METRO ANALYTICS – TEAM 2

Predicting the Invisible: Risk-Based Modeling of Human Trafficking at JFK Airport & Surrounding Communities

The Team

Syed Hashim Raza
MBS - Analytics
Team Lead

Derya Kirca
MBS - Analytics

MBS ADVISORS

Prof. Brian Petrus
MBA, PHR
Associate Professor/HRM Marketing Coordinator

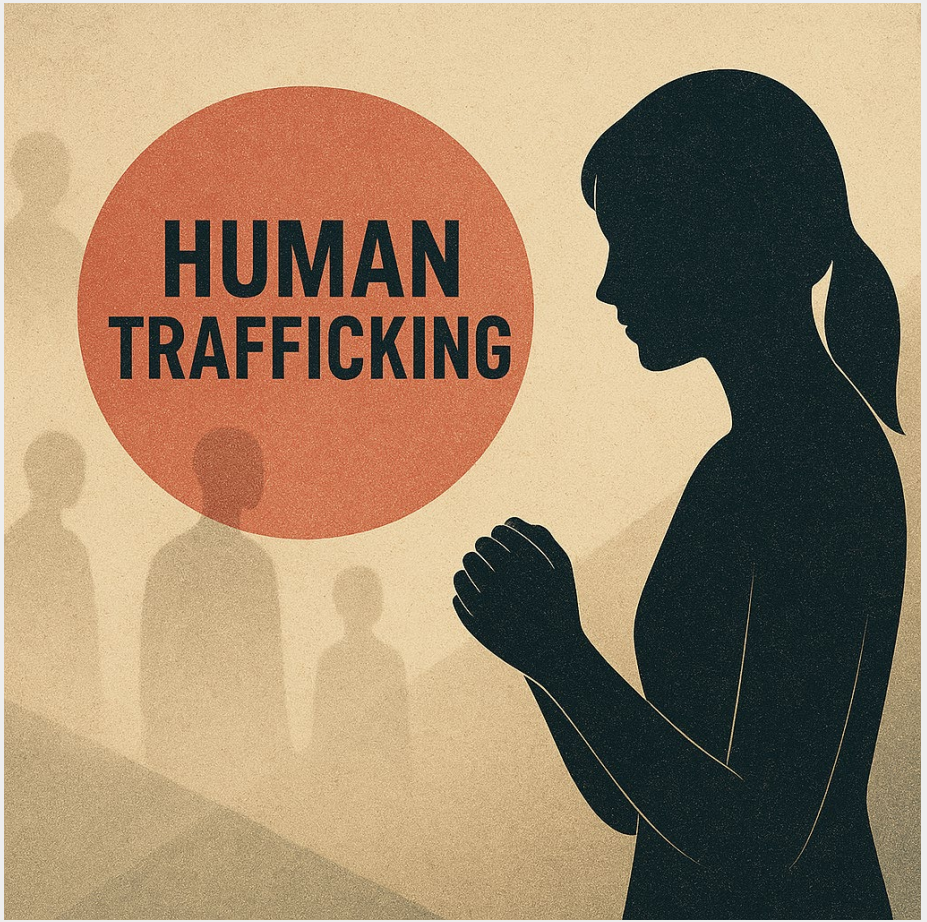
PROGRAM MENTORS

Prof. John Betak, Ph.D.
Sr. Operations & Management Consultant
Metro Analytics

Prof. Felipe Aros-Vera, Ph.D.
Transportation Planner
Metro Analytics

Background

- **50 million people** are living in modern slavery globally (*Walk Free Foundation 2021 Estimate*)
  - **Forced labor:** 27.6 million
  - **Forced marriage:** 22 million
- This equals **nearly 1 in every 150 people** worldwide
- Many trafficking cases go **unreported or misclassified**, as victims often do not disclose their exploitation
- Traffickers often target individuals who are vulnerable due to **poverty, limited education, or unstable living conditions**



https://www.walkfree.org/news/2022/50-million-people-worldwide-in-modern-slavery/

Problem Statement/Goals

Problem Statement

- Airports serve as **key transit hubs**.
- Despite access to tech, data, and extensive screening **there's no integrated system to predict and validate trafficking risks**.

Goals

1. Use spatial-temporal data to build **deep learning models** that identify **trafficking hotspots** in NYC, which can also be integrated into a **real-time detection system.**

2. Validate predicted hotspots by incorporating **socioeconomic data, land use, and vulnerability scores** to guide targeted interventions.

Building the Dataset

NYC OpenData

PORT AUTHORITY NY NJ

METRO ANALYTICS

NYC Crime Data

Observations: 32610
Date Range: 2021-2024

Filtered to only include crimes related to Human trafficking (i.e., kidnapping, child endangerment, drugs, prostitution, etc)

Attributes: Date, Location (Lat/Lon), type of crime

NYC Permitted Events

Observations: 996301
Date Range: 2021-2024

Filtered to only include moderate to large events in the NYC (i.e., Parades, concerts, festivals, sporting events)

Attributes: Date, location, event type, street closures

JFK Flight Inbound

Observations: 48
Date Range: 2021-2024

Monthly passenger count inbound to JFK

Attributes: Count of passengers monthly, Date, Total flights

Human Trafficking Reports

Observations: 8046
Date Range: 2021-2024

Dataset cleaned and filtered to only include cases reported in NYC

Attributes: Date, location (Lat/Lon), trafficking type (i.e., child exploitation, sex exploitation, forced criminality), victim gender

