About me:

Name: Argho Chatterjee

Sex: Male

Age: 30

Current Location: Bangalore

Contact no: +91 8898015070

Git: [https://github.com/argho-pabitra-chatterjee](https://github.com/argho-pabitra-chatterjee%20)

*LinkedIn*: <https://www.linkedin.com/in/argho-chatterjee-954b27ab/>

Email: [argho.chatterjee.001@gmail.com](mailto:argho.chatterjee.001@gmail.com)

*Brief Description:*

With a research background, I am enthusiast to solve hard prediction & reasoning problems using data to help business achieve and provide directions to solve hard business problems and it reflects in my work experiences.  
   
My recent work experiences includes working in data science teams for ecommerce companies[Pepperfry], telecom[Mobileum] and finance&SupplyChain[Zycus] domains and would like to continue my career in this field of advanced analytics. Thereby, I am looking forward to join and work in a team of core statisticians with data science background and ML experts working on similar hard prediction problems.

I have recently joined a company in Bangalore as a Machine Learning Scientist but found that the team is at a Naïve stage of formation and there isn’t clear approach and use cases to help business apply machine learning, in addition to **lack of data & data collection activity** process initiated, thereby I am looking forward to a relevant opportunity in my field area of work.

Professional Details:

*Skillset / Techniques:*

Machine Learning, Statistical Modelling, Big Data, Text Mining, Time Series Analysis, Natural Language Processing, Data Structures [List, Maps, Trees, Sets]

*Modelling Techniques Used:*

Regression Models [Multiple Linear Regression, Logistic Regression, Elastic nets],

Bayesian classifiers [Naïve Bayes, Mixture models],

Dimensionality reduction (PCA, LDA, QDA, FDA),

Tree Models (Decision Tree, Random Forest & XGBoost),

Kernel classifiers-Linear, Polynomial & RBF (Support Vector Machine),

Deep Learning [CNN, RNN, Deep Belief Networks],

Time Series [ARIMA, ARMA],

Association Mining [Apriori]

*Data Management techniques used:*

Hadoop [Hive, Pig, MR], SQL, Spark, MongoDB, Solr/Lucene, Memcached

*Languages and Tools used*

R, Python, Java, Azure ML.

*Area of Interest*

My company owned research areas includes working in areas of text mining (using Naive Bayes, SVM) and NLP (NLTK using standard Stanford Parsers) for structure discovery and feature extraction from raw texts/queries/customer emails and tuning them to rank in preference order of results to maximize a business metric.

Used statistical analysis and sampling (Regression, One&Two tailed Hypothesis tests) for exploratory analysis and inferences to find relationships between various groups of customer in churn, segmentation and targeting in ecommerce and telecom domain.

Used XGBoost and later RNN[Recurrent Neural Nets] for Multi-class classification problems for product click predictions in my ecommerce click prediction use case.

I got an opportunity to work on datasets for product ranking using conditional occurrences of keywords where I had modeled problems related to head to head ranking Competitors through Bayesian approach (Bayes Ranking and Gaussian Mixture Modes).

Applied Dimensionality reduction techniques to various analysis projects - PCA as a preprocessing step to customer segmentation and clustering analysis [reduces noise], applied LDA&QDA to increase class separation to separate overlapping Gaussians to help in multinomial classification problems.

*Previous Project Experience*

I started off my career working for Product Organizations building solutions to scale and handle large number of requests with optimum response time. I had used Java and other distributed database technologies to scale to large number of real time requests and performed space-time complexity on various parts of my code to improve response time with optimum resource consumptions.

I was moved to work on Analytics projects soon thereafter, post which most of my work involved working on Big Data Technologies to train Models for various analytics and product building activities. It has been now 4.5+ years in this field.

As a part of my Business Analytics course work at ISB Hyderabad, I had worked on many regression models, Bayesian classifiers, deep neural network on large datasets apart from my Capstone projects. ***To be brief, I am mentioning recent projects where I had worked.***

*Genpact Headstrong:*

***User Intent Mining:***

Mining intent form customer emails and text data has a lot of value to tap potential revenue and avoid revenue leakage.

I had built a 28 class classifier to predict the sentiment of the intent of the customer using multiple models [one using word2vec, other using 3-gram TF-IDF vectors]. Some Class labels were better predicted with word2vec, some class labels were better with TF-IDF feature vectors. We used Feature-Weighted Linear Stacking to make model prediction results better. This was deployed in Production.

#### *PEPPERFRY :*

#### Personalized Ecommerce Search: Ranking item / Product Click Prediction

This problem in the field of machine learning used multi-model approach using information retrieval techniques. I had trained models to rank-determine the likelihood of a product being clicked for a search query to improve the click to conversion ratio.

