Data-Driven Market Entry Strategy for EdTech in Tier-2/3 Cities

Using a 5,000-student JEE Aspirant Dataset

Ayan Bhardwaj

June 2024

Abstract

We evaluate the market opportunity for EdTech expansion into semi-urban India (Tier-2/3) using a dataset of 5,000 JEE aspirants. We conduct exploratory data analysis (EDA), build dropout-risk models (Logistic Regression, Random Forest), and segment learners via KMeans. Insights are translated into a go-to-market (GTM) play optimized for affordability, access, and retention. We also outline a risk-aware recommendation system that personalizes content, nudges, and pricing. This report is written for a Business Analyst / Product Management audience and mirrors a structured academic-style format.

Contents

List of Figures

List of Tables

1 Introduction

India's EdTech sector is large and dynamic, but penetration is uneven across regions. After aggressive growth in metro markets, national players face saturation and higher acquisition costs. Tier-2/3 markets (semi-urban & rural) represent a substantial base, but success requires affordability, vernacular depth, and hybrid (offline+online) delivery.

Goals.

- Quantify dropout risk drivers among JEE aspirants.
- Compare Tier-1 vs Tier-2/3 patterns to surface access and support gaps.
- Build actionable segments to inform product & pricing design.
- Propose a risk-aware recommendation system and a differentiated GTM vs. incumbents (e.g., BYJU's, PhysicsWallah).

Contributions.

1. Tier-wise EDA and heatmaps (peer pressure, mental-health).
2. Predictive modeling with *Random Forest* (AUC $\approx$ 0.90) and *Logistic Regression*.
3. KMeans segmentation ($k = 3$) highlighting affordability and support needs.
4. TAM–SAM–SOM funnel and retention-led GTM for Tier-2/3.
5. A practical, risk-aware **recommendation system** blueprint tied to operations.

2 Related Context & Motivation

Large national brands have optimized for metro learners, premium SKUs, and long-form courses. Tier-2/3 learners show different constraints: lower ARPU ceilings, inconsistent bandwidth, stronger role of schools/mentors, and higher sensitivity to peer pressure and exam stress. A data-driven, retention-first design can unlock sustainable growth outside metros.

3 Dataset

Sample. $N = 5000$ JEE aspirants (post-Class 12). Key fields used:

- location type: Urban / Semi-Urban / Rural (mapped to Tier-1 / Tier-2 / Tier-3)
- dropout: 0/1 (target)
- family income: Low / Mid / High
- parent education: categorical (mapped to years)
- peer_pressure level: Low / Medium / High (mapped to 1–3)
- mental_health issues: Yes/No (mapped to 0/1)
- coaching institute: None / Some (mapped to 0/1)

- daily study_hours: numeric
- JEE performance fields (optional usage)

Tier proxy. Urban → Tier-1, Semi-Urban → Tier-2, Rural → Tier-3.

Table 1: Feature dictionary (selected)

Feature	Description / Encoding
tier_std	Derived from location_type: {Tier-1, Tier-2, Tier-3}
is_dropout	Target (0/1) from dropout
income_level_num	Low/Mid/High → {1,2,3}
parent_edu_years	Map education level to years (10/12/15/17)
peer_pressure_num	Low/Med/High → {1,2,3}
mental_issues_01	Yes/No → {1,0}
has_coaching	None/Some → {0,1}
study_hours	Numeric (hours/day)

4 Methodology

Preprocessing. Lower-casing, punctuation normalization, ordinal encodes, boolean flags, and numeric coercion for hours. Tiers are derived from location type.

EDA. Tier-wise rates and group summaries; two heatmaps: (i) mental-health vs tier; (ii) peer pressure vs tier.

Modeling. Train/test = 75/25 with stratification. Pipelines with StandardScaler for numeric features and OneHotEncoder for categoricals.

- **Logistic Regression** (interpretable baseline).
- **Random Forest** (400 trees, non-linear performance).

Segmentation. KMeans with $k = 3$ on affordability & context features (income level, parent education, mental-health, peer pressure, coaching, study hours). Profiles are interpreted qualitatively for product fit.

Evaluation. Accuracy and ROC-AUC; qualitative feature-importance narrative for RF; operational interpretability from LR signs.