Worked on by whole team

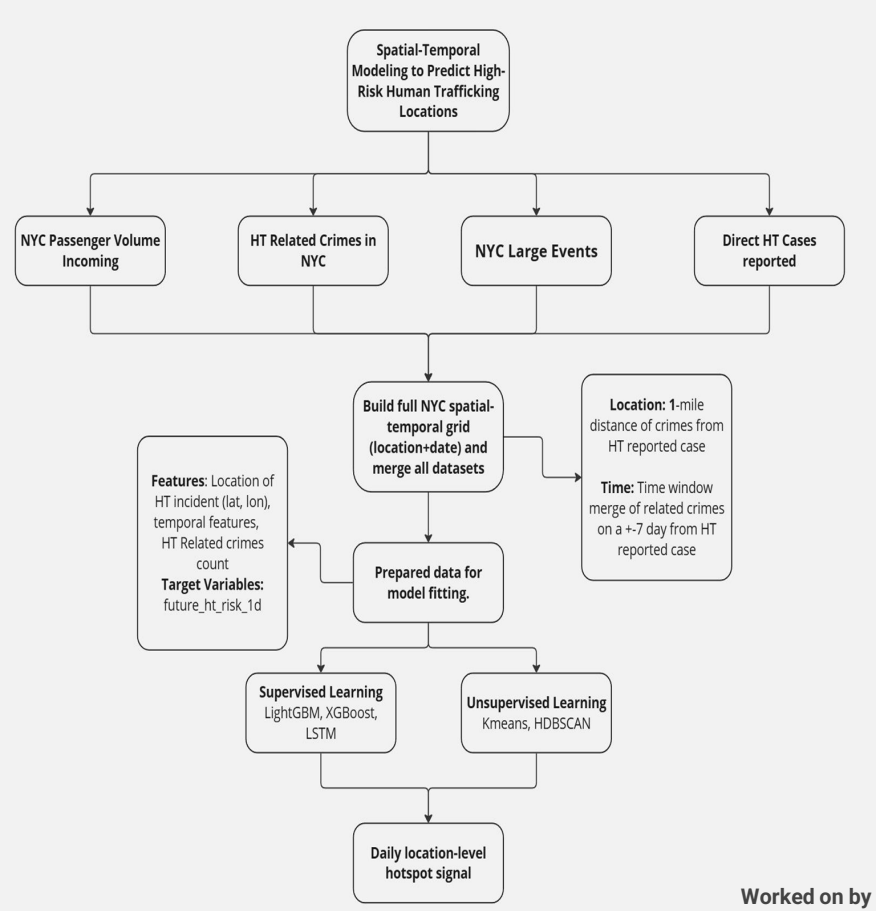
THE DATA: Grid Construction – Modeling Human Trafficking Risk Spatially

1. **Merged on Shared Context:**
   - **Location Proximity** (< 1 mile) and **time window** (±7 days) from know Human Trafficking cases

1. **Spatial - Temporal Grid Creation:**
   - Daily NYC grid (lat/lon x date) from **2021-2024**
   - ~**7.8M rows** - each = one location on one day

1. **Feature Engineering:**
   - **Spatial** : Location (Lat/Lon)
   - **Temporal** : Date, Day of week, holiday/weekend flags, seasons
   - **Crime Trend** s: 7-day rolling crime count, night crime ratio
   - **Passenger volume** : Monthly scaled inflow count
   - **Event count**

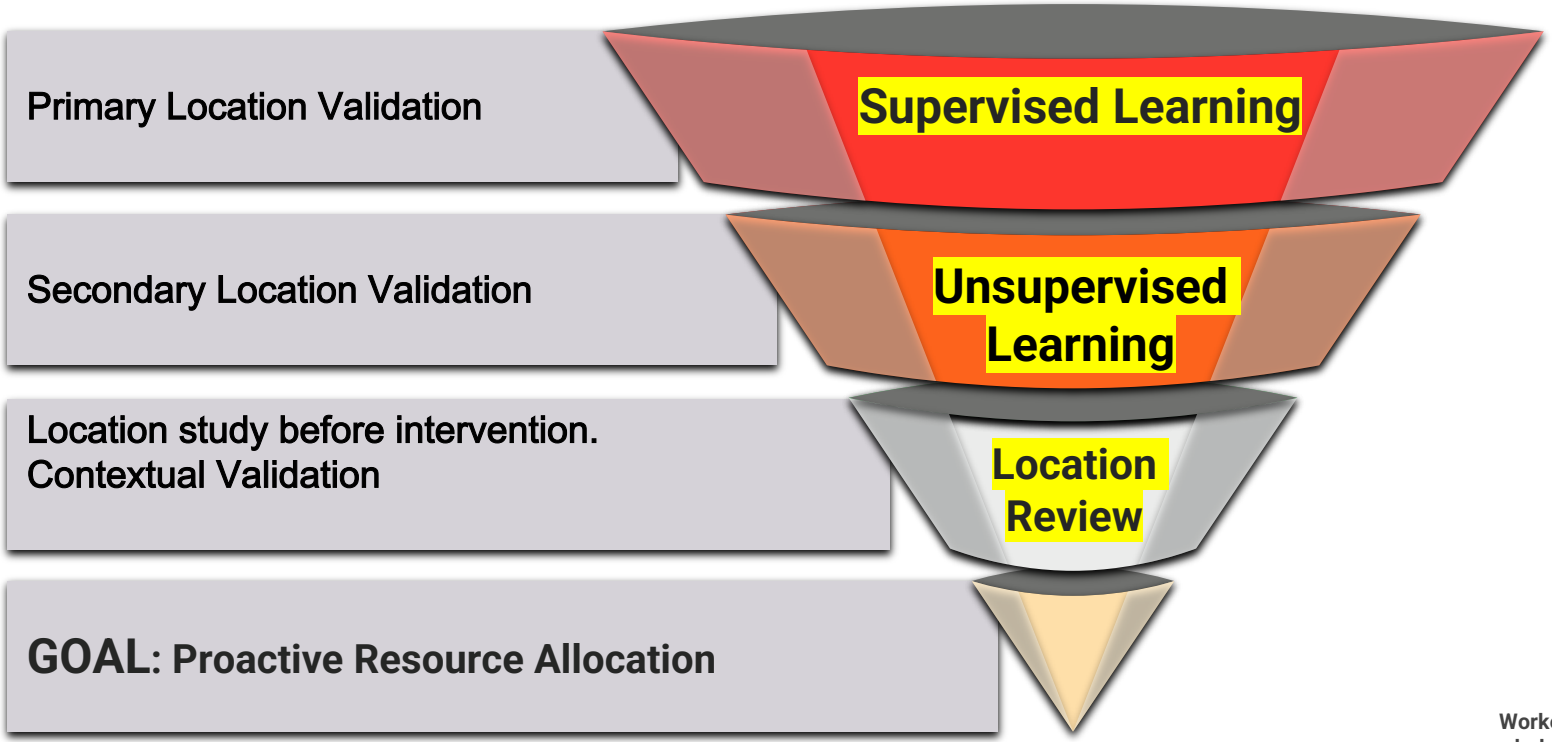
1. **Target Variable**
   - **Daily risk score per location**



