Intro to AI | CS 534

MED-BR: THE MULTI-MODAL EQUITY-DRIVEN BUSINESS RECOMMENDER

Undergraduate Computer Science, Worcester Polytechnic Institute

Keerthana Jayamoorthy kjayamoorthy@wpi.edu

AGENDA

1. Introduction
2. Background
3. Methodology
4. Results
5. Demonstration
6. Conclusion

I. INTRODUCTION

The Problem

- Business success is critically dependent on location.
- 60% of businesses making a location error experience a decline in revenue.
- In Massachusetts, the first-year business failure rate is 9.2%.

I. INTRODUCTION

Motovation

- Boston, Massachusetts, ranked as the third most intensely gentrified city in America during the 2013-2017 study period.
- Experienced 20% or more of its eligible neighborhoods undergoing gentrification during the 2013-2017 period.
- Was included in the list of the 20 "intensely gentrifying" metro areas. This group of cities contained at least ten neighborhoods that gentrified during that time.
- Was among the top 30 cities for intensity of gentrification during the earlier report period (2000-2012)

J. Richardson, B. Mitchell, and J. Edlebi, "Gentrification and Disinvestment 2020." NCRC. [Online]. Available: Motovation. [Accessed: Dec. 11, 2025].

I. INTRODUCTION

Current Gaps

- **Gap 1:** Time-Awareness: SOTA models are often static, failing to model daily demand cycles.
- **Gap 2:** Socio-Economic Equity: Current methods lack an objective function to maximize equitable access

Proposed Solution

- Develop the MED-BR, a specialized Graph Neural Network (GNN).
- **Core Innovation:** Use a Dual-Objective Loss Function (L_total) for constrained optimization.