5 Exploratory Data Analysis (EDA)

5.1 Dropout Rates by Tier

Table 2: Dropout rate by city tier

Tier	Dropout Rate
Tier 1	22.35%
Tier 2	20.49%
Tier 3	19.34%

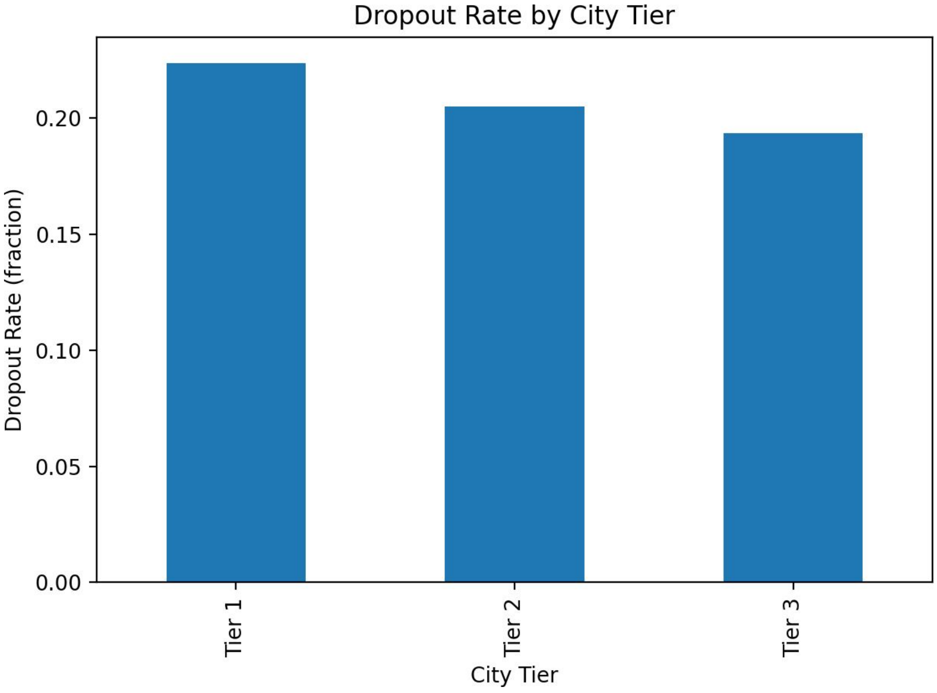

Figure 1: Bar chart: dropout rate (fraction) by city tier.

Observation. Baseline risk is broadly similar across tiers (slightly higher in Tier-1). Opportunity in Tier-2/3 is therefore *not* about higher inherent risk but about addressing access, affordability, and support gaps.

5.2 Access & Context by Tier

Table 3: Coaching enrollment (%)

Tier 1	77.2%
Tier 2	75.3%
Tier 3	74.9%

Table 4: Peer pressure (avg, 1–3)

Tier 1	2.00
Tier 2	2.01
Tier 3	2.00

Table 5: Mental-health issues (% “Yes”)