Worked on by whole team

Project Approach

Multi-Layered Risk Model



Worked on by whole team

Baseline Models

LightGBM & XGBoost

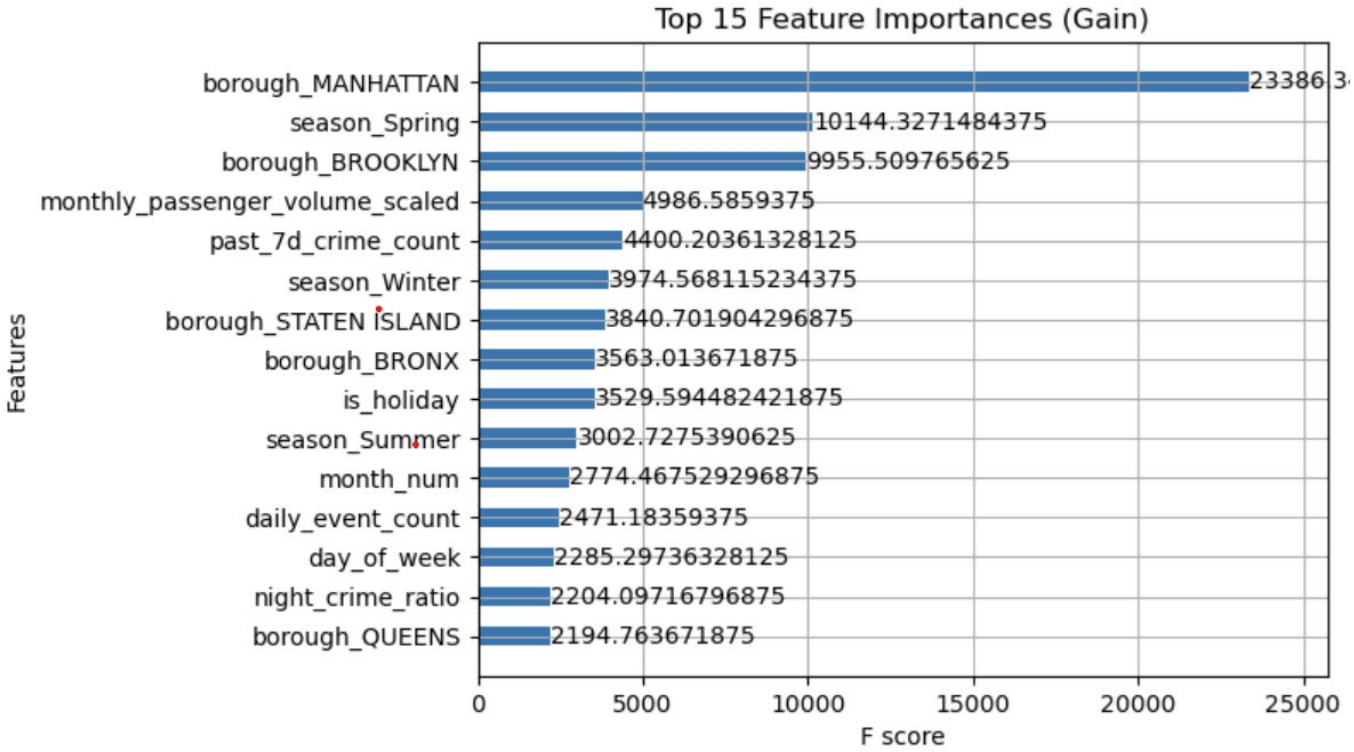
- Modeling Data Split: Training Data: 2021 - 2023 | Test Data: 2024
- **High accuracy and moderate AUC** — able to separate risky vs. non-risky locations
- Good for **benchmarking** and **feature insights**, but not sequence learning

	Accuracy	Recall	Precision	F1 Score	AUC
LightGBM	0.982	0.0162	0.0003	0.0005	0.5069
XGBoost	0.982	0.0243	0.0004	0.0007	0.7155

Key Insight:

We needed a model that could learn from **sequential patterns**

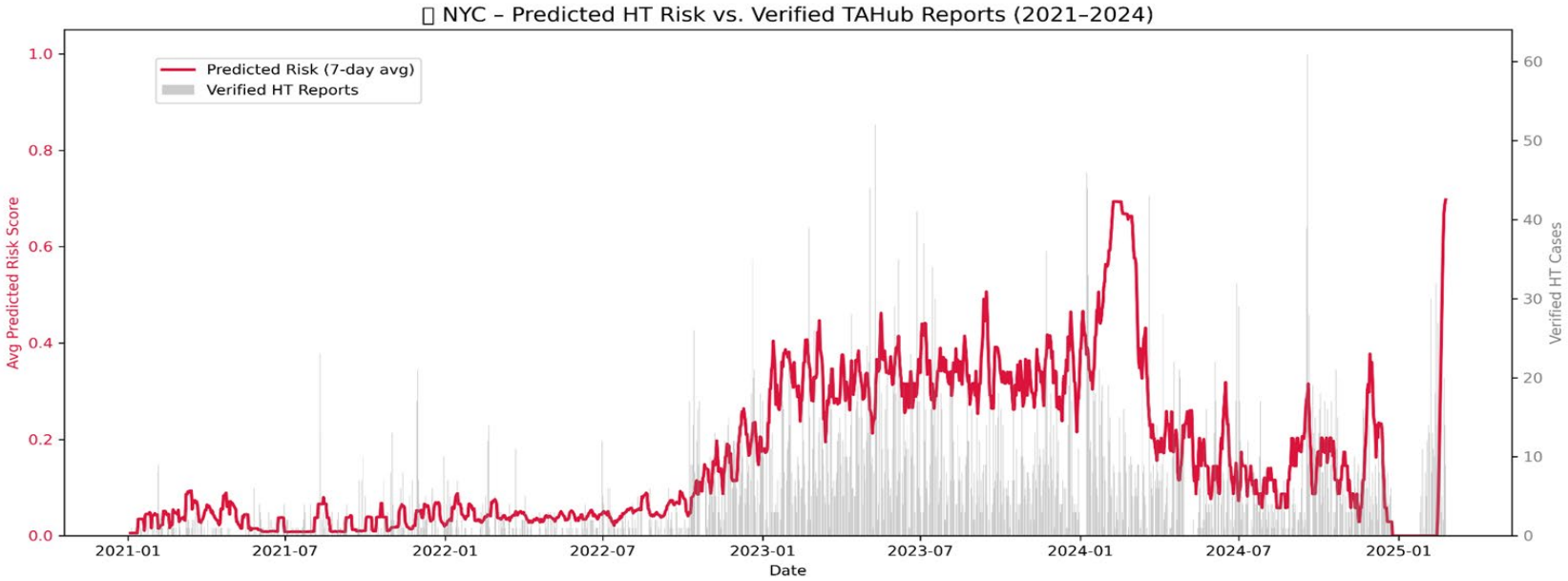
Feature Importance via XGBoost



Based on XGBoost's gain metric - how much each feature contributed to reducing model error.

