Ankit Malik

Data Scientist

Email: ank.dsm@gmail.com
Location: New Delhi, India
Mob: +91-8368447396
LinkedIn: https://www.linkedin.com/in/ankitmalikds

A Data Scientist with 10 years of experience in Machine Learning, Computer Vision, and NLP with domain expertise in Marketing, Supply Chain, Retail, Advertising, and Automation.

A **US B1 visa** holder, and a Technical Reviewer with Packt Publishing. Expert in Python and SQL.

Professional projects (Jul'11 - present)

WorkFusion, India

Data Scientist (Aug'18 – present)

Leading data science team to drive innovation and development by conducting ML workshops and journal discussions.

1. **Automating Extraction for Claim Settlement**:
   Data: Scanned Images of claim forms (CMS 1500) – grayscale, coloured, noisy, scaled and tilted.
   Pre-processing: correction for scale and tilt (SIFT), background removal (pretrained autoencoder)
   Extraction: Field coordinate scaling, OCR and IOU based information retrieval
   Post-processing: Regex based on field type
   Result: Precision (97%) and Recall (93%) against the benchmark of 70% each
2. **In-house AutoML product:**
   Objective: Generic product for information extraction from documents
   Data: Invoices, Bill of Lading, Emails (service requests) etc.
   Process: Tokenization, Annotators, Feature extraction, model selection, and post-processing
   Hyper-parameter Optimization: Model selection (v1) and Feature selection/elimination (v2)
   Tech Stack: Python, Libsvm, Java
   Use case example (Email Service Request automation)
   Email ingestion > Classification > Extraction using above > SR generation
3. **Signature Detection:**
   Objective: Flag missed signatures
   Data: Scanned contracts
   Meta data extraction: Computing word priority list per template per signature area
   Dynamic Snipping: Snipping signature areas using meta data and OCR results
   Handwriting recognition: CNN based classifier (handwriting vs machine text)
   Results: 97% (F1 score)
4. **Model Risk Management for Payments Sanctions screening:**
   Objective: Stress testing, explainability, sensitivity analysis and reporting.
   Techniques: SHAP, LIME, PDP for explainability, hyperparameter sensitivity, and threshold sensitivity

Elsevier (Parity Computing), India

Data Scientist (Nov'16 – Aug'18)

1. **Institution profiling for Elsevier** is being utilised as one of the key products to compile Global University Rankings based on research orientation and contribution to society. The data is utilised by Times Higher Education (THE) and US NEWS among others.
   The idea was to build and deploy a solution for attributing research to institutions.
   High level steps involved in the solution:
   - **Data:** more than 10 million research papers, journals and articles
   - **ETL:** From articles to attribution string (xml)
   - **NER:** Linear chain sequential model (ex: CRF)
     - Tokenization
     - Feature engineering
     - Cross Validation and tuning for variable attention
   - **Matcher:** Matching to orgDB
     - Primary match
     - City, state, country and street address score for variant match
     - Token level fuzzy matching
     - N-gram matching + classifier on match features

- **Clustering:** Cluster of unmatched profiles
  - cosine distance similarity
  - in-house custom, efficient clustering algorithm
- **Results:** 94%+ Precision and 93%+ automation rate (production)

The process was completed on opensource data, and the model was then tuned for customer data. The solution was built in python and ported to java.

VYMO, India

Sr. Data Scientist (May'16 to Nov'16)

AI enabled Sales Assistant: One of its kind in Sales assistant space that involved designing a lead scoring system to help sales force in targeting the leads with a higher likelihood of conversion. AI enabled assistant (with a F-beta of ~82%) led to 32% increase in conversion rate within first month of deployment. Tech stack involved MongoDB, ElasticSearch and Kafka.

MIQ (MediaIQ), India

Data Scientist (Jul'15 - May'16)

1. **Marketing Channel Effectiveness:**
   USDA required insights into their offline (TV, radio, billboards, subway, bus postings, newspaper) vs online media.
   Data Prep: Sales/website traffic vs channel expenditure (joined over date)
   Pre-processing: adding ad-stock effect + normalization
   Modelling: Polynomial regression
   Results: p-value lower than significance value, coefficients as channel attributions
2. **Click Prediction Analytics:**
   Data: Impression and Click labels (highly imbalanced), timestamps, and url/ad features
   Tech: RUSBOOST (Random Under-sampling with Boosting)
   Result: Factors impacting click decisions

Harman Connected Services (Symphony Teleca), India

Data Scientist, (Feb'15 – Jul'15)

iCREATE AWARD winner for multiple successful POCs

Oracle, India

Analyst (Jul'13 - Feb'15)

Markdown Optimization: Built a recommendation engine using seasonality and price elasticity to markdown depth and time for a U.K. based retailer client. Achieved ~14% cost savings following the model implementation.

Blue Yonder (JDA Software), India

Business Consultant (Jul'11 - Jun'13)

Transportation Solution for Mexican Beverage Company: Worked on the technology consulting using i2 TMS as base product.

Education

Degree	Year	Institute	Major and Specialization	CGPA / %
MBA	2009-11	Delhi School of Management, DTU	Supply Chain & Marketing	8.2 CGPA
B. Tech	2005-09	MAIT, GGSIPU, Delhi	Electrical & Electronics	80 %
XII	2004-05	CBSE	Physics, Chemistry, Math, English	81 %
X	2002-03	CBSE	English, Science, Maths	85 %

Certifications

- Deep Learning Specialization – DeepLearning.AI
- Mathematics for Machine Learning Specialization - Imperial College London
- NSE Certified Market Professional - Level 2 - National Stock Exchange of India Limited
- Financial Markets (with Honours) – Yale University
- R Programming - The Johns Hopkins University
- Python Programming - University of Michigan
- APICS CPIM Level 1