Tier 1	48.9%
Tier 2	47.1%
Tier 3	49.6%

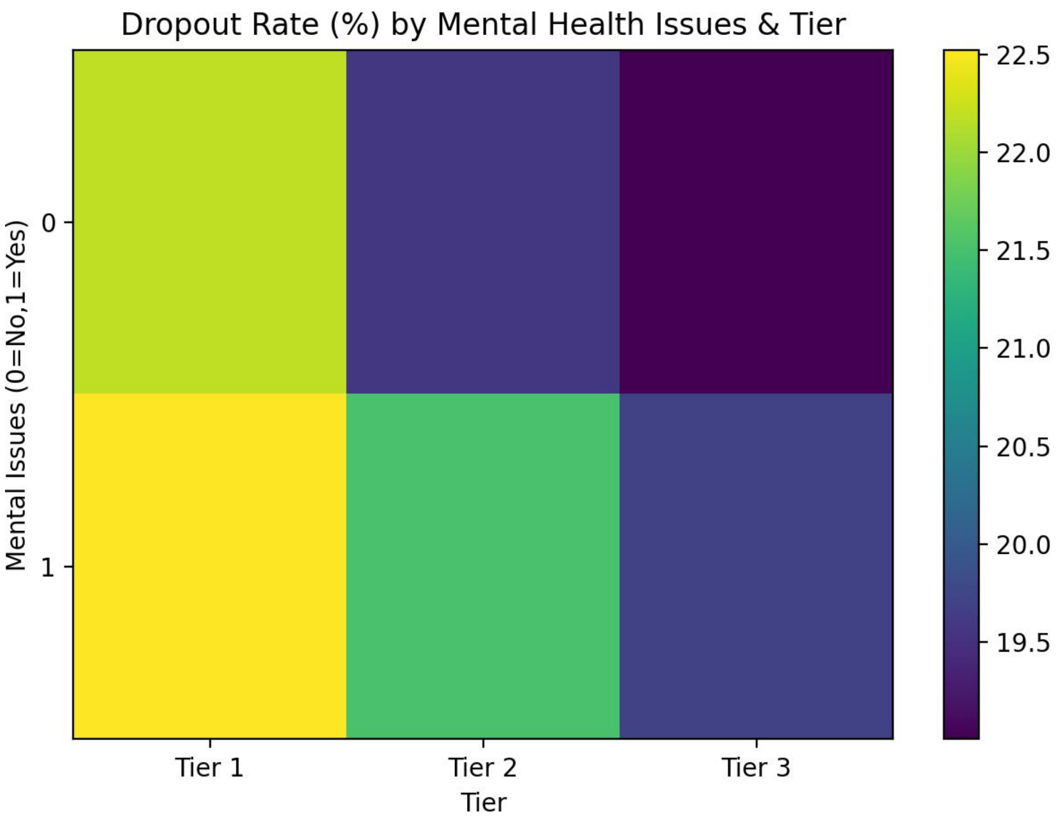

Figure 2: Heatmap: Dropout rate (%) by mental-health issues (0/1) and tier.

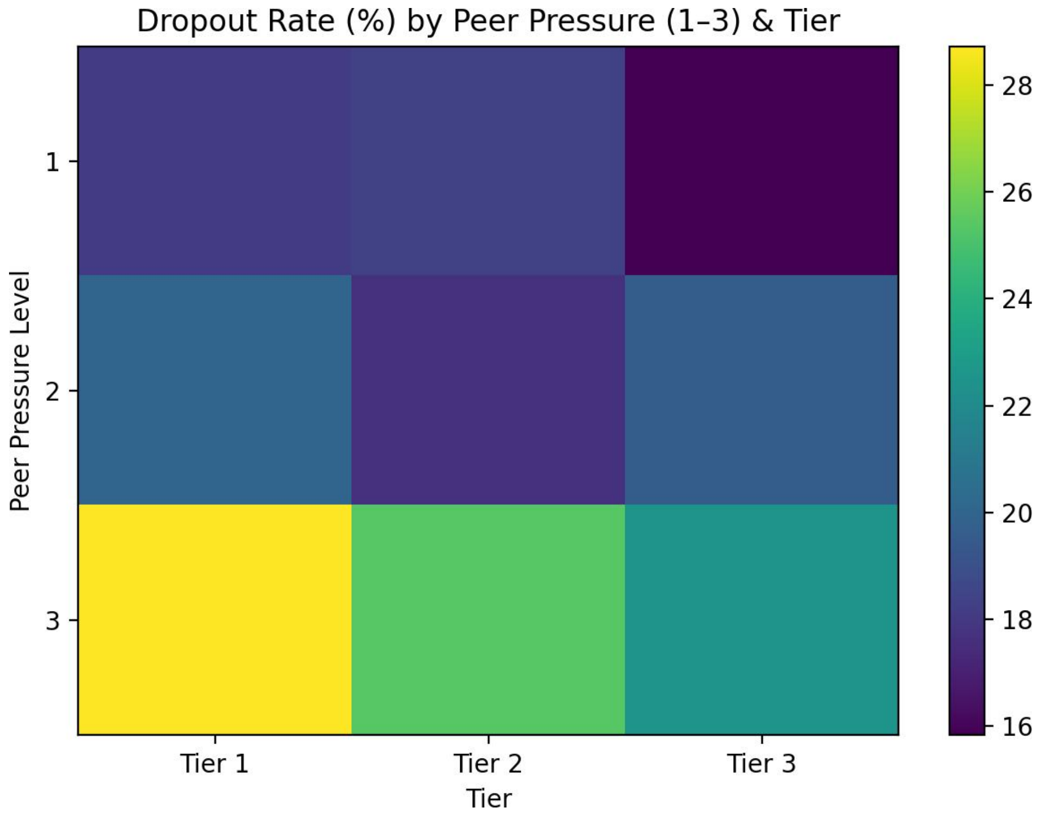

Figure 3: Heatmap: Dropout rate (%) by peer pressure (1–3) and tier.

Insight. Moving from low→high peer pressure increases dropout materially across tiers (e.g., Tier-1 roughly 18%→29%). Mental-health "Yes" cells consistently show higher risk.

