Nashit Baber (Data Scientist)

Business minded, Coursera certified data scientist with demonstrated ability to mine hidden gems in the ocean of data.

nashit93@gmail.com

+91-8421000030

Skype Id: **nashit93**

About Me: nashit93.github.io

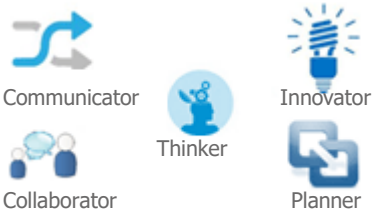
Key Skills

- Natural Language Processing, NLP
- Data Exploration & visualization
- Statistics & Probability
- Data Science & Predictive Modelling
- Requirement Gathering
- Hadoop Ecosystem (Mainly Pyspark)
- Deep Learning (Dense,CNN, RNN)
- Structured and Unstructured Database
- New Product Development
- Python

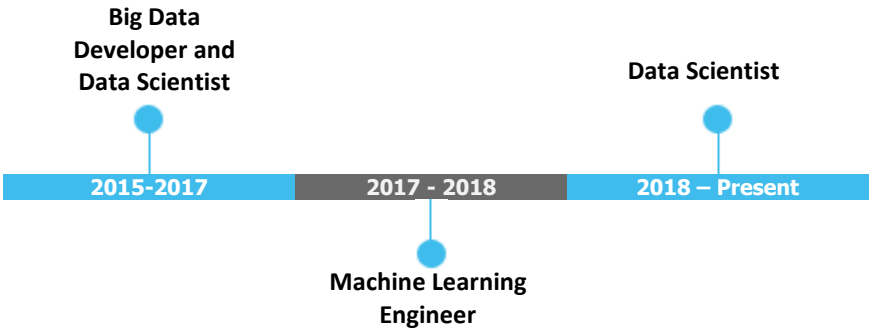
Profile Summary

- A competent professional with **3.6 years** of rich experience in **Data Science (Healthcare, Legal and Compliance and Financial domain).**
- Proficiency in Python, Shell Scripting, Pyspark, Keras, Deep Learning, Hadoop Ecosystem (including HDFS , Hive , Sqoop , Impala), Tableau , R, Regression, Classification, Clustering, Pandas, TensorFlow, NumPy, CNN, RNN, Word2vec, Doc2Vec, NLP, NLTK , Genism, Spacy , SciPy , NumPy , BS4 , Selenium, MY SQL,LXML.
- Effective in implementing data mining and statistical learning solutions to various business problems in Finance /HealthCare/Legal such as Stocks price prediction of Indian Stock Market, Suggesting Entry and Exit points in the given market condition, forecasting Market Conditions, Real World Evidence Apps and LegalView Bill Analyser.
- Keen analyst with excellence in gathering and understanding requirements of clients & other multiple stakeholders (at strategic and tactical levels), followed by translation into functional specifications as well as provisioning of suitable solutions that impact key business metrics as well as managing the business processes in an efficient and cost-effective manner.
- Comprehensive knowledge of **Regression, Optimization, Decision Trees, Boosting, Clustering, Classification Techniques, Deep Learning, Text Mining and Natural Language Processing, Over Sampling/ Under Sampling Techniques such as SMOTE, ENN etc.**

Soft Skills



Career Timeline



Education

2015 **Bachelors of Engineering** from Modern College, Pune University (61%).

Notable Accomplishments Across the Career

- **Text Processing/NLP - Researching and applying various Deep Learning, Machine Learning, Under sampling and Over Sampling techniques to improve the model's precisions and recall.**
- **Migrating a production ready code from Python to Pyspark to improve code execution time.**
- **NLP - Document summarization and extracting keywords using Rake, Text Rank and TF-IDF which helps attorneys to read and to get the theme of the document easily.**
- **Classification - Analyzing and designing the model for lien priority and positioning using Pyspark and Random forest.**
- **Creating stocks prediction system from scratch using various Deep Learning Techniques.**
- Design, Prepare and setup cost Effective, Fully Automated web scraping system to automatically extract, Daily Stock Data, Daily News, Bhav Copies, Advances and declines, Balance Sheets, Cash Flows.
- Designing and Developing a completely automated system for News Sentiment Analysis for stocks, also using state of the art LSTMs on historical Data to predict stock future prices.
- Worked with world renowned pharmaceutical companies like **Astellas Pharma** and many more and delivered great success stories in terms of providing solution to their real problems.
- Provided thought leadership & technical management in defining, implementing, ensuring, measuring & continuous improvement of Data Science and Analytics Practice with quality standards, frameworks & tools.

