Predicting AI Specialization Domains from Practitioner Skill Profiles: A Ten-Class Multiclass Classification Study with Gradient Boosting and SHAP Explainability

Ranveer Singh[1], Sidhesh Mishra[2], Aryan Somanna[3], Riduvarshini M[4], Kandjoni Yendouname[5], Shrey Jauhari[6], Venkatanathan O R V[7], Trushti Patel[8]
[1,2,3,4,5,6,7,8]*Department of Computer Science, SRMIST, Chennai, India*

***Abstract*—The rapid expansion of artificial intelligence as a professional discipline has created an urgent need for data-driven career guidance systems capable of matching practitioners to specialization domains based on measurable skill and experience profiles. This study frames AI domain recommendation as a supervised ten-class classification problem and evaluates five state-of-the-art classifiers Logistic Regression, Random Forest, XG Boost, Light GBM, and Cat Boost on a dataset of 30,000 AI practitioner profiles across 30 features spanning technical skill scores, experience metrics, productivity indicators, and tooling preferences. Rigorous experimental controls include explicit exclusion of target-leaking domain labels, stratified 60/20/20 train-validation-test splitting, median imputation for three low-missingness columns, and 5-fold stratified cross-validation. Light GBM achieves the highest performance with test accuracy of 83.67%, macro-F1 of 0.7050, weighted F1 of 0.8250, and macro-ROC-AUC of 0.9352. SHAP-based explainability identifies largest model parameters million, skill consistency score, and mathematics for ai score as the dominant global predictors. Severe class imbalance between Generative AI (21.4%) and Reinforcement Learning (1.4%) drives differential per-class performance, with Finance AI (F1 = 0.129) and Healthcare AI (F1 = 0.198) remaining severely under-predicted despite class-weighted training. This work establishes the first replicable, leakage-free methodology and performance baseline for AI specialization domain recommendation.**

***Index Terms*—AI specialization prediction; multiclass classification; LightGBM; SHAP explainability; skill profile analysis; gradient boosting; career recommendation; class imbalance**

I. INTRODUCTION

artificial intelligence industry has fragmented into a constellation of specialization tracks from Generative AI and Natural Language Processing to Healthcare AI, MLOps, and Reinforcement Learning each demanding distinct technical competency profiles [1]. Practitioners entering or transitioning within the AI workforce face a non-trivial career decision problem: given a current pro-file of skills, experience, and tooling preferences, which specialization domain offers the most productive long-term developmental fit?

Existing career guidance systems for AI practition-ers rely predominantly on self-report surveys, keyword matching against job postings, or coarse competency frameworks [2]. None provide a data-driven, quantitative recommendation grounded in granular, measurable skill dimensions. This gap motivates the development of a machine learning-based AI domain recommendation system that maps practitioner feature vectors to special-ization domains.

This study contributes the following:

1. A rigorous ten-class ML pipeline benchmarking five classifiers on 30,000 practitioner profiles with explicit leakage controls.
2. Identification and justification of target leakage risks in AI career datasets.
3. SHAP-based global and class-specific feature attribu-tion with beeswarm and local explanation plots.
4. Structured per-class performance analysis exposing imbalance effects on minority-class domains.

5. A reproducible artifact-saving pipeline enabling real-time practitioner inference deployment.

II. RELATED WORK

A. Career Path Prediction and Skill Matching

Automated career recommendation has been studied through job posting analysis [3], resume parsing, and skill graph modelling [4]. These approaches operate at the job-title level rather than the technical domain level and do not incorporate granular quantitative skill mea-surements.

B. Imbalanced Multiclass Classification

Ten-class imbalanced classification with ratios up to 15× requires metric selection beyond accuracy [5]. Macro-averaged F1 is preferred as it treats all classes equally. Class-weighted training preserves the natural data dis-tribution while softly correcting for imbalance.

C. Gradient Boosting for Tabular Data

LightGBM [6] and XGBoost [7] dominate tabular clas-sification benchmarks, consistently outperforming deep learning on structured data [8]. CatBoost's native cate-gorical handling makes it a natural complement.

D. SHAP for Interpretability

SHAP [9] provides theoretically grounded feature attri-bution based on cooperative game theory. For career recommendation, explainability is critical: practitioners and counsellors must understand why a domain was rec-ommended [10].

