Akshay **Sehgal**

www.akshaysehgal.com
akshaysehgal2005@gmail.com
Linkedin, Github, StackOverflow
+91-9916594778

Data Scientist with over nine years of experience, currently working as a **Manager, Data Science** at **AI Garage, Mastercard**, where I research, design, train and deploy ML models powering enterprise scale platforms and products. I have experience in working closely, business leadership, researchers, full stack developers, product teams and dev-ops to map out and productionalising AI/ML models in cloud-based products & platforms for Risk, Finance, Retail and Agritech domains.

I hold the current top 0.1% global ranking on Stack Overflow for Python, Numpy, Pandas, Sklearn and Tensorflow

I am also a veteran educator in the field of AI and Data science, have been teaching advance ML and deep learning since 3+ years now, including being the lead faculty for PGD AI/ML @ IIIT-B+UpGrad.

Previously, I was heading the Data Science initiatives as a **General Manager, Data Science** at **Reliance Industries** where I managed and mentored a large team of 15+ data scientists. I have significant entrepreneurial experience as **co-founder/Head of data products** for a 40 Cr valuation startup in ML domain called iPredictt Data Labs. My career in the data science domain started off in Mu-Sigma, a pure play analytics firm where I was groomed to solve large scale business problems with data for Fortune 500 companies.

Stack & Algorithms

- **Deep Learning –** Graph neural networks, GCN, GraphSAGE, Temporal Graphs, Computer vision, Image segmentation, Language modelling, NLP, RNNs, LSTMs, Word2vec, GloVe, Transformers, Encoder-decoder with attention, Variational Auto-encoders, U-net, DeepFM, GANs, Genetic algorithms, Reinforcement learning, Deep belief networks, Self organizing maps, RBMs, Deep dream networks.
- **Classic ML -** Generalized linear models, Ensemble models (Stacknet, Xgboost, Catboost, RF), Tree based models, SVMs, K-means, Gaussian mixtures, Heirarchial DBScan, TSNE, PCA, Matrix factorization, Probabilistic models, Network analysis, Markov models, Conditional random fields, Forecasting (ARIMA, S-ARIMAX)
- **Libraries -** Numpy, Pandas, Collections, Itertools, Scikit learn, Tensorflow2, Keras, Pytorch, Scipy, Django, Flask, Multiprocessing, Pyspark, Numba, Selenium
- **Languages -** Python3, R, SQL
- **Deployment -** Ngrok, Docker, AWS EC2, AWS Lambda (Zappa), Azure
- **Tools -** Anaconda3, Jupyter Notebooks, Pycharm, Sublime, Homebrew, Tableau, PowerBI

Outside office I explore metaphysics, epistemology, theoretical physics, recreational mathematics, and graphic designing. I have been a guitarist for over 12 years now and my lockdown hobby is baking.

Experience

Year (Start)	Company/Institution	Role
2020 (Current)	MasterCard	Manager, AI Garage (Senior AI Specialist)
2017	Reliance Industries	General Manager, Data Science (SME-1)
2015	iPredictt Data Science Labs	Co-Founder / Head of Data Products
2012	Mu-Sigma	Sr. Decision Scientist
2008	Pune University	B.E Comp Science (2012)

Achievements

- Under top 0.10% global ranking on Stackoverflow (0)
- Adjunct Faculty with IIIT-B/UpGrad PGD, AI and previously educator with INSAID & Digital Vidya for AI/ML(1)
- Frequent contributor on Digital Vidhya, Code Gladiators, Kaggle (medalist). (1)
- Participated as a speaker at multiple tech events across India. (2, 3, 4)
- Have 3 technology patents under my name (201721005644, 201621034521, 201621034522)
- Have 7 patent pending papers in the domain of Deep learning and ML
- Have been interviewed multiple times as a leader in AI/ML industry. (5, 6)
- Delivered multiple lectures as an expert speaker. (7, 8)

Projects

Fraudulent “Ghost” Merchant prediction for Acquirer network

June 2021 – Ongoing

Developing a network score to help Acquirer banks identify fraudulent merchants who pop-up in the network for limited time and contribute to direct or indirect fraud (subsequent fraud transactions on a card after testing). The score is subscribed to by Acquirer banks to help protect them against a variety of new-age fraudsters and is a core C&I (Cyber security & Intelligence) COE product offered by Mastercard globally. Currently working on building data pipelines using spark to identify ghost merchants in historical data in order to train robust fraud detection models based on a merchant activity in the first month of their onboarding.