Work Experience

Mar'18 – Present **Wolters Kluwer Financial Services, Pune**
Data scientist

April'17 – Mar'18 **Hezkrost Technologies (Exadatas), Pune**
Growth Path:
Co – Founder and Machine Learning Engineer

Aug'15 – April'17 **Saama Technologies, Pune**
Growth Path:
Intern, Associate Consultant.

Key Result Areas:

- Data exploration, Anomalies detection, Featuring Engineering, Text mining, modelling.
- Designing and developing predictive model to predict lien priority, LegalView Bill Analyzing.
- Designing and coding algorithm which does sentiment analysis for around 1800+ stock's News data.
- Designing and developing Stock Price Prediction and Recommender system.
- Developing data modeling and analysis for researching and evaluating emerging tools and technologies which helps our organization to stand in the competitive market.

Professional Project Details

Saama Technologies

Project: Cohort Builder (Oct'15 – Jan'16)
Objective: Managing Target Patient Populations
Cohort Builder is a cohort generation solution that identifies patient populations based on inclusion-exclusion criteria. Cohort Builder organizes real world patient information based on diagnosis, drug therapy, and procedures to pinpoint population groups easily that fit specified criteria. The solution summarizes incidence rate ad co-morbidities of a disease and various other target metrics across a target cohort.
Technologies Used: HDFS, Sqoop, Pyspark, hive, impala-shell, python, Shell Scripting.

Project: Treatment Pathway (Jan'16 – Mar'16)
Objective: Select and Analyze patient cohorts in longitudinal observational patient records
Treatment pathways application helps in understanding the patient treatment sequence for any disease and the standards of therapeutic care provided in real world. It can also compare and contrast clinical guidelines with real world patient treatment utilization. It also helps in understanding the duration of treatments, switches and treatment sub-sequence. Also, helps in building your knowledge of relevant patient sub-groups.
Technologies Used: HDFS, Sqoop, Pyspark, hive, impala-shell, python, Shell Scripting.

Project: Population Characteristics (April'16 – July'16)
Objective: Understand incident rate and co-morbidities of a disease across a target population/cohort.
Population Characteristics helps in understanding the patient incident rates and also focus on the other disease that a target population may have.
Technologies Used: HDFS, Sqoop, Pyspark, hive, impala-shell, python, Shell Scripting.

Project: Real World Evidence (Aug'16 – Dec'16)
Objective: Integrating Social Media Data into Real World Evidence.
Using twitter feeds to build a rapid, inexpensive way of assembling large amount of data reporting patient's real experience of treatment, including drug efficiency, safety and drug impact.
Technologies Used: HDFS, Sqoop, Pyspark, hive, impala-shell, python, Shell Scripting.

Project: Data Migration (Dec'16 – April'17)
Objective: Migrating fresh and incremental data from Teradata to HDFS and creating scripts for Extraction, Transformation and Loading using Pyspark.
Technologies Used: HDFS, Sqoop, Pyspark, JDBC Drivers, Shell Scripting.

Hezkrost Technologies (Exadatas)

Project: RPA and Web Scraping (April'17 – March'18)
Objective: Preparing, Planning, setup and execution of a cost-effective ways of automatically scraping, cleaning, processing stocks data for 1857 stocks, their Bhav copies, Daily high/lows, Advances/ Declines, Bulk Deals, Income Statements, Balance Sheets, Cashflows etc.
Technologies Used: Python, Selenium, Beautiful Soup(bs4), lxml, pandas, NumPy.

Project: Data Mart Preparation and Maintenance (April'17 – March'18)
Objective: Fresh and Incremental loading of the above scrapped data into HDFS. Then cleaning and validation of daily updated stocks.
Technologies Used: Python, HDFS, Hive, Pyspark.