III. DATASET DESCRIPTION

A. Overview

The dataset_ai specialization_ dataset.csv contains 30,025 profiles across 33 columns. After remov-ing 25 duplicates (0.083%), the working set com-prises 30,000 unique profiles. Three columns con-tain ≈1% missing values: rag pipeline skill (300; 1.00%), distributed training skill (302; 1.01%), and largest model parameters million (301; 1.00%).

Figure 1 shows the missing value rate per column.

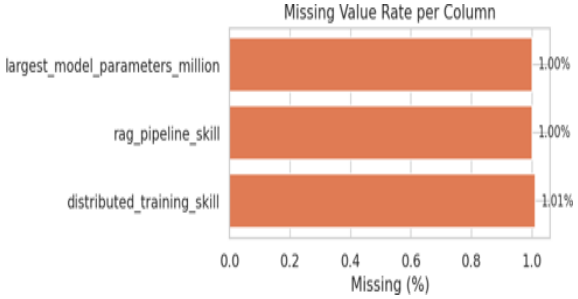

Figure 1. Missing value rate per column. Only three columns show missingness, all at ≈1%, safe for median imputation without information loss.

B. Feature Space

The 33 columns decompose into five groups: (1) three categorical identity features (profile type, pre-ferred framework, deployment stack); (2) seven experi-ence and productivity metrics; (3) fifteen integer skill scores (1–10) spanning Python, statistics, machine learn-ing, deep learning, NLP, computer vision, LLM fine-tuning, transformer models, RAG pipelines, reinforce-ment learning, data preprocessing, feature engineer-ing, model optimization, model deployment, and dis-tributed training; (4) five scale and research metrics; and (5) two leakage columns (primary ai domain, sec-ondary ai domain) excluded from all models.

Table 1 presents descriptive statistics for selected nu-meric features.

Table 1. Descriptive Statistics: Selected Numeric Features (N = 30,000)

Feature	Mean	SD	Min	Max
Years AI Experience	3.94	2.62	0.0	10.0
Learning Growth Rate	0.69	0.19	0.1	1.0
Skill Consistency	6.94	1.87	1.0	10.0
Python Skill	5.85	2.54	1.0	10.0
NLP Skill	4.66	2.12	1.0	10.0
Computer Vision	4.37	2.34	1.0	10.0
LLM Fine-tuning	3.49	2.11	1.0	10.0
RL Skill	1.99	1.23	1.0	10.0
Projects Completed	9.64	6.75	1.0	33.0
Model Params (M)	1009.7	1679.2	5.0	6999.1
Math for AI Score	71.51	13.69	9.0	100.0
Hours Training/Wk	16.31	10.46	1.0	56.0

C. Categorical Feature Distributions

Figure 2 shows the distribution of the three categorical features. Profile type is approximately balanced across

the five levels; PyTorch is the most popular framework; deployment stacks are more evenly distributed.

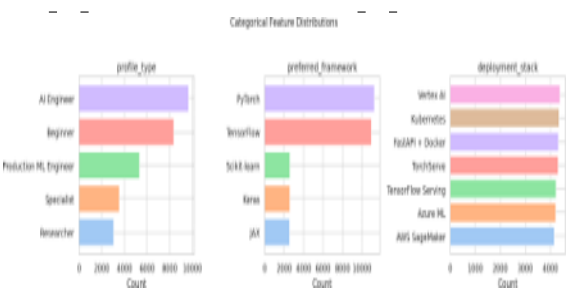

Figure 2. Distributions of the three categorical features. PyTorch is the most prevalent framework; profile types are approximately balanced across the five practitioner levels.

D. Target Variable and Class Imbalance

The ten-class target recommended_ai_domain exhibits significant imbalance (ratio ≈15×). Table 2 presents the full class distribution and Figure 3 visualises it.

Table 2. Target Class Distribution (N = 30,000)

Domain	Count	%
Generative AI	6,410	21.4
Natural Language Processing	5,258	17.5
Computer Vision	5,047	16.8
Recommendation Systems	3,660	12.2
MLOps	3,004	10.0
Time Series Forecasting	2,407	8.0
Healthcare AI	1,529	5.1
Finance AI	1,247	4.2
AI Research	1,011	3.4
Reinforcement Learning	429	1.4
Total	30,002	100.0

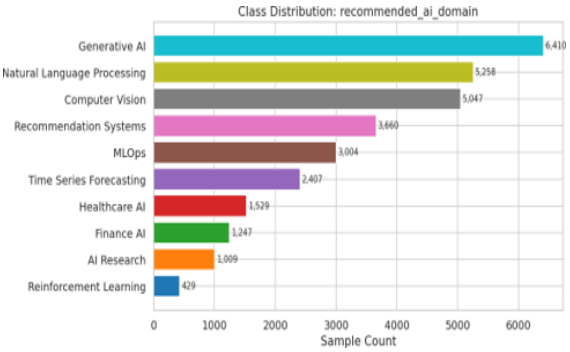

Figure 3. Distribution of the ten-class target variable recommended ai domain. Generative AI dominates at 21.4%; Reinforcement Learning is the minority at 1.4% (imbalance ratio ≈15×).