Deep Learning Time-Series Model: LSTM

- Modeling Data Split: **Training Data: 2021 -2023 | Test Data: 2024**
- **AUC: 0.77 - Recall: 75% -** strong separation of high/low-risk areas
- Output: **Daily Location -level risk scores (0 -1)** for ranking and threshold tuning



Predicted Risk (Aggregated) vs Actual Human Trafficking Reports for NYC (All Boroughs) 2021 -2024

Example of Model Prediction

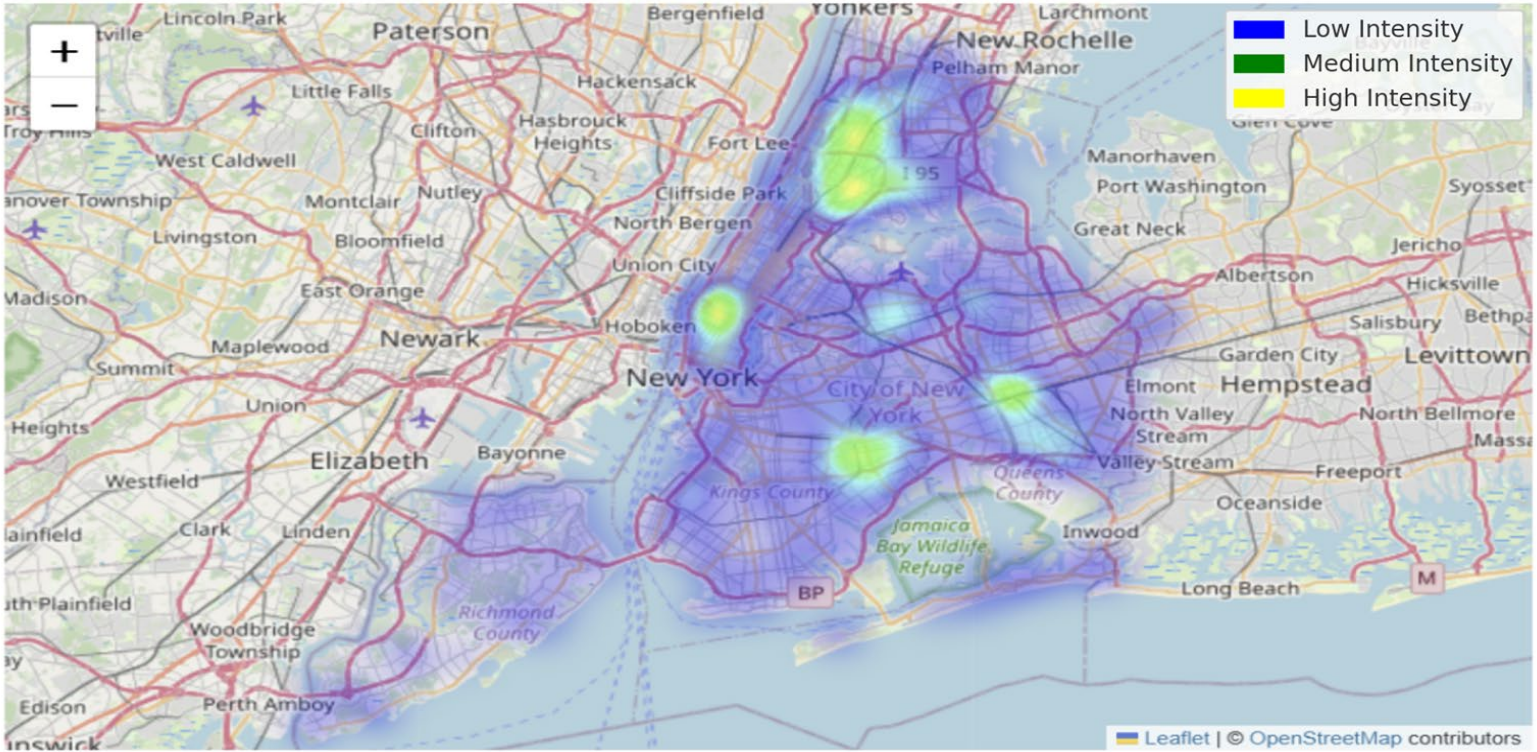


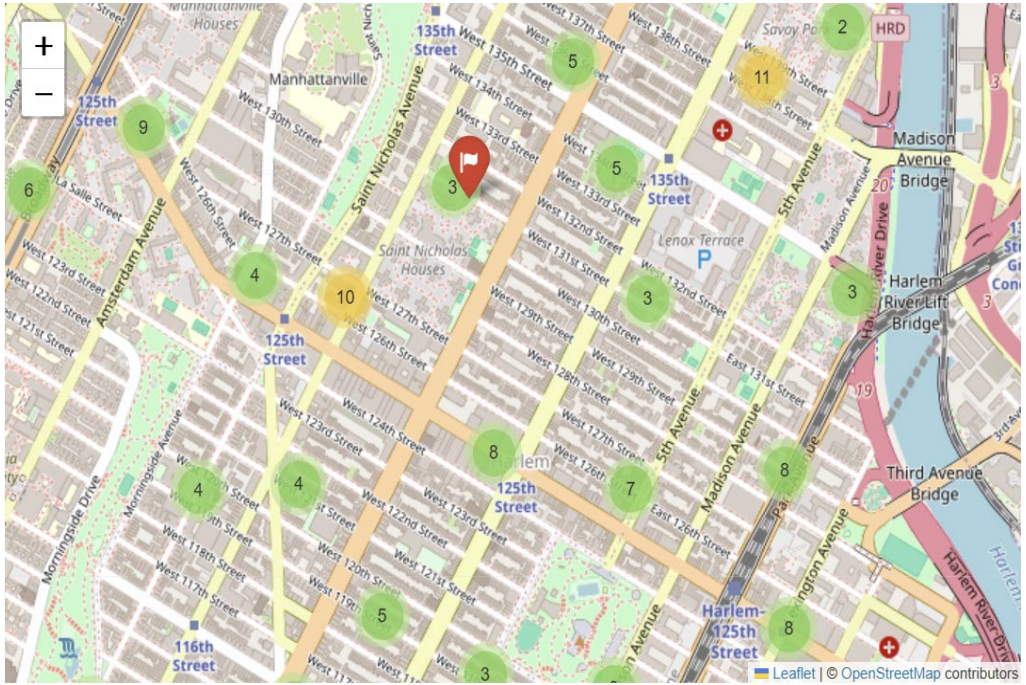
Figure shows predicted heatmap for 14-05-2024
Heatmap changes each day based on the data the model receives.