Tools used: Python3, Pyspark, Hue, Impala, Fraud Detection, Neural networks, Binary classification, Temporal networks

“Craving” subgraph clustering for Food Recommendations

July 2021 – Ongoing

Building a general-purpose recommendation engine model utilizing graph neural networks to cluster ingredient sub-graph of the items currently in user’s basket. Based on exploitation/exploration strategy, the neural network’s goal is to identify what a customer is likely to order based on ingredient composition of their basket, in order to promote cold-start recommendation as well as food exploration.

Tools used: Python3, DGL, Pytorch, Tensorflow2, Sub-graph embeddings, GAT, GCN, LSTM, Node2vec

Virtual Cold-Chain assistant for Produce & Farm Management

June 2021 – Ongoing

Collaborating with Data.org to build a computer vision powered virtual assistant for Indian farmers to optimize decision on farms produce and gain access to sustainable off-grid cooling storages for reducing loss of produce and recover operational costs. The model is part of a largescale MasterCard initiative to help farmers estimate most energy and cost-efficient strategy to store/sell their produce. The architecture involves a multi-regression-based spatio-temporal graph model to forecast market rates for commodities, in conjunction with a computer vision-based pipeline to identify and categorize available produce with a farmer, flowing into a non-linear optimizer under constraints and a natural language generation system to make recommendations via an application interface.

Tools used: Python3, DGL, Pytorch, Tensorflow2, Temporal-GraphSAGE, Temporal-GAT, VGGNet, U-Net, Self-attention, encoder-decoder LSTM, Word2vec

Adversarial Attacks on Graph Neural Networks

Feb 2021 – Ongoing

Developing a solution to identify compromised merchants and cards based on transaction data. Based on the graph connectivity of merchant and customer nodes, we model potential fraud by gauging known fraud nodes and utilizing label propagation methods to induct a component of risk to adjacent nodes. Exploring Adversarial attacks on the graphs with topological attacks and node poisoning attacks on the graph data, to further create defence strategies to protect the fraud detection GNNs against such attacks.

Tools used: Python3, DGL, Pytorch, Tensorflow2, Multihead Attention, GraphSAGE, Graph attention networks, GAT, GCN, Deepwalk

Recommendation engine for Restaurant drive-thru’s

Mar 2020 – Feb 2021

Developed a recommendation engine with for multiple major US based restaurant chains using Deep factorization machines initialized with embedding representations and designed a custom soft-switch equation to optimize revenue vs conversion maximization. The model uses n-item recall for conversion maximization as an objective function and models the interactions between items, weather, time of the day and product metadata. The model is currently in production testing and has a potential of $2 margin gain over each transaction on average based on a simulation study we performed. Currently submitted a paper to a top tier conference on the problem that uses a variation of graph neural networks and a self-designed hierarchical attention mechanism to predict multi-level recommendations.

Tools used: Python3, Tensorflow2, Matlab, DeepFM, DeepFFM, Skipgram encoder, Catboost, Random Forest, Multihead Attention, GraphSAGE, GCN.

CIKM Adversarial Challenge on Object Detection (Competition)

July 2020 – Oct 2020

Competed in an AI challenge by Alibaba-Tsinghua on fooling an object detection algorithm using patched images. The competition involves over 1500 teams and our current ranking stands at 80. The competition marks an important step in improving defences against adversarial attacks over image classification and object detection algorithms. Landed in the 10% of the winning solutions.

Tools used: Python3, Pytorch, Yolo4, Fast-RCNN, Dpatch, Adversarial Robustness toolbox (ART)

Proctoring video interviews using image processing

Mar 2019 – Jan 2020

Built a video analysis platform for recordings collected from python based conference tool Jitsi. The capabilities include face detection, face landmarks, face orientation, face recognition and matching, audio extraction, transcript extraction, audio pause detection, topic modelling and other minor features. The video frame level signals act as an input for unsupervised anomaly detection methods to detect out of ordinary behaviour and compared against a rule engine to flag areas of interests for the hiring manager.

Tools used: Python3, CV2, open-cv, DBlib, Haar classifiers, U-Net encoder-decoder based anomaly detection, Tensorflow2.

Natural Language querying of databases

Dec 2018 – May 2019

Built a python framework which allows natural language querying on small-medium scale databases by using seq2seq neural networks to translate a query into a SQL query. The model is capable of predicting search and condition columns, conditions and aggregations needed in the sql query which is then run on the given database. The result is used with natural language generation to respond to the user as an answer to the query.

Tools used: Python3, NLP, Word embeddings, Seq2Sql, Seq2sql with attention, SimpleNLG, Xsql framework, Keras, Scikit-learn.