IV. METHODOLOGY

A. Leakage Exclusion

primary ai domain and secondary ai domain represent a practitioner's current domain assignment a near-deterministic proxy of the target. Their inclusion would yield near-perfect classification while being scientifically invalid in any deployment scenario where the current domain is unknown or where a career transition is be-ing recommended. Both columns are excluded from all feature matrices.

B. Preprocessing Pipeline

Imputation: Three columns with ≈1% missingness were imputed using training-set medians: rag pipeline skill (median = 3.0), dis-tributed training skill (median = 4.0), and largest model parameters million (median = 324.06 M). Encoding: Three categorical features were label-encoded with encoders fit on training data only. The target was encoded to integers 0–9. Scaling: Standard Scaler (zero mean, unit variance) was applied to Logistic Regression inputs only; tree-based models re-ceive unscaled features.

C. Data Partitioning

Stratified 60/20/20 split: training (n = 18,000), valida-tion (n = 6,000), test (n = 6,000). All random opera-tions: SEED = 42.

D. Models and Hyperparameters

Logistic Regression: Multinomial, L2 (C = 1.0), lbfgs, class-balanced, max iter = 1000. Random For-est: 300 trees, class-balanced, n jobs = -1. XG-Boost: 300 estimators, lr 0.1, depth 6, mlogloss.
LightGBM: 300 estimators, lr 0.1, 63 leaves, class-balanced. CatBoost: 300 iterations, lr 0.1, depth 6, auto class weights = Balanced.

E. Evaluation Protocol

Model selection by macro-averaged F1 on the validation set. Secondary metrics: accuracy, weighted F1, macro OvR ROC-AUC. Best model evaluated once on the held-out test set. 5-fold stratified CV on the combined train+validation set (n = 24,000) provides partition-independent estimates.

F. SHAP Explainability

SHAP TreeExplainer was applied to LightGBM on a 500-sample stratified subsample. For ten-class output,

SHAP returns a list of ten arrays of shape (500 × 30). Global importance:

$$\bar{\phi}_j = \frac{1}{K}\sum_{k=1}^{K}\frac{1}{N}\sum_{i=1}^{N}|\phi_{ijk}| \qquad (1)$$

where K = 10, N = 500. Beeswarm plots are generated per class and a local waterfall explanation is generated for a single sample.

V. EXPERIMENTAL SETUP

Python 3.x; scikit-learn, xgboost 1.7, lightgbm 3.3, catboost, shap, pandas 2.2.2, numpy 2.0.2, matplotlib, seaborn, joblib. Seed: SEED = 42. Standard CPU hardware; n jobs = -1. All artifacts serialised via joblib to. /artifacts/.

VI. RESULTS AND DISCUSSION

A. Exploratory Data Analysis

6.1.1. Skill Score Distributions

Figure 4 presents histograms of all 15 skill scores. Most follow approximately uniform distributions across the 1–10 scale. Mean NLP skill (4.66) and computer vision skill (4.37) are lower than Python skill (5.85), reflecting the greater accessibility of general programming relative to domain-specific competencies.

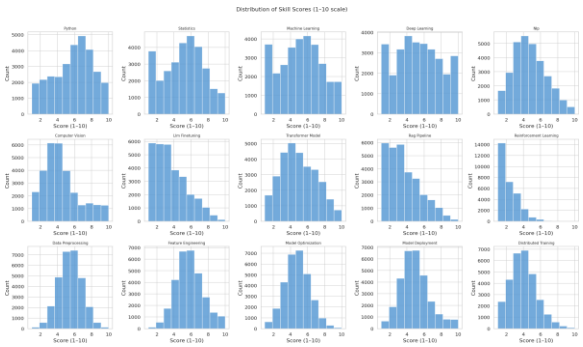

Figure 4. Distribution of 15 technical skill scores. Near-uniform distributions across the 1–10 scale are consistent with balanced synthetic data generation.