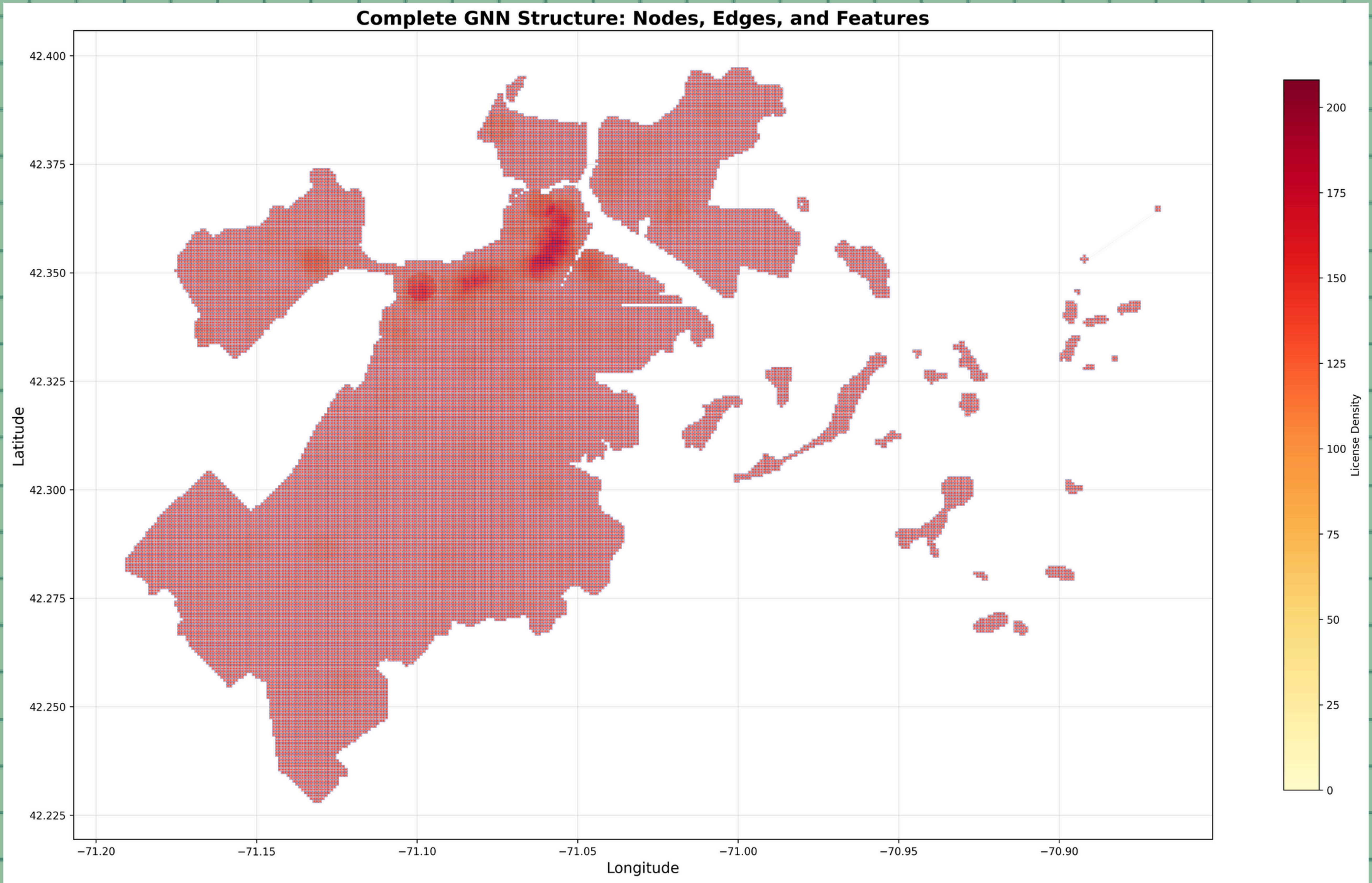
II. Background

Method	Baseline Function (Advantage)	Critical Gaps
Random Forest Classifier (RFC)	Strong Non-Spatial Baseline	Fails to account for Spatial Dependence
Dynamic Huff Model	Addresses basic Time-Awareness	High reliance on proprietary data; Lacks an Equity Objective
DBSCAN + MMNL	Filters viable locations based on constrained Accessibility	Lacks any mechanism for Social Fairness or Equity optimization
Graph Convolutional Networks (GCN)	Superior at incorporating Spatial Context (Spatial SOTA)	Mostly Static; Lacks Explicit Fairness Goals

II. Background

MED-BR FOUNDATION

Component	What is it?
Node (v_i)	A 100 x 100 grid cell in Boston.
Edges (E)	Connections to the 8 nearest neighboring nodes.
Feature Vector (X)	The 16 multi-modal data inputs aggregated to this location.
Output Score (S_i)	The final Survival Probability score ($\in [0, 1]$).



II. Background

MED-BR FOUNDATION

Component	Description	Data/Metric Source & Processing	Associated Regularizer/Objective
Node Definition (V)	A single, viable 100 x 100 grid cell in Boston	**Source:** Parcels (2024) dataset. Processing: Used to filter out non-viable areas (e.g., water, major parks).	N/A
Edge Definition (E)	Connections to its k=8 nearest neighbors.	**Source:** Calculated based on the spatial distance between node centroids. **Processing:** Defines the local spatial dependencies crucial for the GCN.	N/A
Input Feature Vector (X)	16 heterogeneous data inputs aggregated to the 100m level.	**Sources include:** Transit stops, licenses, violations, census data , BPDA ACS demographics, Time-Segmented Ridership (from MBTA transit data). **Processing:** Aggregated using spatial joins and density calculations (e.g., competition density within 500m).	L_success
Equity Weights	Weights derived to calculate equitable access for vulnerable populations.	**Sources include:** Median Income and Displacement Risk features (from BPDA ACS income features). **Processing:** Used to calculate the Kolm-Pollak EDE metric (E).	R_equity
Output Score (S_i)	The network's final predicted survival probability score	**Processing:** Output of the simple stack of GCNConv layers. Used to calculate the loss function components.	L_total

III. METHODOLOGY

The Dual Objective Loss Function

$$L_{total} = \underbrace{\mathcal{L}_{\text{success}}}_{\text{Business Success}} + \lambda_{\text{Time}}\mathcal{R}_{\text{time}} + \lambda_{\text{Equity}}(1 - E)$$

- **Purpose:** To successfully optimize business profit and social equity simultaneously, forcing them to compete for resources during training.
- **Impact:** Mathematically ensures the model pays a "measurable cost" to achieve mandated social constraints alongside predictive accuracy.

III. METHODOLOGY

Equity Regulizer

$$\text{Minimize: } \lambda_{\text{Equity}}(\mathbf{1}-\mathbf{E}) \longrightarrow \boxed{E = \left(\sum_{i=1}^{N} w_i S_i^{\rho}\right)^{\frac{1}{\rho}}}$$

- **Purpose:** To guide site recommendations specifically towards vulnerable and high-need neighborhoods defined by ACS income features.
- **Impact:** Achieves a substantial gain in equitable access (EDE value) compared to unconstrained, profit-only SOTA models.
- **Optimization Goal (The Loss):** We minimize the weighted penalty term, which is equivalent to maximizing the EDE value.
- **EDE Calculation (The Metric):** The Kolm-Pollak EDE value (E) uses need-based weighting (w_i) derived from Median Income and Displacement Risk features to measure accessible social welfare:

III. METHODOLOGY

Time Regulizer

$$\mathcal{R}_{\text{time}} = \frac{1}{|T|} \sum_{t \in T} (S_{i,t} - \bar{S}_i)^2$$