Hashim Raza

Model Backtesting

Harlem 2024 Case Study

- Human Trafficking case in Harlem, Sept 2024 - *Sex trafficking of a child*
- Model Flagged this location on Sept 4th and Sept 6th 2024 - 4 and 6 days before incident
- Risk scores 0.81 and 0.76 assigned to those days respectively



Red marker indicates actual location of HT incident. Green marker next to it represents model's prediction before the incident.

http://www.brooklynda.org/2024/02/28/brooklyn -husband-and-wife-indicted-for-sex-trafficking-of-a-child/

Hashim Raza

Dataset Used for Unsupervised Learning Models

For the **KMeans and HDBSCAN**, we used raw spatial data instead of grid-based inputs to capture natural groupings.

The data sources included:

- **NYC crime data**
- **NYC permitted events**
- **JFK inbound flight records**
- **Human trafficking reports**

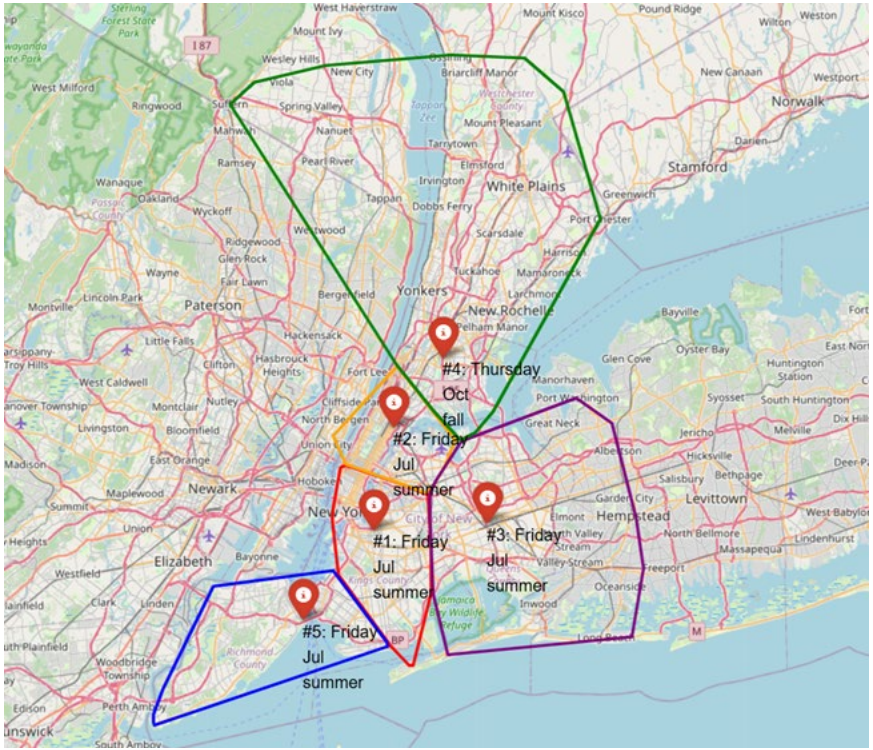
30,691 observations

15 features

Worked on by whole team

Top 5 HT Hotspots – KMeans Clustering

Identification of persistent Human Trafficking Zones



- **Clustering & Risk Mapping**
  - Used **KMeans** on location data to identify the top 5 high-crime zones.
- **Weekday Trends**
  - All clusters show more activity on **weekdays** than weekends.
- **Seasonal Peaks**
  - Summer, especially **July**, is the peak period in most clusters.
- **Actionable Insight**
  - These trends can **help focus law enforcement** and outreach efforts on **high-risk times and areas**.

Derya Kirca

Clustering Quality Metric KMeans

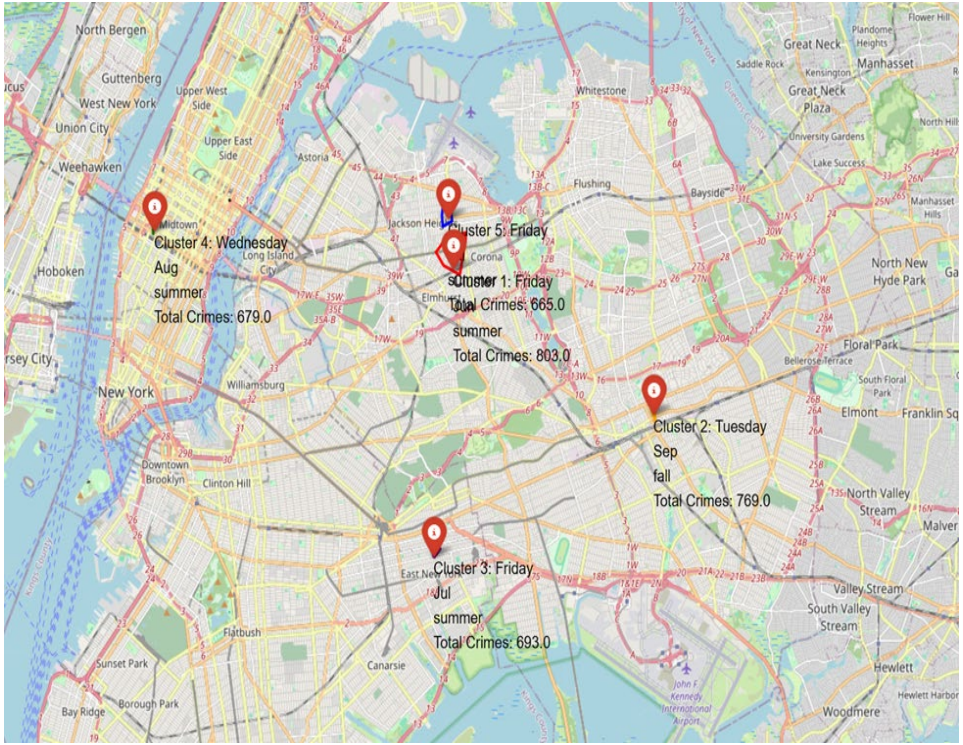
Average Silhouette Score for Top 5 Clusters with Interpretation:

Cluster 0: 0.4147 🟡 **Moderate**
Cluster 1: 0.3796 🟡 **Moderate**
Cluster 2: 0.4062 🟡 **Moderate**
Cluster 3: 0.3848 🟡 **Moderate**
Cluster 4: 0.2906 🟡 **Moderate**

Silhouette Score Range	Interpretation
>0.5	**Strong** - clusters are well separated and dense
0.25 - 0.5	**Moderate** - clusters are somewhat well defined
0.0 - 0.25	**Weak** - possible overlap or uneven density
<0.0	**Poor** - Likely misclassified points

Top 5 HT Hotspots – Precision Mapping with HDBSCAN

Time-Patterned Hotspots



- **HDBSCAN** – helps **filter noise and avoid misleading groupings** from less meaningful or isolated incidents (unlike KMeans which forces all points into clusters).
- Only high-density **clusters with 15 or more crimes** are shown.
- Each cluster is labeled with: Most common day (e.g. Friday), Top month, Season, Total crime count
- Provides **localized insight** into **when and where** crimes tend to occur.

