Arbitrage City: AI-Powered Real Estate Investment Engine

Mohamed Diab, Bayan Khateeb **Data Collection & Management Lab — Technion**

Goal: A Universal Investment Engine

To develop a scalable, city-agnostic pipeline that automates the discovery of Real Estate Arbitrage opportunities by bridging disconnected data sources.

The Global Problem: In every major city, the market for buying homes and the market for short-term rentals (Airbnb) operate in silos.

- Investors lack a unified view of **Cost vs. Revenue**.
- Manual analysis is slow and misses high-yield pockets.

Methodology

We propose a 4-stage automated pipeline applicable to any metropolitan area.

1. **Ingestion:** Robust scraping of local real estate listings.
2. **Enrichment:** NLP extraction of value-drivers (amenities, size).
3. **Geospatial Mapping:** Matching specific addresses to rental zones.
4. **Predictive Modeling:** Using historical rental data to forecast yield.

Case Study Implementation: NYC

To validate our framework, we deployed the engine on **New York City** due to its high data complexity and strict anti-bot protections.

Data Sources:

- **Buy Side:** Scraped *Properstar.com* (5 Boroughs).
- **Rent Side:** Official *InsideAirbnb* dataset (29k+ records).
- **Volume:** 1,755 Unique Listings processed.

System Architecture

1. Target Source (e.g. Properstar)

↓

2. Batch Scraper (Playwright + Session Rotation)

↓

3. NLP Cleaning (Feature Extraction)

↓

4. Geo-Matching (ArcGIS Address → Lat/Lon)

↓

5. ML Valuation Model (Random Forest Regressor)

Scalability Feature: The scraper uses dynamic session rotation to bypass generic WAFs, making it adaptable to different real estate portals.

Proof of Concept: NYC Map

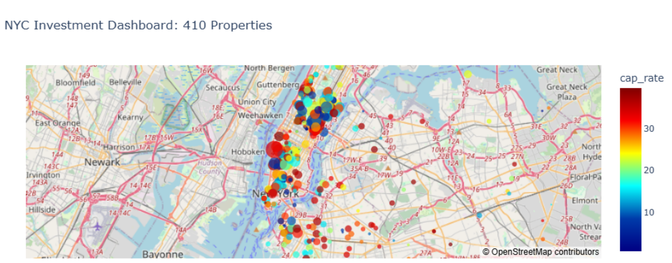

Figure 1: The engine applied to NYC data. Red/Orange areas indicate arbitrage opportunities (¿10% Yield).

Case Study Results (NYC)

Deploying the model on New York City revealed significant market inefficiencies.

Model Accuracy:

- **Algorithm:** Random Forest Regressor (100 Trees).
- **Performance:** MAE of ≈$1,200 on annual revenue.

Findings:

Metric	Value	Outcome
Total Ingested	1,755	Raw Data
Geo-Matched	1,755	100% Coverage
High-Yield Ops	**451**	**Arbitrage Found**
Avg Top ROI	22%	Outperforms Market

Conclusions & Future Work

- **Generalizability:** The pipeline successfully automated the ”Blind Spot” problem. While tested on NYC, the logic holds for any city with Airbnb presence.
- **Strategic Insight:** In our case study, outer boroughs (Brooklyn/Queens) offered 2x the ROI of Manhattan due to lower entry costs.
- **Next Steps:** Integrate Computer Vision (CNN) to score listing photos for ”Luxury” vibes to improve global price predictions.

Tools: Databricks, Azure, Python, Plotly