Amit Sharma

C: 1-312-479-5727 | asharm36@hawk.iit.edu

PROFESSIONAL PROFILE

Experience with **Artificial Neural Network**, **Deep Neural Network**, **Natural Language Processing**, **Chatbots**, TensorFlow, **Prediction**, **Clustering**, **Recommendation Systems** and Narrative Science. Research and Development of machine learning pipeline, design of **Optical Character Recognition** (hand written) and **Anomaly Detection** System using **Multivariate Gaussian Model**. Designing and developing consistent reporting packages and dashboards using various BI solutions. Key team player and working with management and software development team to develop data science models as per business requirements with the Agile Methodology.

EDUCATION

- **Master of Computer Science: Specialization in Data Science** — **May 2017**
  Illinois Institute of Technology — Chicago, Illinois
- **Bachelor of Technology: Electronics Instrumentation & Control** — **May 2014**
  Rajasthan Technical University — India

TECHNICAL SKILLS

- **Machine Learning:** Linear Regression, Logistic Regression, Decision Tree, SVM, Naïve Bayes, kNN, K-Means, Random Forest, Factor Analysis (EFA, CFA), Principal Component Analysis (PCA), Gradient Boosting, Word2Vec.
- **Natural Language Processing:** Stemming, Lemmatization, Word Embedding, POS tagging, Named Entity Disambiguation and Recognition, Sentimental Analysis, Semantic Text Similarity, Text Summarization.
- **Programming Languages:** Python (Numpy, Scipy, Pandas, Matplotlib, Scikit-Learn, TensorFlow, Keras, NLTK, Gensim, Scrapy, NetworkX, Bokeh, Seaborn, Flask, Django, PyQT), R, SQL, PL/SQL, java, C++.
- **Big Data Technologies:** Amazon Elastic MapReduce, Apache Hive, Apache Spark, PySpark, Apache Pig, MapReduce, Hadoop, MongoDB.
- **BI and Analytics Tools:** Tableau, SAS, WEKA, H2O, DataRobot, Rapid Miner, MATLAB, Qlikview, Qliksense, Tableau, Anaconda, jupyter notebook.
- **Databases:** Oracle, MySQL, SQL Server, ProstgreSQL, DynamoDB.
- **Web Application Development:** HTML, CSS, Javascript, JQuery, Bootstrap.
- **Other Skills:** Microsoft Office, AWS, Linux, Descriptive Statistics, Inferential Statistics.

WORK EXPERIENCE

Data Science Consultant at Cognive.com — **July 2018 – Present**

- Developing a data pipeline using AWS S3, Sagemaker with boto3 SDK for python.
- Assisting in developing False Positive Reduction and Unknown Threat Detection System.
- Forming an Entity Resolution Model with **Deduplication, Record linkage, Canonicalization.**
- Performing bagging for creating dummy data and handling highly biased dataset.

Data Scientist at LTI — **Client: Citi Group (Oct, 2017 – Present)**

Project Name: Case Review Automation (CRA) in Anti Money Laundry Department.

Anti-Money Laundering teams have the responsibility to monitor all activities occurring throughout their institution in search of behavior consistent with money laundering. Automatizing the Case Review process in Anti Money Laundry for possible SAR escalations and case closings.

- **Data processing** and cleaning using **Pandas**, **NLP**, **R** and **SQL.**
- Developed **Entity Resolution System** using **Logistic Regression**, **Python** and **MongoDB** for escalating high risk transactions.
- Standard Industry Practices for utilization of financial systems, reporting systems, research tools and data extraction activities using **Hive**, **PySpark**, **Hadoop**, **SAS** and **SQL**.
- Developed an algorithm to classify transactions as a case of structuring or non-structuring with **Time Series Analysis** and **Time Value of Money** by developing **ARIMA** model using **Python (numpy, pandas, Scikit-Learn, matplotlib, seaborn).**
- **Clustering** and **Association** models were developed to stablish a relationship between transaction amount and type of transactions.
- Developed geo-spatial clustering model to find geographic and cultural correlation between transactions using **NewtorkX**, **googlemaps API**, **K-means**.
- Models were creating on **data robot** and **H2O** for quick deployment and testing for feedback.
- Successfully implemented CRA and decreased the time of the process by **18%**.
- Visualization and reporting using **Matplotlib**, **Seaborn**, **Bokeh, SAS** and **Tableau**.
- Appling Machine Learning Algorithms like **Clustering**, **Regression**, **Classification** and **Recommendation**.
- **Web Scrapping** social & political information about the site using Python's **Beautiful Soup** and **Scrapy**

Project Name: Natural Language Generation in Cognitive Automation.

Harvesting power of AI and Natural Language Processing to develop a hybrid system for generating reports from data points in Anti Money Laundry Case Review System. Applying NLG to the narrative component of Suspicious Activity Reports can assist in providing new insights into the fraudulent activities.