We created ranking functions by training models on dataset, with the goal to optimize the performance of IR applications in terms of various evaluation business measures and we got statistically significant improvements using learning based methods such as RankNet, RankSVM, and RankBoost.

I and my team had even tried out with XGboost and its boosted trees just were amazing, both in terms of efficiency and effectiveness.

We applied Neural nets as a part of improvement process where other classifiers were hard to train or critical to business measure improvements. This was deployed in Production.

***Bot detection pattern mining system:***

The aim of this project was to identify crawlers/bots that crawl the website in real time and display a Captcha to non-human user interfaces/systems.

The dataset (click logs) were unlabeled and cannot train a classifier to predict the result of a particular click from an IP address.

Approach used was to perform clustering on PCA transformed feature set to find out relevant cluster resembling homogeneous groups of different Bots and human users. Features like click rate, http referrer, unique URL hits, percentage of website coverage etc. post clustering were identified to be useful features to train classifiers on. Post cluster data analysis, we assigned individual labels to clusters to identify them training of classifier.  
Once cluster labelling was done, I used this new labeled data to train a model which learns the patterns of respective clusters to identify bot clusters from human users in real time. This was deployed in Production.

***Customer Segmentation***

We used hierarchical clustering to help business understand the structure of various group of our customers. Each customer was represented as a feature vector, based on the business metrics, post which we ran clustering on multiple customer sample dataset, to ensure generalization of our analysis.

*MOBILEUM:*

***Churn Prediction:***

Used SVM to predict if a customer is about to churn a telecom operator based on his historical and current behavior like subscription model type, travel pattern, class of subscriber, call drops, age of customer-relationship, min/max usage in last 7/30/60 days, etc. This was deployed in Production for 60 days as a part of POC on 7 months dataset. It was later productized as engineering product release.

***Bill Shock Model:***

The model tries to discriminate between high usage customers from bill shock customers, based on their usage and factors like the class of subscriber, continuous data usage factor level[normal/abnormal], usage threshold, etc. This model uses SVM to train. This was used as an input to churn analysis, as well to improve customer churn model.

***Fraud Detection:***

The model tries to predict fraud occurring in telecom network. Anomaly detection was used on a dataset in addition to feature selections using significance test, on a 270000 observations, to give us a reasonably good ROC curve for our Model.

*Previous Working Experience*

|  |  |  |
| --- | --- | --- |
| Company | Domain | Period |
| Genpact | Finance | Apr 2017 |
| Pepperfry | Ecommerce | Feb 2016 – Apr2017 |
| Mobileum Inc. | Teleco Big Data & Analytics | Jan 15 – Feb 16 |
| Zycus Inc. | Procurement Domain (Analytics team) | Dec 12 – Jan 15 |
| TCS | Finance (ecommerce Products of GE Fleet) | Dec 11- Dec 12 |
| AG Technologies | High Performance Engineering Materials | Oct 10 – Dec 11 |
| YSC | Banking and Finance Domain Products | May 09 – Oct 10 |

*Recent role/Job responsibility details:*

Genpact: Work involves building intelligent reasoning systems which uses Data Mining techniques at scale for Financial and Accounting Domain of Customers.

Pepperfry: Worked in data science team at Pepperfry for building their Search module, bot detection system and Card Fraud Analytics. models using Machine Learning and big data technologies.

*Mobileum*: Worked in Big Data and Machine Learning for predicting Customer Churn, telecom Fraud analytics, Bill Shock Prediction for roaming customers.

*Zycus*: Worked in automatic classification engine as a part of AI research team which automatically performs spend analysis as a SAS module for various customers.

*TCS*: Worked on backend systems using Java and implemented a Regression model used for predicting Quotations to end customer based on various User defined parameters, this was a part of the Quotation module for Client GE [General Electrics].

*AG Technologies*: Worked on back end development of products for Saint Gobain using core Java [involved usage of Concurrent Data Structures] which used to summarize various Research modules which Saint Gobain used to work.

*YSC*: Was member of development team of core banking solution development. YSC had its own MVC framework built in Java. I was involved building the backend system of CBA[Core Banking Application] and had good exposure to Data Structures, scaling the backend[Had improved space-time complexity across various parts of my code base]

Education Details:

*Business Analytics, Indian School of Business [1 year advanced analytics course], 2016-17*

It is a 12 months course [3 terms] with focus on advanced analytics and use of machine learning to model various challenges with the understanding of the associated statistics & required math.

*Bachelor of Computer Science Engineering, Mumbai University, 2004-08*

A general computer science engineering course program covering areas of hardware, software, networking, data structure, algorithms and other areas of computer science.

Other Details:

*Past time Activities*

* Football (before on the field, now a days on my xbox system)
* Kaggle (I have my own team for Kaggle, we participate, learn and enjoy thinking over new issues to be solved in the datasets)