Clustering Quality Metric HDBSCAN

Average Silhouette Score for Top 5 Clusters with Interpretation:

Cluster 1: 0.9183 🟢 **Strong**
Cluster 2: 0.9994 🟢 **Strong**
Cluster 3: 0.9994 🟢 **Strong**
Cluster 4: 0.9999 🟢 **Strong**
Cluster 5: 0.9931 🟢 **Strong**

Silhouette Score Range	Interpretation
>0.5	**Strong** - clusters are well separated and dense
0.25 - 0.5	**Moderate** - clusters are somewhat well defined
0.0 - 0.25	**Weak** - possible overlap or uneven density
<0.0	**Poor** - Likely misclassified points

Spatial Validation with Unsupervised Learning

Do LSTM-Predicted Risks Match Real Hotspots?

- **What we did:**
  - Overlaid heatmap from **LSTM** prediction
  - Applied **KMeans** to detect persistent hotspots
- **What we found:**
  - **Strong overlap between predicted risks and known hotspots**
  - Weekday peaks and summer surges matched model predictions

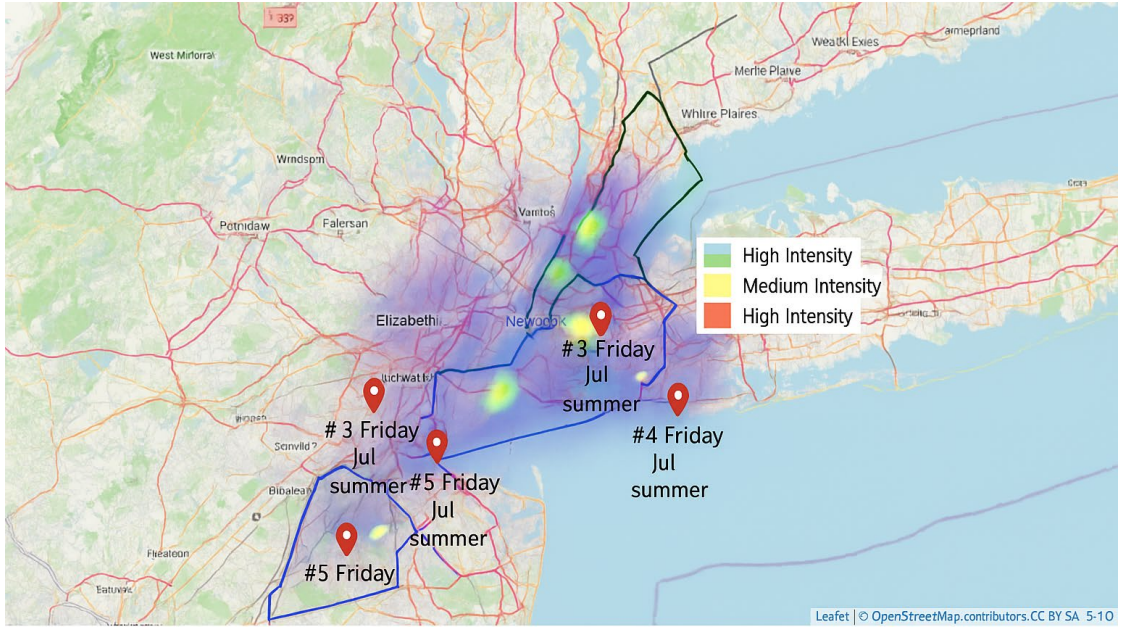


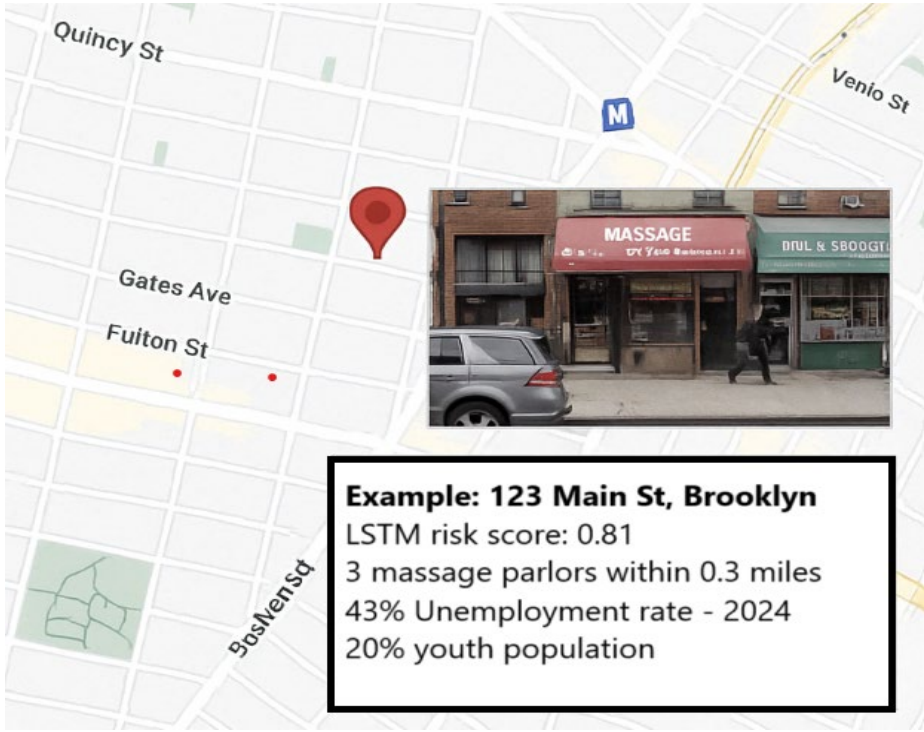
Figure shows predicted heatmap via LSTM prediction for May 14, 2024. Overlaid are the clusters identified via KMeans clustering with most common day, month, and season.