- **Purpose:** To ensure the model's predicted survival score (S_i) is stable across daily demand cycles (morning, afternoon, and evening).
- **Impact:** Prevents the model from relying on static urban features, making the recommendation robust and truly time-aware.
- **Mechanism:** Minimizes the variance of the predicted scores across the three user-defined peak time periods

III. METHODOLOGY

Data Procurement

Data Component	Source Data Set(s)	Processing Method	Used for...
Graph Structure (V,E)	Parcels (2024) dataset, Spatial distances.	Discretization into 100m x 100m grid cells. Filtering out non-viable areas (e.g., water, parks).	GNN Node/Edge Definition
Time-Aware Features	MBTA transit data.	Aggregated as Time-Segmented Ridership.	$\mathbf{X}$ Input / $\mathcal{R}_{\text{time}}$ Regularizer
Socio-Economic/Equity Features	BPDA ACS Demographics.	Aggregated as Median Income and Displacement Risk features.	$\mathbf{X}$ Input / $\mathcal{R}_{\text{equity}}$ Regularizer
Business/Competition Features	Cleaned data (licenses, violations, census).	Density Calculations (e.g., competition density within a 500m radius).	$\mathbf{X}$ Input / $\mathcal{L}_{\text{success}}$ Objective

III. METHODOLOGY

GNN Creation

$$H^{(l+1)} = \sigma(\tilde{D}^{-\frac{1}{2}}\tilde{A}\tilde{D}^{-\frac{1}{2}}H^{(l)}W^{(l)})$$

Variable	Description
$H^{(l)}$	Input feature matrix at layer l
$H^{(l+1)}$	Output feature matrix at layer l+1
$W^{(l)}$	The weight matrix being learned
$\tilde{A}$	The normalized adjacency matrix (captures neighbor connections)
$\tilde{D}$	The degree matrix of A
σ	The activation function

- **Model Type:** Shallow Graph Convolutional Network (GCN), implemented using PyTorch Geometric (PyG).
- **Configuration:** The model uses a simple stack of GCNConv layers with hidden_dims=[64, 128, 64].
- **Operation:** The core is the GCN convolutional operation, which transforms node features between layer l and layer l+1.

III. METHODOLOGY

Overall Data Flow

1

Raw Data Ingestion:
Heterogeneous data (Transit, ACS, Licenses) are integrated

2

Graph Construction:
Data is spatially joined and aggregated to the 100m x 100m node level

3

GNN Input:
The feature vector (X) enters the GCN layers

4

Forward Pass:
The GCN processes X and outputs the survival score S_i

5

Loss Calculation:
S_i is fed into the L_total function, where the R_time and R_equity terms apply the constrained optimization.

6

Backpropagation
The total loss gradient is used to update weights, ensuring the dual objectives (profit and equity) are optimized simultaneously

IV. EXPERIMENTAL RESULTS

Spatial SOTA Baseline

BCE Loss ($\mathcal{L}_{\text{success}}$):

$$\mathcal{L}_{\text{success}} = -\frac{1}{N}\sum_{i=1}^{N}\left[y_i \log(p_i) + (1-y_i)\log(1-p_i)\right]$$

F1 Score:

$$\text{F1} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

- **Success Metric:** We measure the final predictive skill using the F1-Score.
- **Baseline Result:** This profit-only approach yielded a baseline F1 Score of 0.7783