6 Predictive Modeling Results

6.1 Performance

Table 6: Model performance (test split)

Model	**Accuracy**	**ROC-AUC**
Logistic Regression	0.801	0.857
Random Forest (400 trees)	0.822	0.898

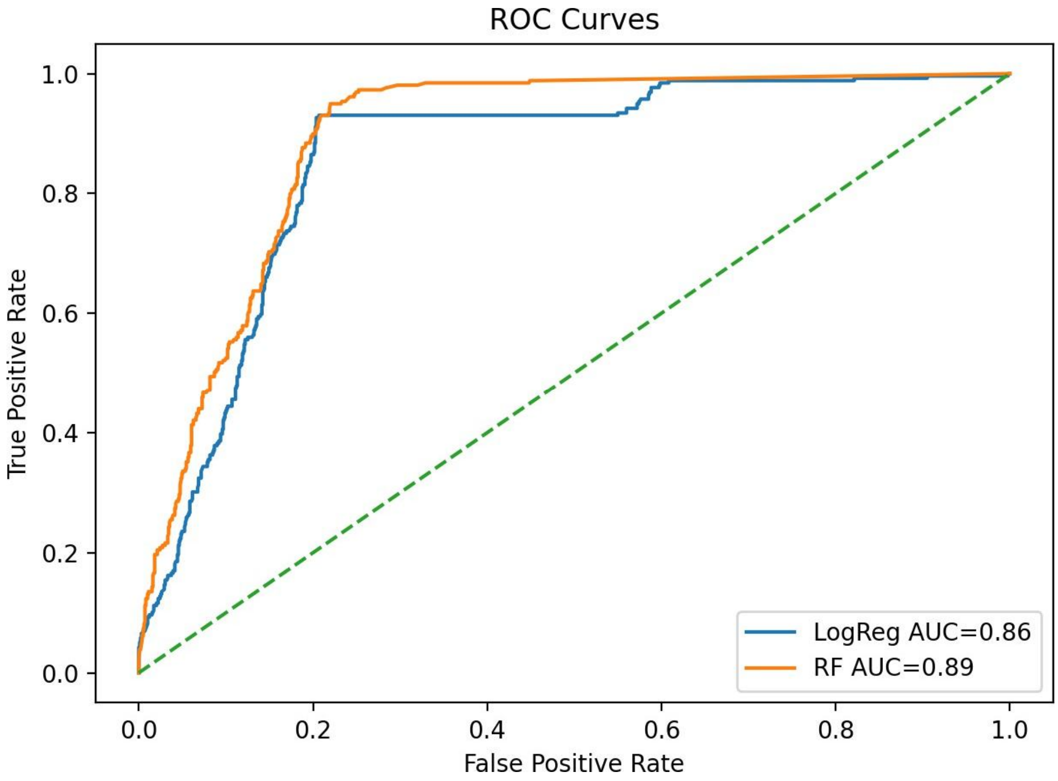

Figure 4: ROC curves for Logistic Regression and Random Forest.

6.2 Interpretation & Drivers

Directionally (from LR signs & RF importances):

- *Risk up:* higher peer pressure, mental-health issues.
- *Risk down:* more study hours, higher parent education.
- *Modest net effects:* income level, coaching flag (after controlling for other variables).
- *Context proxy:* tier picks up access and support differences.

Operationalization. Use the RF risk score to trigger: (i) targeted nudges, (ii) mentor outreach, (iii) personalized content and pricing (see Recommendation System).

7 Segmentation (KMeans, $k = 3$)

Clustering on affordability & context features yields three actionable profiles:

1. **Cluster A: Coaching-heavy, low MH flags** — seeks advanced practice, fine-grained tests; upsell is topic deep-dives and mock test packs.
2. **Cluster B: Coaching-heavy, high MH flags** — benefits from stress management, habit

loops, and peer de-pressure; need mentor check-ins and short daily routines.

3. **Cluster C: No coaching** — needs access, affordability, and community study pods; emphasize low-cost bundles, vernacular explainers, and weekend doubt clinics.

All clusters show ~21% base dropout; retention levers differ by segment (content depth vs. support vs. access).

8 Market Sizing (TAM–SAM–SOM, Illustrative)

We propose an assumption-driven funnel to bound opportunity in Tier-2/3:

- **TAM** (Total Addressable): all Tier-2/3 Class 12 STEM learners interested in engineering prep.
- **SAM** (Serviceable Available): those with smartphone access and willingness to pay micro-bundles.
- **SOM** (Serviceable Obtainable): realistic share via school tie-ups, micro-centers, and referrals in year 1–2.

Table 7: Illustrative funnel (assumptions tuned per state)

Stage	Share	Count (example)
Tier-2/3 STEM learners	–	20.0M
Smartphone + WA/SMS reachable	70%	14.0M
Willing to pay (200–500/m)	50%	7.0M
Reachable via schools/micro-centers	40%	2.8M
Initial obtainable share (SOM)	20%	0.56M

This aligns with a **medium-term target of 10–12M learners** served cumulatively across states as distribution deepens (school networks, mentors) and product expands to boards and vernaculars. All figures are *illustrative* and should be localized by state.