Worked on by whole team

Contextual Validation

Understanding the Socio-Economic Factors Behind the Risk

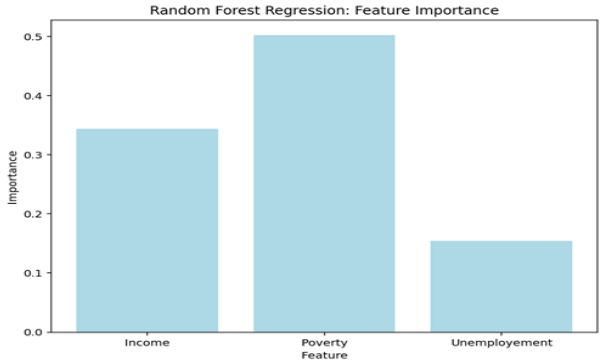
- **What we did:**
  - Reviewed **LSTM-Predicted Hotspot**
  - Analyzed **Socioeconomic and environmental** factors
- **Key Observations:**
  - **High poverty** and **unemployment** rates
  - **High-risk businesses** (e.g., massage parlors, motels)
  - **Mixed-use zoning** (residential and commercial
  - **Limited** community oversight
  - Demographic vulnerabilities present



https://data.census.gov/table/ACSDT1Y2023.B23025?q=B23025

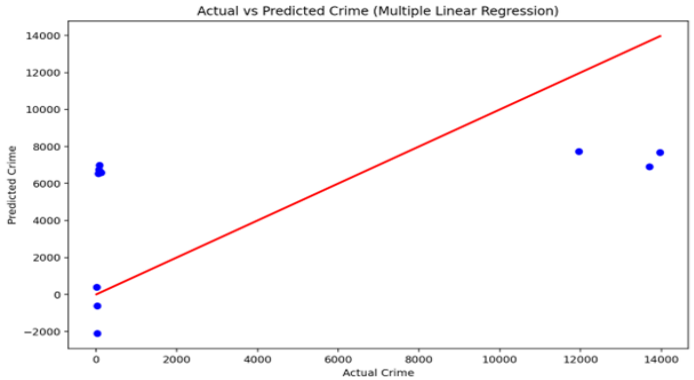
Worked on by whole team

Socio-Economic Data – What Didn't Work



	Feature	Importance
0	Income	0.343683
1	Poverty	0.502359
2	Unemployement	0.153957

R-squared on Test Data: 0.9967310828327505
Mean Squared Error on Test Data: 73411.1261027232

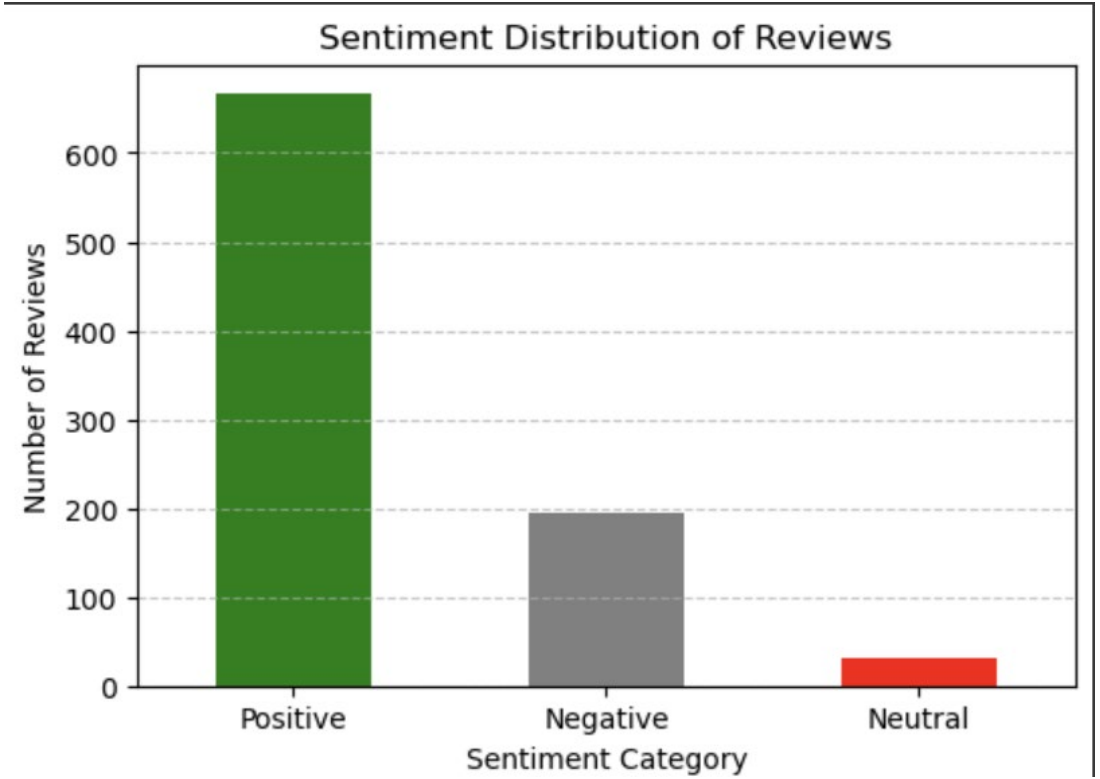


- Collected borough-level data for **Brooklyn, Queens, Bronx, Staten Island, and Manhattan**
  - Timeframe: from 2021–2024
  - **Income**, **Poverty**, **Unemployment**
- Yearly income and poverty data were **converted to monthly estimates**
- Models such as **Random Forest**, **Multiple Linear Regression**, and **ARIMAX** were tested, but socio-economic features showed **minimal predictive value**.
- Estimates were **too simplified** and lacked the **temporal detail** needed for daily prediction.

Derya Kirca

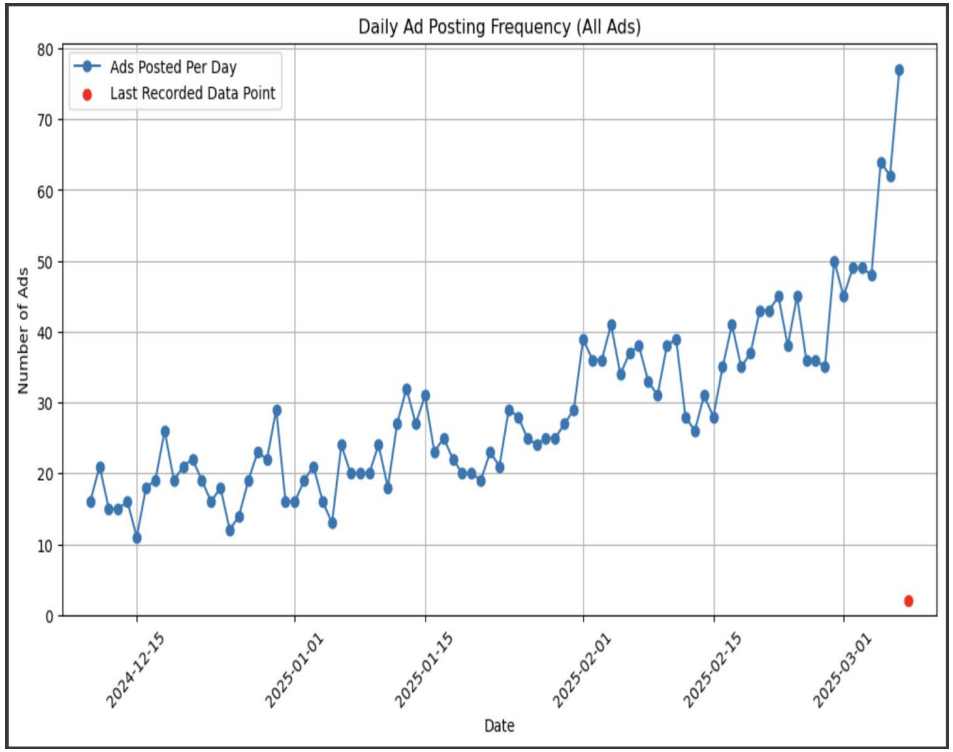
Sentiment Analysis of High-Risk Businesses