JD-CV matching algorithm for candidate shortlisting

Aug 2018 – Feb 20

Built a CV sourcing and shortlisting platform that allows hiring managers to access a ranked order of profiles matching the requirement. These profiles are enriched using multiple data sources and are parsed to extract education, experience, skillsets, project and personal information from the profile. This is followed by document clustering to obtain relevant domain cluster, and document similarity (ranking) algorithms to match JD document to profiles. A reinforcement learning layer is being added to capture and personalise hiring manager preferences and behaviours, while ensuring company standards and requirements.

Tools used: Python3, NLP, Word embeddings, t-SNE, Doc2vec, PCA, Spacy, fuzzy matching, GMM, document classification using LSTMs, reinforcement learning, Keras.

Distributed Virtual Assistant Development Toolkit

Jun 2018 – Dec 2019

Developed a python based tool allowing non-technical users to design, train and deploy closed domain virtual assistants using a GUI. The bots were integrated into a meta-model that allows intermediate intent switching to an intent on another bot deployed on some other server. Also, allows users to integrate APIs during any part of the conversation (for assisting user by fetching data, validating user inputs against database or completing a transaction on a service such as travel bookings, leave/regularisation systems, HR queries etc). Integration with live systems and applications is ongoing.

Tools used: Python3, NLP, NLG, RASA framework, entity extraction, Markov chains, LSTM based neural networks, Django, Docker, nginx, Keras, Scikit-learn.

Course Recommendation Engine for Reliance LMS

Oct 2017 – May 2018

Productionalised a course recommendation engine for 30,000+ employees which integrates various businesses at user end and various learning partners of Reliance at content end. Utilised employee demographics and organisational data to create multiple recommendation systems integrated via a multi-arm bandit-based architecture to personalise each user's experience. Have used matrix decompositions, fuzzy logic, collaborative filtering, association models, context clustering and reinforcement learning.

Tools used: Python3, Text analysis, NLP, Collaborative filtering, SVD, Search strategies, Multi-arm Bandits, Reinforcement Learning, Scikit-learn.

Employee Car-pooling service using Geo-Spatial clustering

Sep 2017 – Oct 2017

Designed and deployed an unsupervised model over employee address database across Mumbai to create geo-spatial clusters based on density of residence across the map. This was followed a route optimisation algorithm using network analysis of the graph of clusters and then a match making model for route matching which estimated polygon similarity between optimal (estimated) routes of the passengers and car owner. This model is currently being housed into a B2B employee services module called Share-a-Ride.

Tools used: Python3, Text analysis, Google API, Hierarchical DBScan, Network centralities and route optimisation, Polygonal similarity techniques.

Viewer interest prediction on Rental Listings on Renthop (Kaggle)

Mar 2017 – May 2017

Objective was to predict how popular an apartment rental listing is based on the listing content like text description, photos, number of bedrooms, price, etc. The data comes from renthop.com, an apartment listing website. I created an ensemble model using xgboost wrapped in a cross validator, stacked over KazAnova's StackNet with random forest and SVM. Features used included basic features, simple calculated features, constructed features over manager_id using tf-idf and clustered longitude-latitude positions, and finally magic feature. Model iterations were done with parameter tuning followed by averaging and geometric mean of predictions. The accuracy measure was log loss and my best model got me top 7% global ranking on Kaggle.

Tools used: Python3, NLTK, SVM, K-Means, Random Forest, XGBoost with Cross Validation, StackNet by KazAnova.

Recruitment decision making tool called Careerletics Enterprise

Jul 2016 – July 2017

Careerletics Enterprise is an intelligent platform for recruiters which assists them with pre-hire decision making and reduces the hiring lifecycle from a few weeks to a few minutes. It assists a recruiter by parsing resume data, quantifying candidate metrics, calculating relevance against a job description and ranking candidates by a metric called employability score. First, an exhaustive database linking industries & functions to skill sets, companies, job positions and colleges was created by using natural language processing over a database of half a million resumes documents (without any specific template). This database was then utilised to identify qualification, skills and experience information from user resumes via a parser. This was coupled with a chatbot to collect missing candidate information directly. Next, a stacked model for filtering, relevance matching, and competitive ranking was developed. Candidates which were finally selected by the recruiter are captured and used as a feedback the self-learning algorithm to adjust parameter weights. The platform and algorithm are patented under iPredictt Data Science Labs.

Tools used: Python3, Expectation maximisation, Gaussian mixture model, Gradient Boosting, PCA, hierarchal clustering.

Analysis of Political Affiliation and Sentiment over Social Media

Jan 2016 – May 2016

The objective was to understand the sentiments for a popular Indian News Network with respect to different political parties over Twitter and Facebook and compare the sentiments of other competitor news networks against it. Tweepy & web scraping was used to pull data via Twitter and Facebook, followed by data cleaning, feature generation, and NLP treatment to generate a sentiment report. The sectors of analysis included comparing political party affiliation, quantifying shared sentiment across newsgroups, detecting targeted negative propaganda over social media and forecasting topic-wise sentiment over Twitter.