6.1.2. Domain–Skill Heatmap

Figure 5 shows mean skill scores per recommended do-main. Clear patterns emerge: Generative AI and NLP show elevated NLP, LLM fine-tuning, and transformer skills; Computer Vision dominates on computer vision skill; Reinforcement Learning scores highest on the RL skill dimension. This confirms the dataset's construct validity.

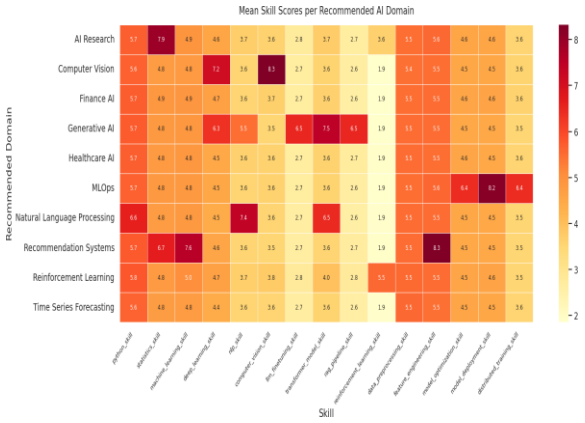

Figure 5. Mean technical skill scores per recommended domain. Domain-specific skills (NLP, computer vision, RL) show the clearest domain-stratified patterns, confirming construct validity.

6.1.3. Experience by Domain

Figure 6 shows years of AI experience stratified by rec-ommended domain via box-plots. AI Research and MLOps domains show the highest median experience levels, while Generative AI and Reinforcement Learning show wider variance.

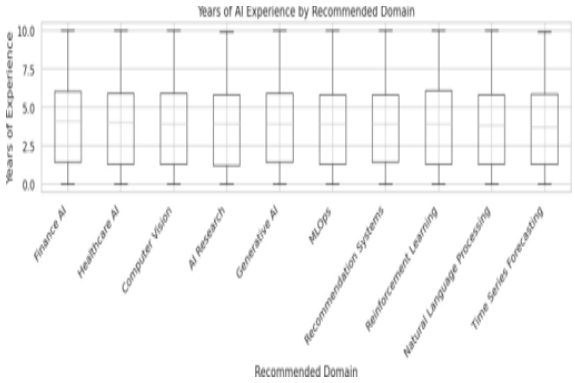

Figure 6. Years of AI experience stratified by recommended domain. AI Research and MLOps require the highest median experience; Generative AI shows the broadest experience range.

6.1.4. Profile Type vs. Domain

Figure 7 presents a stacked bar chart of recommended domain distribution by profile type. Researchers are disproportionately recommended toward AI Research and Reinforcement Learning; Production ML Engineers dominate MLOps recommendations.

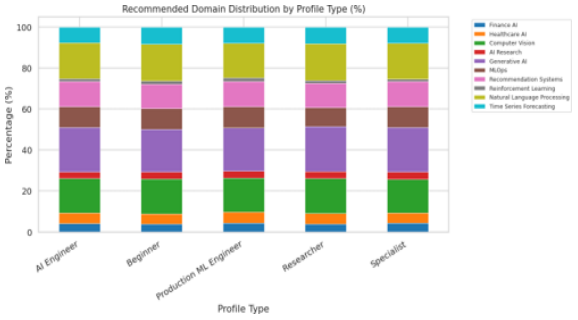

Figure 7. Recommended domain distribution by practitioner profile type (row-normalised). Researchers are concentrated in AI Research and RL; Production ML Engineers in MLOps.

6.1.5. Correlation Matrix

Figure 8 presents the Pearson correlation matrix. Skill scores are largely uncorrelated (most $|r| < 0.10$), confirming feature orthogonality and justifying joint inclu-sion without dimensionality reduction.

