TARAPRASANNA SAHA BABU

Data Scientist | HSBC - Bengaluru

Email : tara2010.p@gmail.com
Skype : taraprasanna_1
Mobile : +91-8695383939

www.linkedin.com/in/taraprasanna-saha-babu

Machine Learning

- Classification
- Regression
- Clustering
- Boosting
- Ensemble methods
- Optimization- Linear& Non Linear
- Feature Engineering
- Exploratory Analysis

NLP

- BERT
- LSTM
- RNN
- Topic Modelling

Big Data

- Hadoop
- Spark
- Scala
- Sqoop
- Hive
- SparkSQL

Cloud Technology

- GCP
- Docker

Analytics Tool/Viz

- SAS
- IBM SPSS Modeler
- Tableau
- Plotly - Dash

Others

- SAFe – Scaled Agile Framework
- Git

Representative Experience

HSBC   Bengaluru, India
Data Scientist   *July 2019 - Present*

NPS Live Chat

- **Objective**: To build a scalable self -service app which can provide actionable insight of NPS (Net Promoter Score) for different geographies (US, UK, APAC) along with text analytic capabilities.
- **Analytical Approach**: Text processing on Live chat/Branch contact survey data - lemma, pos, spell check, synonym replacements are done with spacy and symspell. Unlabeled texts are labeled by considering several steps such as Topic modeling, cluster samples based on soft-cosine similarity than assign top 5 words as a topic for each cluster. After fine-tuned the labels with doccano, BERT was trained with the labeled data for multi-label classification.
- **Business Benefit:** Business is able to identify what matters within the huge volumes of chats/surveys created and shared each day. With millions of chats a month it can be hard to make sense of what is being said and what's important to business. Our solution tracks the information which business needs to guide their communications strategy.
- *Tools/Libraries Used* – Python, BERT, Topic Modeling, Spacy, Gensim, Plotly-Dash, Flask, Symspell, Docker, Google cloud platform.

Premier Attrition Prediction

- **Objective**: To predict customer's likelihood to become inactive or any signs of disengagement (either through drop in TRB or decrease in no. of products held) in the next 3 months.
- **Analytical Approach**: For primary data wrangling, GCP's Big query framework is used. To identify key features, Feature Engineering is carried out. For qualitative interpretation, Shap values and PDP are used. As a classification algorithm, Lightgbm is used. Stability (PSI), intensity (KS, RO) and performance with lift scores were key monitoring metrics that were included.
- **Business Benefit:** The model is used as a part of retention strategy for customers.
- *Tools/Libraries Used* – Python, Pandas, Scikit-learn, Plotly, Google cloud platform

Wealth Management Product Mis-selling

- **Objective**: The Sales Quality (SQ) team is looking to find anomalous behavior of financial consultants to investigate further and assess for mis-selling indicators.
- **Analytical Approach**: Isolation forest is used as a modeling approach to generate an anomaly score for each financial consultant at Wealth Management.
- **Business Benefit:** The new tool either will replace the existing ERBA tool or will be an addition as determined by Global Head of Sales Quality.
- *Tools/Libraries Used* – Python, Pandas, Scikit-learn, Plotly, Google cloud platform

Cognizant Technology Solutions   Kolkata, India
Data Science Associate   *June 2016 – July 2019*

IBM SPSS Modeler to SPARK Machine Learning

- **Objective**: Client wanted to migrate their Machine learning models/streams from IBM SPSS Modeler to Spark – Big data environment which are running in production.
- **Analytical Approach**: Developed a python framework which can extract metadata of ML models, data pre-processing steps (Machine learning pipelines) from IBM SPSS Modeler. Python parser can automatically generate executable spark-scala code by taking inputs from Jython script which ran on SPSS Modeler. Developed complex generic function in Scala to replicate data transformation and modeling steps in SPSS Modeler. Practically, built a solution which can migrate all the streams (Machine learning models/pipelines) from IBM SPSS Modeler to spark-scala environment.
- **Business Benefit:** Reduced time required to run the job. Brought in efficiency of around 80% in migration and high accuracy.
- *Tools/Libraries/Languages Used* – Python, Jython, Pandas, Spark, Scala, Hive, Sqoop, Spark SQL

Kantar Media Clipping Implementation

- **Objective**: The project aims to implement the extraction of data from newspaper, magazine and articles (provided in the form of images) in the form of text boxes and club the text boxes belonging to a particular topic/news story together.
- **Analytical Approach:** The approach consists of implementation of optical character recognition to extract the various components of the newspaper and segregate them as text boxes, headlines and images in xml format. The methods used for clubbing together of similar text boxes are geometry and semantic based. The geometry-based method attempts to club the boxes based on their proximity and relative position.
- **Methodology:** The algorithm used in this case is hierarchical clustering implemented on a distance matrix and evaluated using silhouette score. The semantic based approach considers the use of NLP based text similarity approaches like cosine similarity, doc2vec, word movers distance, NER, topic modelling using latent dirichlet allocation and hierarchical dirichlet process.
- **Business Benefit:** Providing the information clients need to match their brand to the media that will reach their target market. Enabling advertisers, agencies and media owners to identify, target and reach key consumer audiences.
- *Tools/Libraries Used* - Python, Pandas, Scikit-learn, gensim, Seaborn.

- **VIT University**   Vellore, India
  M.Tech in Nanotechnology (Computational Science); CGPA: 9.22   *April 2014 – May 2016*

- **Sathyabama University**   Chennai, India
  B.Tech in Electronics and Telecommunication; CGPA: 8.25   *Aug. 2010 – March 2014*

- Band Gap Engineered Nano Perforated Graphene Microstructures for FET. The simulation has been performed using drift-diffusion semi-classical and tight-binding based non-equilibrium green's function (NEGF) methods. The obtained simulation results are compared with previously reported experimental work. Algorithms: First Principle, Slater-Koster (Python Environment – Quantumwise Atomistix Toolkit). AIP Publication link