Project: Stock Prediction using HFT Algorithms and Machine Learning techniques (April'17 – March'18)
Objective: Coding highly efficient HFT algorithms to find real time Buy/Sell triggers of around 1800+ stocks. Used various technical algorithm such as SMI, MDA, oscillators, Bollinger, Pivot Points, RSI, super trend etc. to make real time buy/ sell decisions. Apart from real time triggers, Sentiment analysis was done on news data using machine learning techniques and reinforcement learning algorithms were used on historic data to predict the future prices and certain weightage was given to these during real time triggers.
Technologies Used: Mainly Python, NumPy, Keras, Deep Learning.

Wolters Kluwer

Project: Keywords Extraction (Mar 2018 – May 18)
Objective: Read the Legal document and try to summaries that using AI. Every time new documents are updated/published lawyers must read the whole document to find the rules. Created a Keywords Extraction algorithms which will process the document and highlight the key phrases. Document Tagging is also required to say it is relevant or not. Created a dictionary for a sematic match of the words domain specific Ontology is required on which we are working.
Technologies Used: Json, Oracle Db, Reg Ex, Rake, TF-IDF, Text Rank, K mean Clustering, word embedding, word2vec, doc2vec and Python.

Project: Lien Priority and Positioning (May 2018 – Aug 2018)
Objective: Every Business unit/ secure party wanted to analyze their Debtors so that they can understand the customer behavior and prioritize the lien. Is it safe or not? Data need to be clean on the name match because of variation. we used Fuzzywuzzy, Levenshtein distance and regular expression to clean the name. After that using Sqoop bulk insert operation of data is performed. CDSW is the platform on which we are analysis our data using Pyspark. Creating multiple trigger on the data to see the lien priority and its positioning.
Technologies Used: CDSW, Pyspark, tableau, Oracle ,Python, Edit distance Algorithms ,clustering (K-means)

Project: Legal Bill Analyzer (May 2018 – Present)
Law firm charges the amount based on the work they done, they generate the invoice which consist of the line items. We need to identify that that line item is charged correctly or not means that invoice need adjustment or not. Traditionally group of lawyers sit and review those bills but now we are very close to solve that problem using Machine learning and NLP. Based on the line item description which is a text field we apply NLP and generating a lot of features using word2vec or doc2vec. And once that feature is generated we apply machine learning technique like Random forest and XGBoost to predict adjusted or not.

Technologies Used: Python, Word2vec, doc2vec, LDA (Topic modelling), Random Forest, XGBOOST.

Project: Legal Bill Analyzer (Model Re-Building Using Deep Learning) (May 2018 – Present)
Law firm charges the amount based on the work they done, they generate the invoice which consist of the line items. We need to identify that that line item is charged correctly or not means that invoice need adjustment or not. Traditionally group of lawyers sit and review those bills but now we are very close to solve that problem using Machine learning and NLP. Based on the line item description which is a text field we apply NLP and generating a lot of features using word2vec or doc2vec. And once that feature is generated, we apply machine learning technique like Random forest and XGBoost to predict adjusted or not.
Technologies Used: Python, doc2vec, Deep learning (Dense layers, CNNs, Bi Directional LSTMs).

Personal Project Details

Project: Home Automation
Objective: Creating a home automation system which will use clusters of Raspberry Pi to control house and daily work something like Jarvis. Based to certain observations it will learn preferences over news, give intelligent reminders for work, meeting etc. Intelligent temperature and appliance controller which will switch off / on your appliances whenever they are not needed.(Still in POC Phase).Trailer Here!

Volunteer Works

Project: FreeJee.com
Objective: Co-founder of FreeJee.com. Although at a very early stage. We are dedicated to providing free education across the globe for all level of students (Absolutely Free! No copy rights or sign ups!)

You can check my current work here!

When I am not doing Matrix Multiplication

- I Love swimming
- Playing badminton & table tennis
- Playing Assassin's Creed
- And I love to Travel

Personal Details

Date of Birth:	07-04-1993
Languages Known:	English, Hindi
Mailing Address:	B2-602, Bramha Emerald County, Kondhwa, Pune-410048, Maharashtra

Links

GitHub : https://github.com/nashit93
Stack : https://stackoverflow.com/users/6323230/nashit
LinkedIn : https://www.linkedin.com/in/nashit-babber/
Me : https://nashit93.github.io/