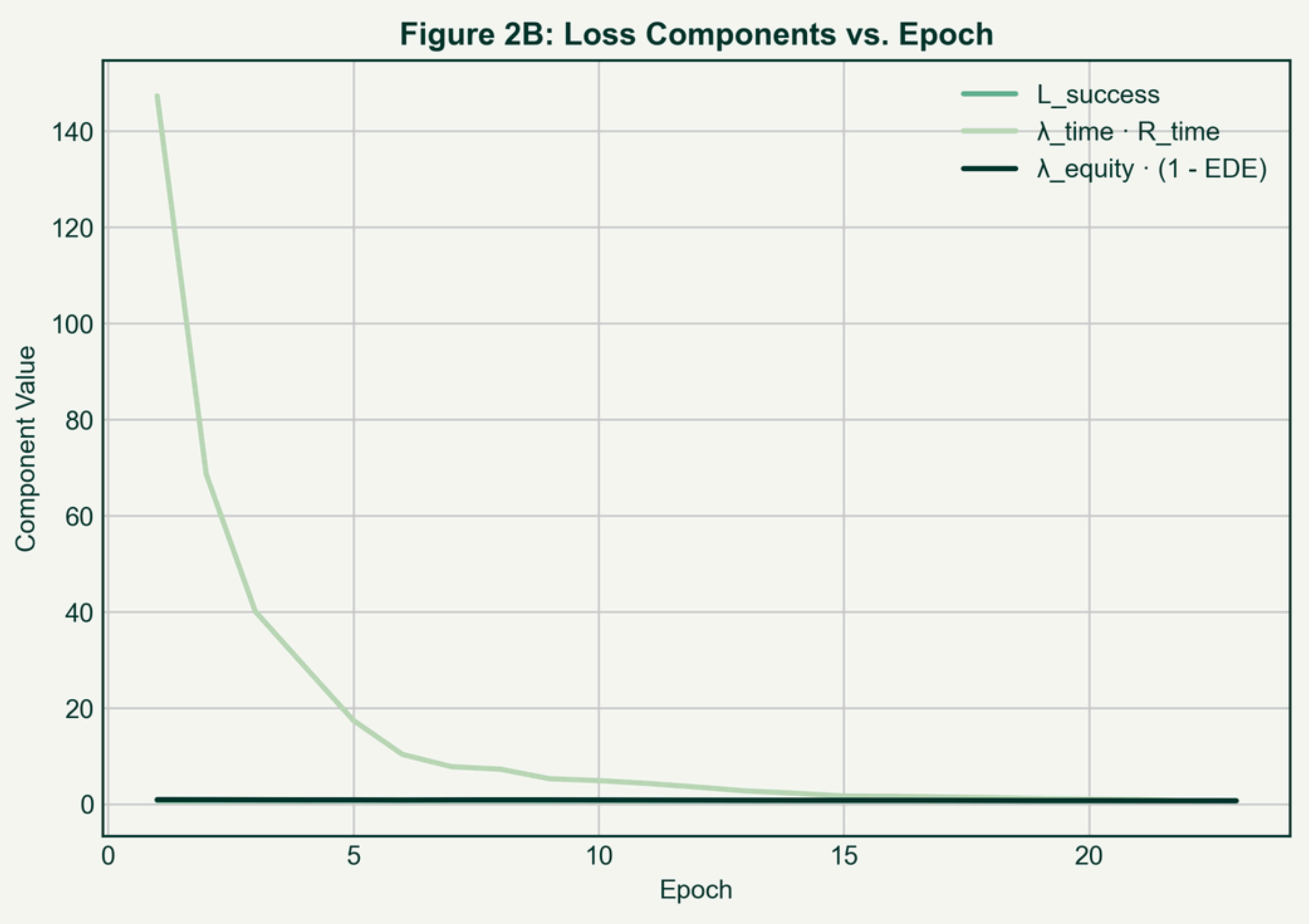
- **Process:** We train this model by feeding all our node features (X) through the GNN and optimizing using the standard Binary Cross-Entropy (BCE) Loss.
- **Purpose:** BCE Loss is the foundation of L_success. It measures the difference between the predicted survival probability (p_i) and the actual outcome (y_i)