9 Recommendation System (Improved, Risk-Aware)

9.1 Architecture

1. **Risk Scoring Layer** (RF probability): flags high-risk learners from behavior + context features.
2. **Policy Layer**: maps risk bands & segments to interventions (content, price, mentor, cadence).

3. **Personalization Layer**: content sequencing (topic gaps), language, difficulty; dynamic pricing within guardrails.
4. **Nudging Layer**: habit formation (daily 15-min), peer de-pressure micro-content, exam-stress playbooks.
5. **Mentor Routing**: auto-assigns mentor slots to high-risk users; triages group vs. 1:1.

9.2 Policy Examples (BYJU's/PW Differentiation)

Table 8: Intervention policies by segment (illustrative)

Segment	**Intervention Set**	**Why it wins vs. BYJU's/PW**
A: Coaching-heavy, low MH	Advanced mocks, performance analytics, topic deep-dives, timed drills	Depth & granularity for high-commitment users; better retention than generic video-first flows
B: Coaching-heavy, high MH	Daily 15-min routines, stress modules, study circles, mentor check-ins	Tackles peer-pressure & anxiety head-on; retention moat often missing in incumbents
C: No coaching	200–500 micro-bundles, vernacular explainers, weekend doubt clinics, school-tie-up packs	Affordable access + offline touch-points; localized delivery vs. large one-size-fits-all SKUs

9.3 Pricing Personalization (Guardrailed)

- Base: 199–299/m micro-bundle; add-ons 49–99 (tests, doubts).
- *Conditional discounts*: scholarship ladder (attendance + improvement → fee relief).
- *Family plan*: sibling at 50%.

9.4 KPI Tie-in

North Star: streak-keeping learners with weekly practice minutes ↑ and risk score ↓.
Guardrails: CAC, refund rate, time-to-intervention; equity checks for pricing fairness.

10 Go-To-Market (Tier-2/3)

10.1 Product

- **Lite vernacular app**: offline-first; SMS/WhatsApp drills; parent dashboard in local language.

- **Peer & mental-wellbeing**: anonymous study circles; exam-stress micro-lessons.
- **Risk-based nudges**: mentor outreach when risk crosses threshold; A/B library of nudges.

10.2 Distribution

- **Micro-centers** at district/subdistrict hubs (weekend doubts & tests).
- **School partnerships**: teacher referrals; 10-min morning practice bell synced to app.
- **Community mentors**: ex-toppers/college students running neighborhood cohorts.

10.3 Positioning vs BYJU's/PW

- **Retention moat**: risk-aware ops + MH/peer de-pressure content.
- **Local depth**: boards & languages tuned to Tier-2/3 realities.
- **Affordability**: modular micro-bundles that grow ARPU with outcomes (not promises).

11 Execution Metrics & Ops

Activation: D7 streaks, first 3 quizzes completion.
Engagement: weekly Q&A, mentor touchpoints, practice minutes.
Risk Ops: time-to-intervention for high-risk flags; resolution rates.
Economics: CAC, ARPU, cohort LTV; center ROI (footfall → paid conversion).
Quality: topic mastery uplift; NPS; refund ratio.

12 Limitations & Ethics

Tier mapping is a proxy for access; self-reported mental-health/peer items are noisy; coaching label is coarse (no hours/intensity). Risks: over-targeting by income segment; pricing fairness; mentor load. We recommend bias audits on the risk model and transparency to learners and parents.

13 Conclusion

Peer pressure and mental-health indicators are the most consistent dropout drivers across tiers, while coaching has a modest net effect after controls. A retention-led strategy in Tier-2/3—vernacular micro-bundles, offline touchpoints, and a risk-aware recommendation engine—can capture demand sustainably while improving learner outcomes.

References (Illustrative)

1. Client-provided reference report (*ed.pdf*).
2. Market/industry overviews (e.g., consulting and industry associations).
3. Standard ML texts for classification and clustering (e.g., Hastie, Tibshirani, Friedman).

A Appendix A: Reproducibility Notes

Figures are generated from the same CSV via a Python script that:

1. Normalizes columns & encodes features.
2. Produces: fig_dropout_by_tier.png, fig_heatmap_mental_tier.png, fig_heatmap_peer_tier.png, fig_roc_curves.png.
3. Trains LR and RF with a 75/25 stratified split and renders ROC curves.