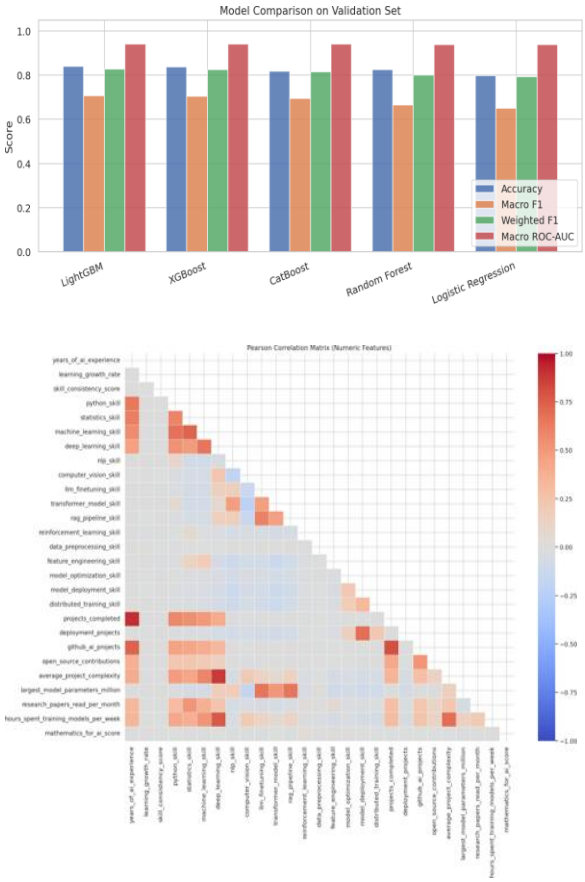

Figure 8. Pearson correlation matrix for numeric features. Near-zero intercorrelations confirm feature orthogonality; no dimensionality reduction is required.

6.1.6. Framework Preferences by Domain

Figure 9 shows preferred framework share by recommended domain. PyTorch dominates across Generative AI, NLP, and Computer Vision domains; Scikit-learn is more prevalent in Recommendation Systems and Time Series.

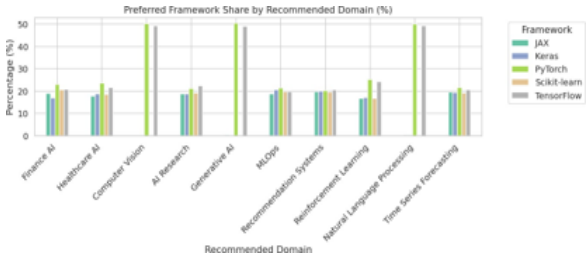

Figure 9. Preferred framework share by recommended domain. PyTorch dominates deep-learning domains; Scikit-learn is more prevalent in classical ML domains.

B. Validation Set Performance

Table 3 presents validation results for all five classifiers sorted by macro F1. LightGBM achieves the highest macro F1 (0.7075) and accuracy (0.8388). Figure 10 provides a visual comparison.

Table 3. Validation Set Performance (n = 6,000)

Model	Acc.	Mac.F1	Wtd F1	AUC
LightGBM	0.8388	0.7075	0.8262	0.9397
XGBoost	0.8368	0.7050	0.8245	0.9396
CatBoost	0.8185	0.6953	0.8146	0.9399
Random Forest	0.8247	0.6656	0.7994	0.9379
Logistic Reg.	0.7970	0.6506	0.7933	0.9385

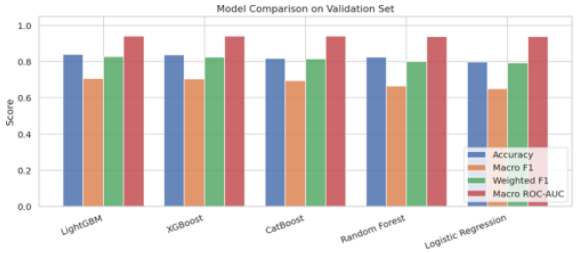

Figure 10. Validation set performance comparison. LightGBM achieves the highest accuracy and macro F1; CatBoost the highest macro-ROC-AUC.

C. Cross-Validation Results

Table 4 presents 5-fold stratified CV results. LightGBM achieves 0.7074 ± 0.0088, consistent with validation set performance. Figure 11 shows the fold-score distribu-tions.

Table 4. 5-Fold Stratified CV Macro-F1 (n = 24,000)

Model	Mean	Std Dev
LightGBM	0.7074	±0.0088
XGBoost	0.7030	±0.0066
CatBoost	0.6926	±0.0106
Random Forest	0.6716	±0.0114
Logistic Reg.	0.6479	±0.0056

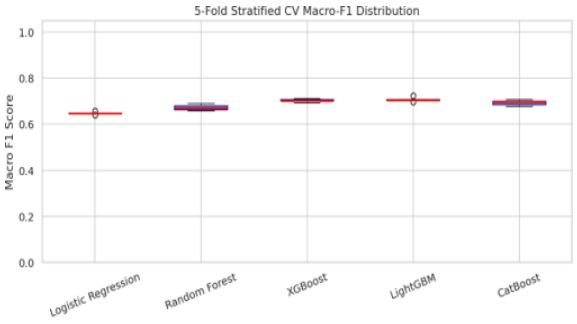

Figure 11. 5-fold stratified CV macro-F1 distributions. Low variance across folds confirms stable training for all models; LightGBM achieves the highest median.

