring groups
- Used statistical model interpretation techniques to explain predictive models
- Improved false negative rate by 30% which enables Optum RQNS to save $70M+ of manual chart coding cost per year

EDUCATION

Post Graduate Diploma in Business Analytics

IIM Calcutta, IIT Kharagpur, ISI Kolkata

📅 Jun 2015 – May 2017

M.Sc. in Mathematics and Computing

IIT Guwahati

📅 Jul 2013 – Jun 2015

B.Sc. in Mathematics and Computer Science

Chennai Mathematical Institute

📅 Jul 2010 – Jun 2013

TECHNICAL SKILLS

- Development, DB

  Python R SQL Shell

- ML Frameworks

  sklearn PyTorch Tensorflow h2o

- Big Data

  PySpark MapReduce Hive Oozie

OTHER RESEARCH ACTIVITIES

- Program Committee member of ECML-PKDD 2020 for Applied DS track
- Paper "*A Large-scale Analysis of the Marketplace Characteristics in Fiverr*" accepted in HICSS 2016
- Paper "*Bayesian Fault Detection Engine for Predictive Maintenance in Data Centre Infrastructure Management*" accepted in ICBAI 2016

OTHER DATA SCIENCE ACTIVITIES

- Developer and maintainer of package *jointtsmodel*
- Part of 4 member team from ISI Kolkata in *International Data Science Game* 2016 finals in Paris and achieved rank 14 among 100+ teams
- 2 Bronze medals in Kaggle NLP challenges in 2020
- Ranked 16 among 50+ in Codalab satellite image classification challenge 2019
- Won $20K+ across various data science competitions in Topcoder between 2017-2020

OTHER ACCOMPLISHMENTS

- Hold KPMG Six Sigma Green Belt
- Recipient of *National KVPY scholarship* during 2009-2015 for pursuing research (top 200 among 50K+ applicants)
- Cleared *Regional Maths Olympiad* from W.B. region in 2008 (top 50 among 5000+)