Akshay **Sehgal**

www.akshaysehgal.com
akshaysehgal2005@gmail.com
https://in.linkedin.com/in/akshay-sehgal-a5071458
+91-9916594778

I am a **Data Scientist**, currently working as a lead (General manager, SME-1) at Reliance Industries, where I design, train and deploy ML models powering enterprise scale platforms and products. I currently lead a team of about 10 data scientists working closely with full stack developers, product teams and dev-ops to map out and productionalizing AI/ML architectures for cloud based applications related to employee management systems and services. Some of my projects include, geo-spatial route matching, recommendation engines, distributed virtual assistants, document semantic matching, anomaly detection on image data and natural language to querying a database.

Previously I was heading the Strategy and Development of data powered products for a startup company called iPredictt Data Labs which I also co-founded. My career in the data science domain started off in Mu-Sigma, a pure play analytics firm. I have worked cross domain including HR, Advertising, Retail and Entertainment industries.

Core Skills

I have experience with **supervised learning methods** such as generalized linear models, decision trees, ensemble models (Stacknet, Xgboost, RF), support vector machines (SVM), probabilistic models, deep learning and **unsupervised models** such as K-means clustering, Gaussian mixture models, hierarchical DBScan for geo-spatial data, PCA, deep belief networks, RBMs and self-organizing maps. This includes applying NLP based models such as Seq2Seq models with attention, word embeddings (fasttext, word2vec, Glove), LSTM/GRUs, 1d convolutional networks. In my time outside office I enjoy exploring GANs, Deep dream networks, Reinforcement learning, Genetic algorithms, computer vision and network analysis.

I have frequently worked with Django and Flask frameworks deployed via Ngrok or Docker on Amazon EC2, Azure VMs, and server-less deployment using Amazon Lambda (Zappa).

Experience

Year	Company/Institution	Role
Current	Reliance Industries	Lead Data Scientist (GM)
2017	iPredictt Data Science Labs	Head, S&A ML products
2015	iPredictt Data Science Labs	Data Scientist Lead
2014	Mu-Sigma	Sr. Decision Scientist
2012	Mu-Sigma	Decision Scientist
2008	Pune University	B.E Comp Science

Achievements

I am a frequent contributor on Digital Vidhya, Code Gladiators, Kaggle[1]. I have strong experience in handling non-technical clients, top management and have participated as a speaker at multiple tech events across India[2] [3] [4]. I also have 3 technology patents under my name (201721005644, 201621034521, 201621034522)

Projects

Natural Language querying of databases

Dec 2018 – Ongoing

Building 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.

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

JD-CV matching algorithm for candidate shortlisting

August 2018 – Ongoing

Building 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.

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

Distributed Virtual Assistant Development Toolkit

June 2018 – Ongoing

Building a python based tool allowing non-technical users to design, train and deploy closed domain virtual assistants using a GUI. The bots are then 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/regularization systems, HR queries etc). Integration with live systems and applications is ongoing.

Skills used: Python, 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

Productionalized a course recommendation engine for 30,000+ employees which integrates various businesses at user end and various learning partners of Reliance at content end. Utilized employee demographics and organisational data to create a 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.

Skills used: Python, 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 a 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 optimization 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.

Skills used: Python, Text analysis, Google API, Hierarchical DBScan, Network centralities and route optimization, 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.

Skills used: Python, 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 utilized 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.

Skills used: Python, Expectation maximization, 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 sentiment of 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.

Skills used: Python, Tweepy, NLTK, Topic Modelling, Sentiment Analysis.

Optimize Ad Exchange networks for increasing campaign value

Jul 2015 – Dec 2015

The objective was to create a platform for a 60cr turnover Mobile Ad Exchange startup to optimize 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 categorize publisher efficiency and thus analyze trends better, while click probability for cookie ids was calculated by building a logistic model. The campaign statistics were visualized using charts and Sankey diagrams over an R-Shiny server.

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

Supply Chain network optimization and planning

Nov 2014 – Mar 2015

Client was a fortune 50 multinational computer technology giant. The project objective was to analyze backlogs and develop a network flow optimization 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 Fulfillment Centers/ Customers. Missing data were imputed using decision trees followed by Linear programming to minimize the objective function of the number of backlogs in each network. The resulting model was visualized using Tableau and shared with 1,000+ stakeholders and executives from Singapore, Austin, Hong Kong, London, Korea and India offices.

Skills used: SQL, R, Decision Trees, Linear 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 generalized 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.

Skills used: Python, 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 analyze 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 realization by the end of the exercise was that the decline was primarily due to cannibalization by a fresh brand they launched themselves but for a higher age category. This allowed them to take major decisions in time to stabilize the curve to around 8% decline in the coming quarter and affected the launch dates of their upcoming brands.

Skills used: R, Deterministic modeling, Web analytics, Generalized 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 behavioral 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 customized targeting. This directly affected client’s top line for specific departments such as gardening and home repair.

Skills 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.

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