IV. EXPERIMENTAL RESULTS

Training The Loss: Initial Issue

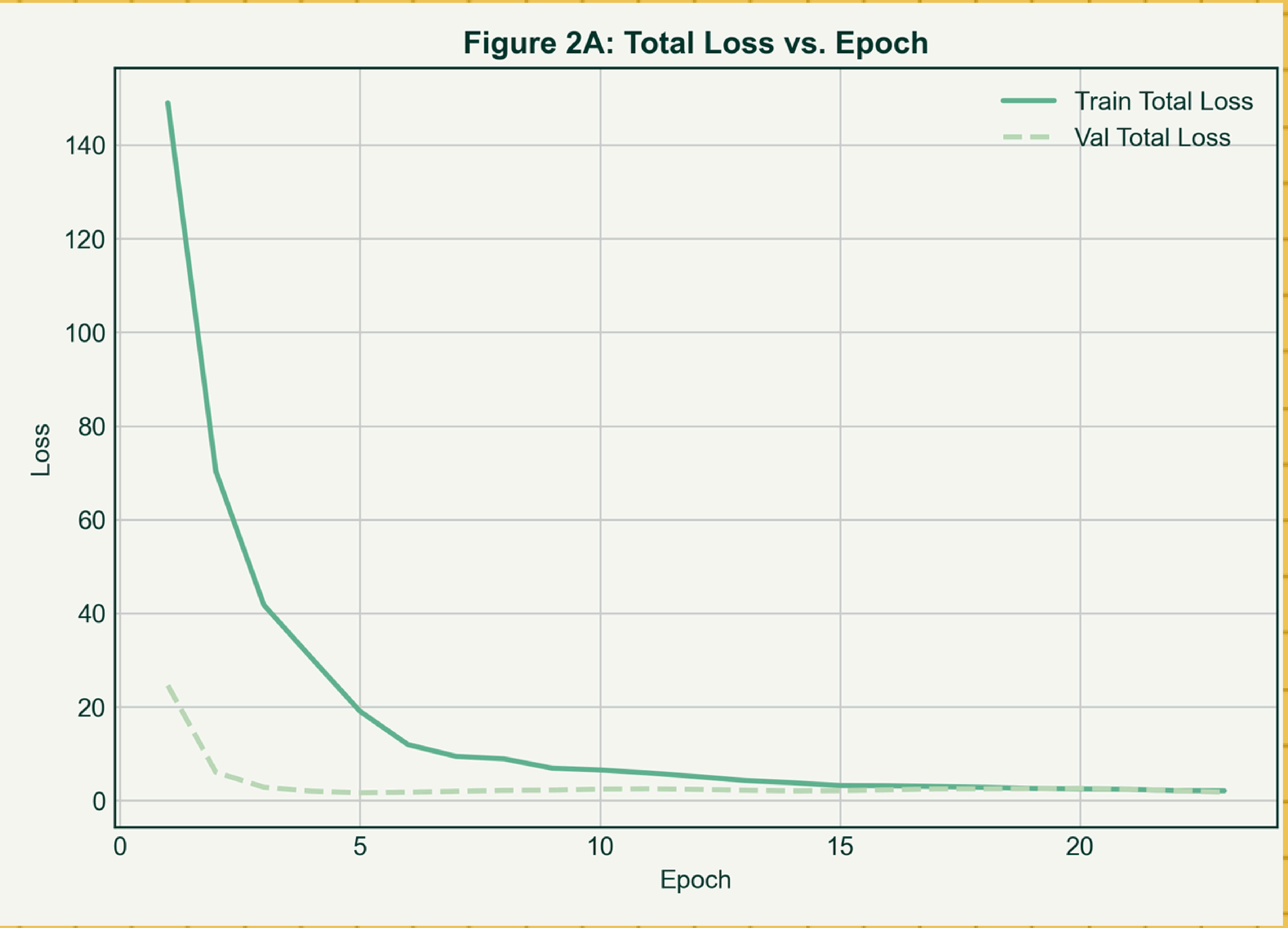
- **The Challenge:** Raw regularization terms (R_time, 1-E) were approximately 10^-5, too small to influence the gradient compared to the BCE Loss (approx 0.1)
- **The Result:** The dual-objective model initially operated as a "silent regularizer," failing to optimize the constraints
- **The Solution:** Applied a large Magnification Factor (e.g., x 50,000.0 for R_time) to scale the terms into the active gradient range
- **Proof of Success:** The sharp drop and convergence of the weighted regularization terms in the loss history prove that the fix successfully forced the model to optimize the dual objectives

Component	Initial Raw Magnitude
L_success (BCE Loss)	0.1
R_time (Variance)	0.00001
R_equity (EDE)	0.00001



IV. EXPERIMENTAL RESULTS

Training The Loss



- **Successful Convergence:** Training and Validation Loss drop quickly and converge smoothly over 23 epochs
- **Proof of Stability:** The consistent tracking of the validation curve confirms the model is stable and did not severely overfit the training data.
- Validates the entire GNN architecture, ensuring the foundation is sound before the final results were measured.