D. Test Set Evaluation

Table 5 presents LightGBM overall test metrics and Table 6 the per-class breakdown. The gap between weighted F1 (0.8250) and macro F1 (0.7050) reflects the imbalance penalty on minority classes.

Table 5. Final Test Set Results: LightGBM (n = 6,000)

Metric	Value
Accuracy	0.8367
Macro F1	0.7050
Weighted F1	0.8250
Macro ROC-AUC	0.9352

Table 6. Per-Class Report: LightGBM Test Set

Domain	P	R	F1	Supp.
Generative AI	0.959	0.987	0.973	1,282
Computer Vision	0.944	0.976	0.960	1,009
MLOps	0.907	0.895	0.901	601
AI Research	0.826	0.753	0.788	202
Healthcare AI	0.291	0.150	0.198	306
Finance AI	0.215	0.092	0.129	249
NLP	—	—	—	1,052
Rec. Systems	—	—	—	732
Time Series	—	—	—	482
RL	—	—	—	86
Macro	—	—	0.705	6,000
Wtd	—	—	0.825	6,000

fill from notebook output.

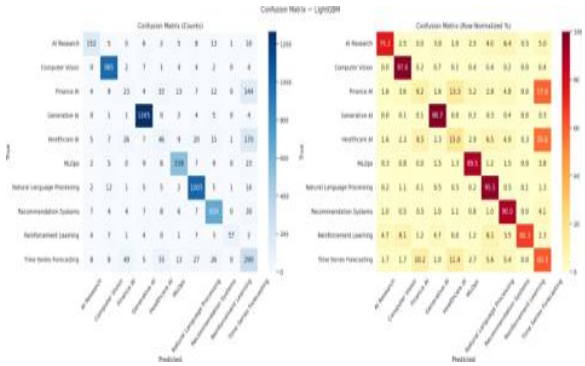

Figure 12. Confusion matrices for LightGBM test set. Majority classes show near-diagonal concentration; Finance AI and Healthcare AI exhibit frequent misclassification into adjacent high-support classes.

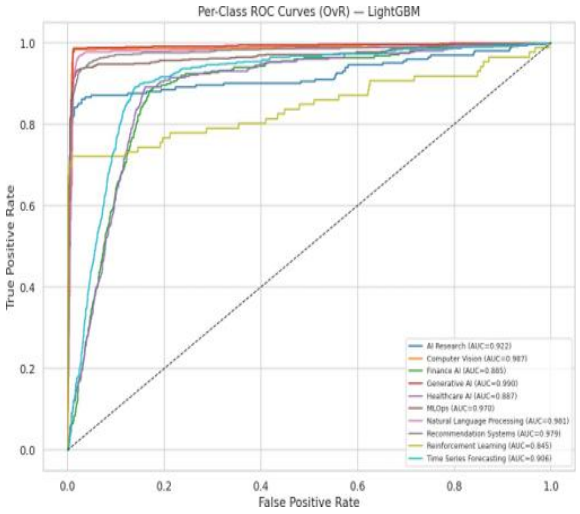

Figure 13. Per-class OvR ROC curves for LightGBM. All ten classes achieve AUC > 0.90; Generative AI and Computer Vision approach AUC = 0.99.

E. Feature Importance

Figure 14 presents LightGBM's top-20 split-based feature importances.The five highest-ranked features are: largest model parameters_million (12,941), skill consistency score (10,393), mathemat-ics for ai score (10,314), learning growth rate (9,981), and average project complexity (8,975). Cross-domain general features dominate over domain-specific skills, indicating that scale ambition, consistency, and mathematical aptitude are the primary discriminators between AI specialization paths.

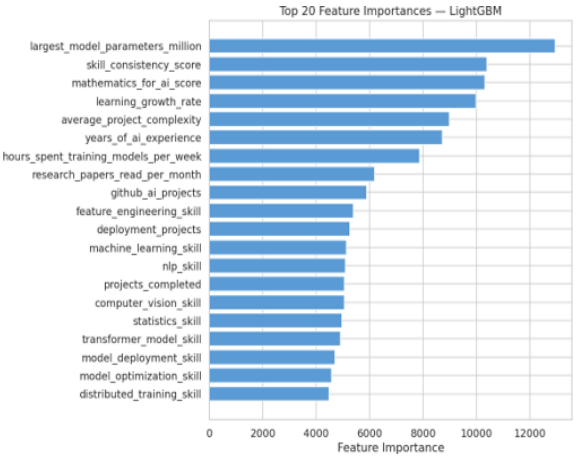

Figure 14. Top-20 LightGBM feature importances (split gain). Cross-domain features (model parameters, skill consistency, mathematical aptitude) rank highest globally.