- Data Ingestion from different sources (structured and unstructured) using **ElementTree XML API, JSON parser/loader, XLRD, Pandas, SAS7BDAT, SQLAlchemy.**
- Data Analysis for **correlation**, temporal & spatial analysis and creating **Prediction** and **Recommendation** Engine.
- Using Natural Language Processing for surface realization using **POS tagging**, **Word Embedding**, **Sentiment Analysis** and **text summarization** using **Hidden Markov Model (HMM)**.
- Creating **REST API** for NLG using Python's **Flask** framework.

Project Name: Informational Services Group (ISG)

ISG provides authoritative sources of reference data to the clients across the Institutional Client Services (ICG) organization in Citi, ISG promotes collection, Storage, analysis, Distribution of products, pricing, accounts and corporate action data. ISG is central authority providing golden sources of data to the bank's critical systems for franchise critical activities.

- Maintenance of **Cloudera Hadoop** cluster.
- Checking and maintaining the docker configuration.
- Troubleshooting performance issue and creating test cases.
- **Data Integration** and **Query optimization**.
- Used JIRA for project tracking, Bug tracking and Project management.
- Involves in frequent communication with system users to determine specific feature expectations, resolution of conflict or ambiguity in requirements as demanded by the various users or groups of users
- Manage the version migration process as code moves through the development and testing phase.
- Work on **SQL** and **UNIX** shell scripting.
- Handle UAT and Production releases.

Data Science Intern at ML Wiz Inc., New Jersey **June 2016 – Oct 2017**

Online and In-class Data Science and Machine Learning platform for data science aspirants. Offers data science assistance to companies or clients as well as learning and training service to employees of the companies or clients and general public.

- Interaction with business contacts and understanding the scope and needs of the project.
- Identify, locate and access relevant data using **Python, R, Hadoop, SQL.**
- Performing big data tasks using Python and R libraries for **Hadoop** (**MRJob**) and **Spark** (**PySpark**).
- **Data processing** and cleaning using **Pandas**, **NLP**, **R** and **SQL.**
- Visualization and reporting using **Matplotlib**, **Seaborn**, **Bokeh, SAS** and **Tableau**.
- Appling Machine Learning Algorithms like **Clustering**, **Regression**, **Classification** and **Recommendation**.

Data Analyst Intern at Jaipur Development Authority, India **July 2014 – Aug 2015**

JDA is responsible for land acquisition and distribution and various other public and private projects. Land prices are decided by the certain parameters as that of the development and facilities in that area.

- Collecting data from government records, news feed and private real estate companies.
- Predicting authenticity of the customers on the basis of their past track record by developing customer-oriented database.
- Generating reports and visualization for further data pipeline.

PROJECTS

CNN model for image classification using TensorFlow

- Created **CNN** model using two sets of 2 convolution layers followed by **Relu** activation and **max pooling** layers and 2 **fully connected** layers and **softmax** with cross entropy and **gradient descent optimizer** achieving 82% accuracy using **Python, TensorFlow, AWS, Numpy.**
- Use of **FC dropout** layers gave 74% and further accuracy was increased by 10 % using **data augmentation**.

Bike-share user prediction using Neural Network

- Created a **Multi-Layer Perceptron (MLP)** model with two hidden layers to predict the bike rental price.
- Developed **Feed-Forward** network and **Back-Propagation** using gradient descent with on **Hidden Layer** to predict number of bike share on a given data using **Python**, **Numpy**, **Pandas**, **Matplotlib**. Accuracy of 81% was achieved by controlling number of hidden neurons, epochs and learning rate.

Real-Time Presidential Vote Prediction

- Real time tweets about Donald Trump were collected using **Twitter Streaming API** to know the thoughts of the people on twitter.
- **Fetched Realtime tweets** with tag "Donald Trump" using **TwitterAPI** and stored in a text file.
- Communities were formed using friend lists of each of the person whose tweets were collected and were plotted using **Matplotlib** and **NetworkX**.
- Tweets were **classified** into positive or negative using **sklearn CountVectorizer and LogisticRegression**.

Content Based Movie Recommendation System

- Using genre of movie as the content for the movie we try to predict and recommend new movie to the user depending upon his reviews on the similar movies.
- Similarity is defined by the **tf-idf (term frequency–inverse document frequency)** which is the weighting factor for our document.
- Storing the data in data frames using **Pandas**.
- Weights that are produced using **tf-idf** are stored in a **csr-matrix.**

Facebook Community Detection and Link Prediction

- “Like” data for Bill Gates was collected using **FacebookAPI** and was done for one more hop. **Girvan Newman** was implemented for **Community Detection** and **Recommending** new links or friends using same approach.

Cloud Enabled Distributed Task Execution Framework

- Developed a **Python** Framework to execute large number of fine granular tasks using **AWS** like **SQS** and **DynamoDB**.
- Built an app which converts image URLs into a 1-minute video and store them in **S3** for users to download.

REST API for store management

- CURD (Create Update Read Delete) API for managing stores and items available at stores using **Python**, **Flask** and **Flask-SQLAlchemy**.

Paper Presentation

Sharma, Amit (Jan – March 2015) “**Transitioning to IPV-6**” International Journal of Computer Science and Technology (IJCST), vol-6.1, v2.

http://www.ijcst.com/vol61/1/42-Pulkit-Gupta.pdf