IV. EXPERIMENTAL RESULTS

Key Performance Gains

Optimal Configuration: The best trade-off was found at lambda_Equity=3.00

F1 Score Improvement

Baseline		MED-BR
0.7783	+17.43% →	0.9526

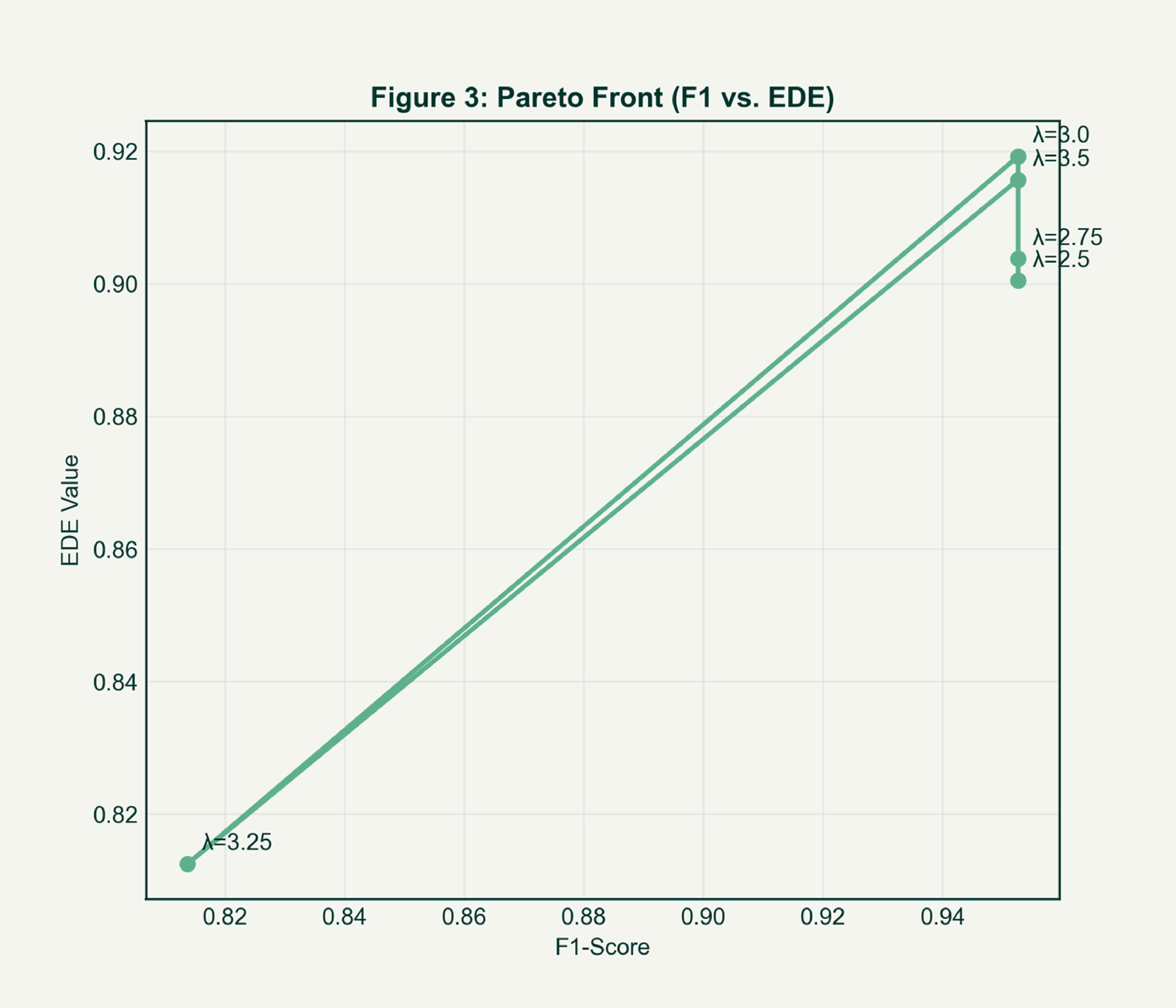
EDE Score Improvement

Baseline		MED-BR
0.8145	+12.85% →	0.9192

IV. EXPERIMENTAL RESULTS

Paretto Front

- **Purpose:** to find the single optimal balance point between two competing objectives (Profit vs. Equity)
- **Mechanism:** The front plots the entire range of potential solutions, mapping the trade-off achieved by varying the weight of the equity constraint
- **Optimal Choice:** maximum achievable EDE gain before F1-Score begins to drop significantly
- **Result:** lambda_Equity=3.00



V. DEMO

Output

- Graph of Suitable Nodes
- Information on Each Node
- Link to coordinates on Google Maps