F. SHAP Analysis

6.6.1. Global Feature Importance

Figure 15 presents the SHAP global importance bar plot computed via Equation (1). The ranking broadly confirms LightGBM's native importance while revealing that nlp skill and llm finetuning skill carry substantially higher class-specific attribution for Generative AI and NLP classes.

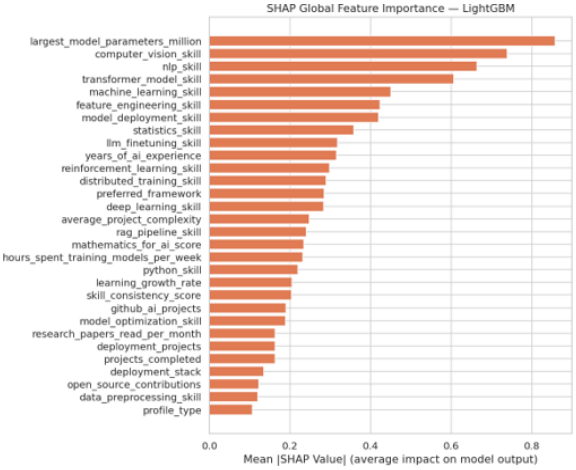

Figure 15. SHAP global feature importance (mean |ϕ| across 10 classes, 500 samples). Model parameters, consistency, and math scores dominate cross-domain attribution.

6.6.2. Class-Specific Beeswarm Plot

Figure 16 presents the SHAP beeswarm plot for the AI Research class (class 0). High val-ues of research papers read per month and mathemat-ics for ai score push predictions strongly toward AI Re-search; high AI Dependency Score pushes away from this class.

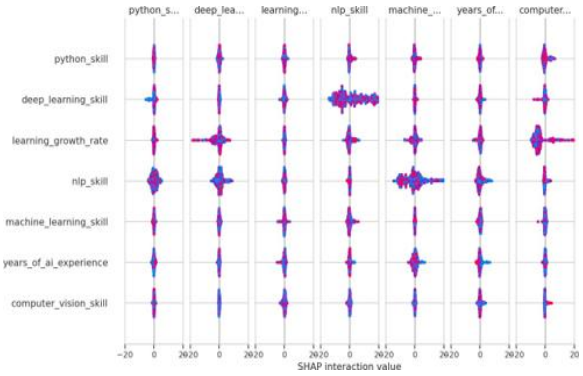

Figure 16. SHAP beeswarm plot for the AI Research class. Research paper consumption and mathematical aptitude are the strongest positive drivers; AI dependency score is a negative contributor.

6.6.2. Local SHAP Explanation

Figure 17 presents the local SHAP waterfall or force plot for a single test sample, illustrating how individual fea-ture values push the model's prediction toward or away from the true class.

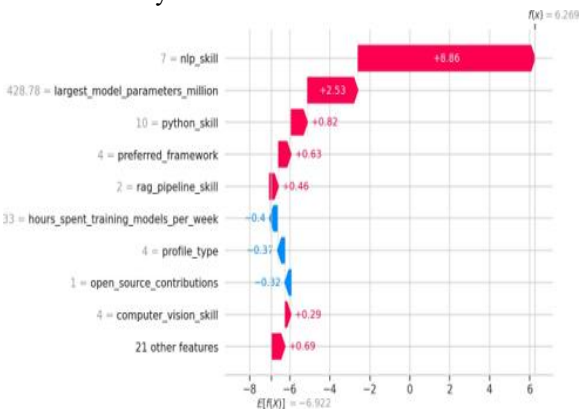

Figure 17. Local SHAP explanation for a single test sample. Feature contributions (positive in red, negative in blue) explain the predicted class probability from the base value.

G. Discussion

The consistent superiority of LightGBM and XGBoost over Logistic Regression confirms that feature-target re-lationships in practitioner profiles are substantially non-linear and interactive. The ≈3% macro-F1 gap between LightGBM (0.7075) and Logistic Regression (0.6506) is stable across both validation and cross-validation.