Tools used: Python3, Tweepy, NLTK, Topic Modelling, Sentiment Analysis.

Optimise Ad Exchange networks for increasing campaign value

Jul 2015 – Dec 2015

The objective was to create a platform for a INR 60cr turnover Mobile Ad Exchange startup to optimise ad campaign time and direction which involves selecting the right publisher for the advertising campaign as a factor of time of the day, conversion rates, customer target category and network type. Variable importance calculated via Decision trees to categorise publisher efficiency and thus analyse trends better, while click probability for cookie ids was calculated by building a logistic model. The campaign statistics were visualised using charts and Sankey diagrams over an R-Shiny server.

Tools used: R, R-Shiny, Decision trees, Random Forest, Logistic regression.

Supply Chain network optimisation and planning

Nov 2014 – Mar 2015

Client was a fortune 50 multinational computer technology giant. The project objective was to analyse backlogs and develop a network flow optimisation model for Americas, EMEIA and Asia logistics team to enhance the efficiency of respective supply chains. A model was built on 3 years of backlog data with stage-wise & SKU-wise flow's starting from Manufacturing to Fulfilment Centres/Customers. Missing data were imputed using decision trees followed by Linear programming to minimise the objective function of the number of backlogs in each network. The resulting model was visualised using Tableau and shared with 1,000+ stakeholders and executives from Singapore, Austin, Hong Kong, London, Korea and India offices.

Tools used: SQL, R, Decision Trees, Ensemble models, Dynamic Programming, Tableau

Theoretical Win prediction for customers of a Casino Giant

Jun 2014 – Oct 2014

A Fortune 500 Casino Giant used certain business rules to calculate ADT (Accumulated daily theoretical win) for each of their customers to decide the category of their marketing spend which had an extremely low accuracy (32%). The objective was to build a regression model to predict ADT values for customers based on gambling spends, wins and trip information. An ensemble model was created based on analysis of variation in the test variable (ADT). A certain segment of the customer population (which was primarily low spend customers) was tackled using generalised linear models while remaining segment (which comprised primarily of high spend customers) was tackled using 11 separate Support Vector Machine classification models. The test variable for these was bucketed into spend categories instead of using a continuous ADT value. The accuracy of this model was much higher than the base model (53%). The exercise was followed by creating a financial modelling simulator using these predictions to generate best and worst-case profit/loss scenarios over variable marketing spends.

Tools used: Python2, ANOVA, K-means clustering, Support vector machines, Monte-Carlo simulation.

Driver analysis for market Cannibalisation

Dec 2013 – May 2014

The 2nd largest toy manufacturer brand showed quarter on quarter ROI decline of 20% which amplified further during the latest holiday season. Clear understanding was required on what were the prime causes of this decline. A five-dimension deterministic model was created to analyse parameters calculated through web analytics. This model was then passed through regression analysis for generating estimates for each parameter as a substitute for driver towards the sales decline. A major realisation by the end of the exercise was that the decline was primarily due to cannibalisation by a fresh brand they launched themselves but for a higher age category. This allowed them to take major decisions in time to stabilise the curve to around 8% decline in the coming quarter and affected the launch dates of their upcoming brands.

Tools used: R, Deterministic modelling, Web analytics, Generalised regression models

Customer Segmentation and Targeting for retail products

May 2013 – Nov 2013

Client was the world's biggest home improvement retail company. The objective was to create customer segments based on their behavioural traits, spend patterns and volatility in purchase categories which would allow the client to understand and target better. Customer segmentation based on transaction data was done using RFM segmentation followed by item-based and user-based collaborative filters to create purchase category recommendations for customised targeting. This directly affected client’s top line for specific departments such as gardening and home repair.

Tools used: SQL, Excel, R, RFM Segmentation, Collaborative Filtering.

Real Time in-store traffic analysis using Brickstream

Nov 2012 – Apr 2013

Client was the world's biggest home improvement retail company. They were on a pilot with Brickstream, which is a Video analytics software which uses aisle camera footage to virtually create trip lines and dwell zones. Exhaustive reports were created for the data collected by Bricksteam enabled cameras on a weekly level. Trip line analysis allowed client to predict traffic hours in real time and accordingly align their store associates for coming days/weeks, thereby improving resource management. Dwell analysis enabled the client to understand dwell times of customers at specific aisles positions thereby enabling them to take decisions on shelf space management.

Tools used: Brickstream, SQL, R, Video processing.