User Input

- Type of Food Establishment
- Short Description
- Number of Locations

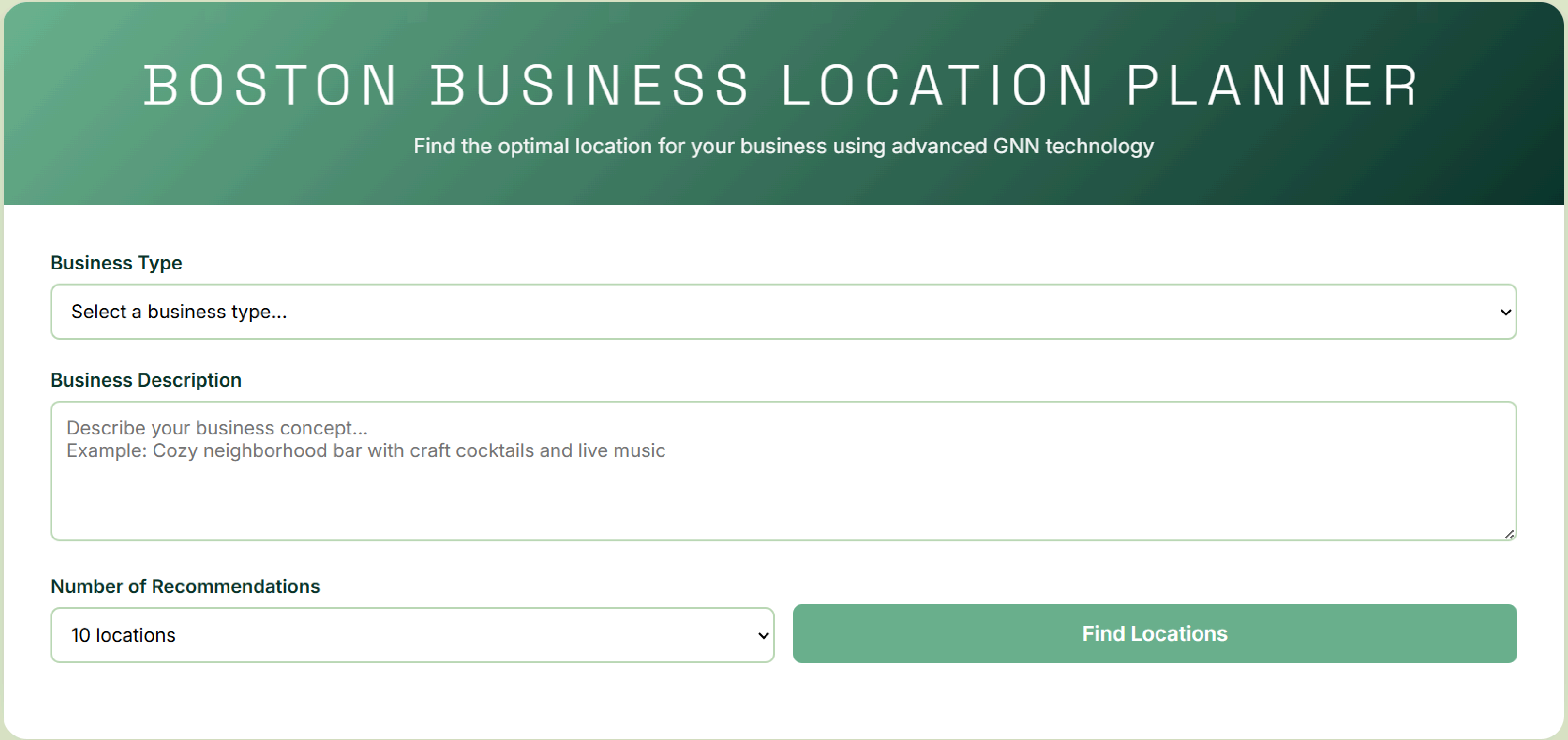
Form collects business type (dropdown), description (text), and number of recommendations. On submit, sends POST to /recommend with form data.

Behind the Scenes

Loads the production GNN model and precomputed spatial features (~6,190 grid cells). Runs inference on all nodes using only spatial features (transit, demographics, competition, violations).

V. DEMO

Input


BOSTON BUSINESS LOCATION PLANNER
Find the optimal location for your business using advanced GNN technology
Business Type
Select a business type...
Business Description
Describe your business concept...
Example: Cozy neighborhood bar with craft cocktails and live music
Number of Recommendations
10 locations
Find Locations

V. DEMO

Text Output

Recommended Locations

Model Accuracy: 81.7% | Equity Score: 79.6%

List View | Map View

Location #1 (Node 2607)

81.1% - Good

Coordinates	License Density	Traffic Signals	Building Violations	Property Quality
42.345657, -71.068792 View on Map →	31 competitors	13	0	0 violations

Location #2 (Node 2608)

81.0% - Good

Coordinates	License Density	Traffic Signals	Building Violations	Property Quality
42.345657, -71.067894 View on Map →	32 competitors	15	0	0 violations