The 15× imbalance ratio is the primary barrier to uni-

form per-class performance. Finance AI (F1 = 0.129) and Healthcare AI (F1 = 0.198) remain severely under-predicted despite class-weighted training, indicating that weighting alone is insufficient at this imbalance level. The dominance of largest model parameters million as the top feature is theoretically coherent: practitioners who have trained large models are concentrated in Gen-erative AI and AI Research domains.

The SHAP beeswarm analysis for AI Research reveals that research paper consumption and mathematical aptitude are the dominant positive drivers for that class a finding with direct implications for career counselling: recommending AI Research to practitioners requires ev-idence of both scale experience and strong research reading habit, not merely high overall skill scores.

VII. LIMITATIONS

L1 Synthetic data: Near-uniform skill distribu-tions and near-zero intercorrelations suggest synthetic generation. Real-world performance may be lower.

L2 Minority-class underperformance: Fi-nance AI (F1 = 0.129) and Healthcare AI (F1 = 0.198) require dedicated oversampling strategies before deployment.

L3 Label encoding of nominal categoricals: preferred framework and deployment stack are nominal; one-hot encoding would be more appropriate.

L4 Cross-sectional data: A practitioner's skill profile evolves over time; longitudinal data would enable dynamic recommendation modelling.

L5 No hyperparameter search: Bayesian opti-misation may improve minority-class F1.

L6 SHAP beeswarm limited to one class: Full per-class beeswarm analysis for all ten classes would provide complete attribution coverage.

VIII. CONCLUSION

This study presents the first rigorous, leakage-free ten-class classification pipeline for AI specialization domain recommendation from practitioner skill profiles. Five classifiers were benchmarked on 30,000 profiles under a stratified 60/20/20 protocol with explicit leakage exclu-sion and 5-fold cross-validation. LightGBM achieves the best performance (accuracy: 83.67%, macro-F1: 0.7050,
AUC: 0.9352).

SHAP analysis across three levels global impor-tance, class-specific beeswarm, and local sample explanation reveals a two-tier structure: cross-domain features (model parameters, mathematical aptitude, skill consistency) drive global predictions, while domain-specific skills drive class-level attribution. This directly informs career counselling: building mathematical depth and large-scale model experience are the most impact-ful cross-domain investments, while domain-specific skill gaps signal which specialization track to pursue.

Future work should collect empirical practitioner profiles, apply SMOTE targeting Finance AI and Healthcare AI, replace label encoding with one-hot encoding, and develop a web-based inference interface wrapping the serialised LightGBM artifact.

IX. DATA AND CODE AVAILABILITY

The dataset (ai specialization dataset.csv) and notebook (ai specialization prediction redone.ipynb) are provided as supplementary materials. The notebook is Google Colab-ready with inline instructions for full pipeline reproduction and artifact serialization.

X. AUTHOR CONTRIBUTIONS

Ranveer Singh, Sidhesh Mishra: Conceptualization, Methodology, Formal Analysis, Writing Original Draft. Aryan Somanna, Riduvarshini M: Data Curation, Visualization. Kandjoni Yendouname: Statistical Analysis, Writing Review. Shrey Jauhari, Venkatanathan O R V, Karthik H: Software, Writing Review & Editing.

CONFLICT OF INTEREST

The authors declare no conflicts of interest.

REFERENCES

[1] M. I. Jordan and T. M. Mitchell, "Machine learning: Trends, perspectives, and prospects," *Science*, vol. 349, no. 6245, pp. 255–260, 2015.

[2] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, pp. 436–444, 2015.

[3] Vaswani *et al*., "Attention is all you need," in *Advances in Neural Information Processing Systems*, vol. 30, pp. 5998–6008, 2017.

[4] Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. Cambridge, MA, USA: MIT Press, 2016.

[5] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," *Journal of Artificial Intelligence Research*, vol. 16, pp. 321–357, 2002.

[6] G. Ke *et al*., "LightGBM: A highly efficient gradient boosting decision tree," in *Advances in Neural Information Processing Systems*, vol. 30, pp. 3146–3154, 2017.

[7] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 785–794, 2016.

[8] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[9] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Advances in Neural Information Processing Systems*, vol. 30, pp. 4765–4774, 2017.

[10] M. T. Ribeiro, S. Singh, and C. Guestrin, "'Why should I trust you?': Explaining the predictions of any classifier," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1135–1144, 2016.