- *Google Places API* - Scraped high-risk establishments in a 30-mile radius around JFK.
- **Total Observations** : 180
- **Total Reviews:** 893
- Types of **businesses include hotels, massage parlors, nail salons & strip clubs** .
- Important for **contextual validation of high -risk locations**



Data Scraping

Backpage Illicit advertisement scrape to explore Ad frequency



- **Source** : *escortalligator* website via beautifulsoup
- **Observations** : 2551
- **Date Range** : 12/10/2024 - 03/08/2025 ~ 3-months
- Primary locations identified as **Manhattan, Brooklyn, and Queens**
- Increase advertisement **activity in Spring** . Corresponds with **increased crime activity in Spring** .

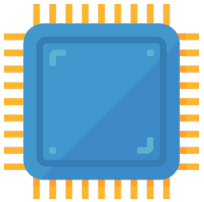
Key Issue : Not enough data to add as feature for modeling.

Worked on by whole team

Key Learnings

Understanding the Problem

Learned that **spatial and temporal context**, and iterative modeling are key in **forecasting rare, high-impact events** like human-trafficking.



Model Selection

Gained hands-on experience with **LSTM,** and **clustering algorithms (KMeans and HDBSCAN)** for spatiotemporal modeling using large-scale (**7.8M+**) **time series data.**

Data Storytelling

Developed a stronger ability to translate **technical insights into actionable strategies** that support public sector interventions.

Worked on by whole team

Recommendations for future work

1. Continue collecting online ads from Backpage-style sites to enhance model pattern recognition.

2. Acquire granular inbound flight data at JFK to strengthen model features.

3. Collect taxi and rideshare route data from JFK to predicted high-risk zones to identify potential trafficking routes.

4. Maintain access to Google Maps, Google Places, and Yelp APIs to enable live contextual validation of high-risk locations.

Worked on by whole team

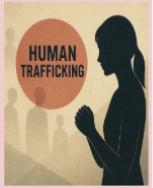
Predicting the Invisible: Risk-Based Modeling of Human Trafficking at JFK Airport & Surrounding Communities

Introduction

50 million people are living in modern slavery globally

Trafficking cases go **unreported or misclassified**, as victims often do not disclose their exploitation

Traffickers target individuals vulnerable due to **poverty, limited education, or unstable living conditions.**

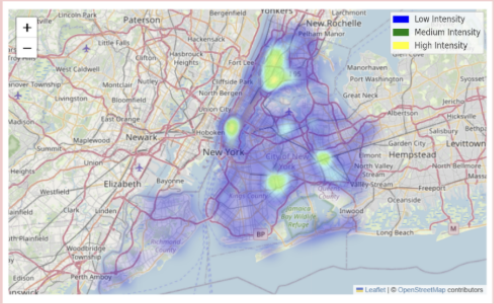


Results

Achieved AUC of 0,77 and 75% recall predicting next-day Human Trafficking risk based on locations.

Using HDBSCAN, determined precise cluster boundaries, with clearer view of when and where risks are highest.

Incorporated location context to spotlight intervention zones and inform policy decisions



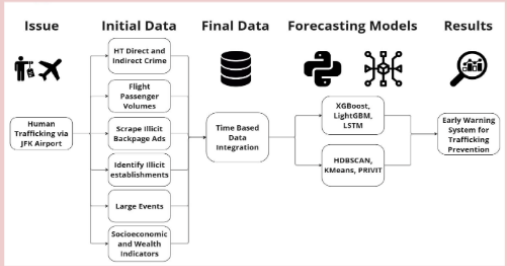
Problem Statement

Despite access to tech, data, and extensive screening there's **no integrated system to predict** trafficking risks.

Goal

Build **deep learning models** that can **identify trafficking hotspots** and be integrated into a real-time **detection system.**

Approach



Key Learnings

Takeaway: Learned the value of iterative modeling, adapting data quality and feedback

Technical: Gained hands-on experience with Deep Learning and Clustering Algorithms

Business: Developed a stronger understanding of how data science can drive public sector interventions.

THANK YOU!

References

1. Esquivel, N., Nicolis, O., Peralta, B., & Mateu, J. (2020). Spatio-temporal prediction of Baltimore crime events using CLSTM neural networks. *IEEE Access*, 8, Article 9251302. https://doi.org/10.1109/ACCESS.2020.3036715
2. Reddi, T., Kusuma, C., & Parvin, S. (2024). *Mapping crime dynamics: Integrating textual, spatial, and temporal perspectives*. Proceedings of the 15th Annual IEEE Ubiquitous Computing, Electronics & Mobile Communication Conference (UEMCON). IEEE. https://ieeexplore.ieee.org/document/10754762
3. Maass, K. L., & Konrad, R. (2022). *Operations Research and Analytics to Combat Human Trafficking: A Review of Methodologies and Applications*. Journal of Operations Research, 58(3), 215-230. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9423650/
4. Li, Y., et al. (2017). *Semi-Supervised Learning for Detecting Human Trafficking in Online Classified Ads.* Security Informatics. https://security-informatics.springeropen.com/articles/10.1186/s13388-017-0029-8
5. Wang, L., et al. (2019). *Sex Trafficking Detection with Ordinal Regression Neural Networks.* ArXiv. https://arxiv.org/abs/1908.05434
6. Cuenca, E., et al. (2023). *Human Trafficking in Social Networks: A Review of Machine Learning Techniques.* ResearchGate. https://www.researchgate.net/publication/374480817_Human_Trafficking_in_Social_Networks_A_Review_of_Machine_Learning_Techniques