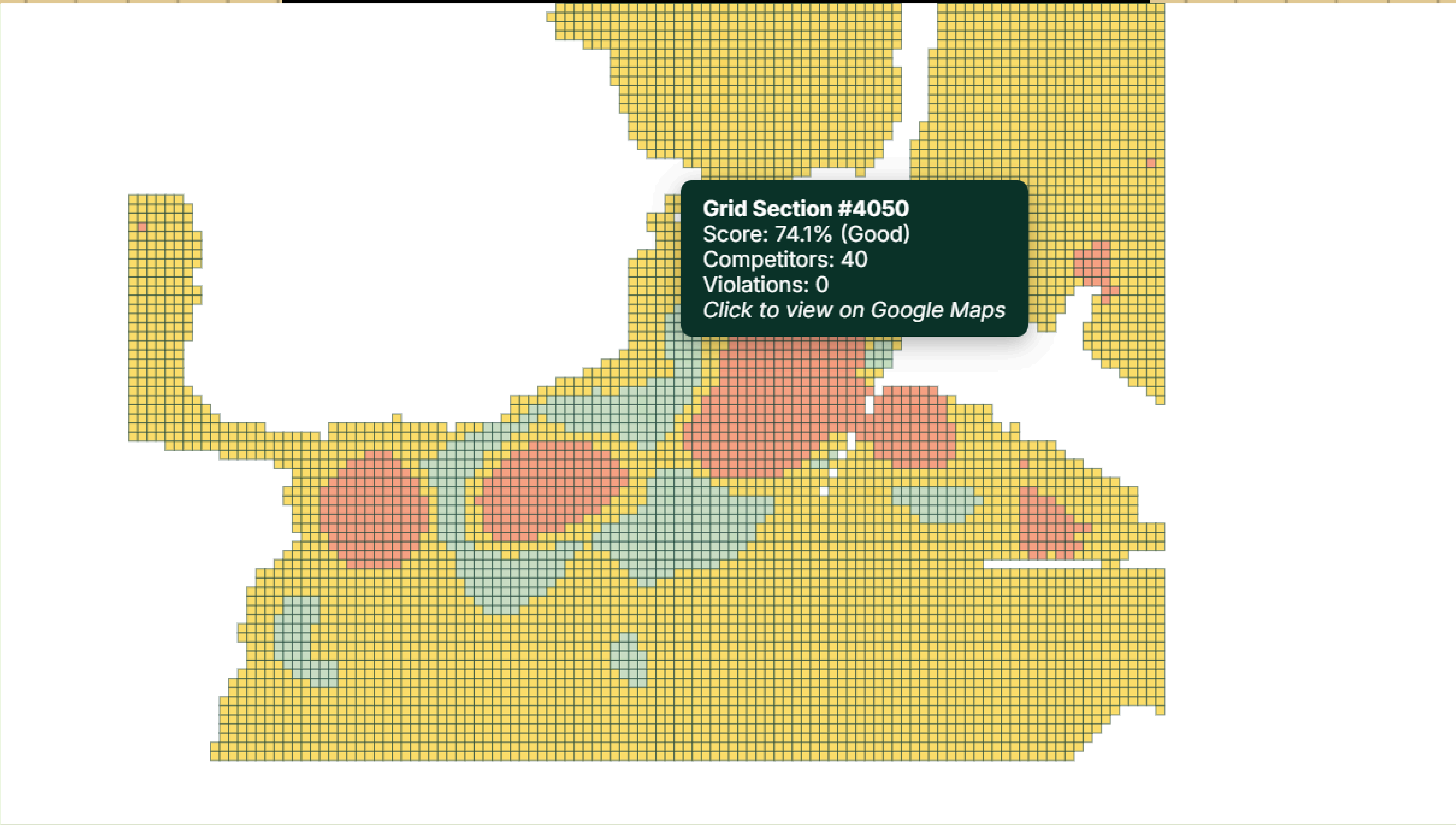
Location #3 (Node 2700)

81.0% - Good

Coordinates	License Density	Traffic Signals	Building Violations	Property Quality
42.346321, -71.068792 View on Map →	25 competitors	15	0	0 violations

V. DEMO

Visual Output



Excellent (90%+) Good (70-90%) Moderate (50-70%) Lower (<50%)

VI. CONCLUSION

Limitations

Limited Scope: The model is currently constrained to food establishments (restaurants/cafes).

Data Constraint: We lacked extensive, high-quality business data and relied on establishment licenses, limiting the scope.

Workaround: We successfully used BPDA ACS Demographics as a proxy, maintaining the objective's integrity despite relying on aggregated survey data.

Proxy Equity Data: The ideal high-resolution NERD dataset was discontinued and inaccessible.

Key Lesson: Diagnosing and fixing the magnitude mismatch (the "silent regularizer" issue) was the most significant technical learning.

VI. CONCLUSION

Future Work

Industry-Specific Benchmarking

Dynamic Data Integration

Refining EDE

Larger Dataset

VI. CONCLUSION

Citations

Alhazzaa, M. S., et al. (2025). Exploring AI Applications for Optimal Business Location Selection in Smart Cities: A Literature Review. IEEE Transactions on Intelligent Transportation Systems.

U.S. Bureau of Labor Statistics. (2024). Business Employment Dynamics. [Data extracted for the Massachusetts one-year failure rate.]

Mahmud, K. R., et al. (2024). A Comprehensive Business Location Choice Model Leveraging Machine Learning in Systematic Choice Set. Urban Computing and Applications.

Lan, T., Cheng, H., Wang, Y., & Wen, B. (2022). Site Selection via Learning Graph Convolutional Neural Networks: A Case Study of Singapore. Remote Sensing, 14(15), 3579.

Liang, Y., Gao, S., Cai, Y., Foutz, N. Z., & Wu, L. (2020). Calibrating the dynamic Huff model for business analysis using location big data. Transactions in GIS, 24(3), pp. 681-703.

Richardson, J., Mitchell, B., & Edlebi, J. (2020). Gentrification and Disinvestment 2020. NCRC. [Online]. Available: ncrc.org/gentrification20/.

Chamber of Commerce. (2024). Small Business Statistics. [Online]. Available:

Thank You!

Github Repo

Research Paper