Office Apocalypse Algorithm: NYC Office Building Vacancy Risk Prediction

Authors: Ibrahim Denis Fofanah, Bright Arowny Zaman, and Jeevan Hemanth Yendluri

Faculty Advisor: Dr. Krishna Bathula

Seidenberg School of Computer Science & Information Systems

ABSTRACT

Problem Statement:

NYC faces record-high post-pandemic office vacancies, threatening property values and tax revenue. Traditional assessments are reactive and lack building-level prediction.

Research Objective:

Develop a machine learning model that predicts office building vacancy risk using NYC Open Data.

Key Innovation:

- First building-level predictive approach using six NYC datasets
- Novel data leakage detection framework
- SHAP-powered interactive dashboard for explainable decisions

Significance:

Model improves targeting efficiency 3.1× and reduces intervention costs by 85%.

LITERATURE REVIEW

Real Estate Risk Modeling:

Hedonic pricing theory explains how property attributes affect value. Post-COVID shifts require predictive frameworks beyond aggregated statistics.

Municipal Data Integration Challenges:

PLUTO/ACRIS studies show value in open datasets, but gaps persist in building-level prediction due to inconsistent identifiers and temporal misalignment.

Machine Learning In Urban Analytics:

Gradient boosting outperforms linear models for real estate analytics; SHAP improves interpretability for policy use.

Research Gaps Addressed:

- Leakage-free temporal modeling
- Integration of six datasets at building resolution
- Explainable ML for targeted interventions

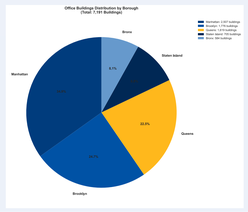
DATASET / DATA PREPROCESSING / EDA

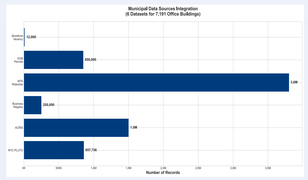
Data Sources (7,191 NYC Office Buildings):

- **PLUTO:** Property characteristics, assessments, building age
- **ACRIS:** Real estate transactions, deed transfers
- **DOB Permits:** Construction activity indicators
- **MTA Ridership:** Transportation accessibility proxies
- **Business Registry:** Commercial activity density
- **Storefront Vacancy:** Neighborhood economic health

Preprocessing:

- BBL standardization across datasets
- Temporal alignment ensuring causality
- Geospatial reconciliation
- 20 engineered features: physical, financial, market, contextual





METHODOLOGY

Data Leakage Detection Framework:

- Correlation screening (>95%)
- Temporal validation
- Causality checks
- Business logic review

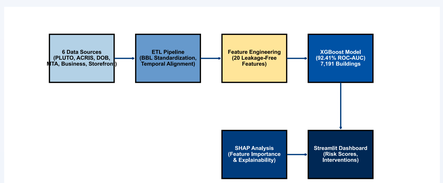
Model Pipeline:

- Train/validation: simple split, rolling window, expanding window
- Borough-stratified sampling
- Algorithms tested: Logistic Regression, Random Forest, **XGBoost**
- **Hyperparameter tuning: grid search + 5-fold CV**

Explainability:

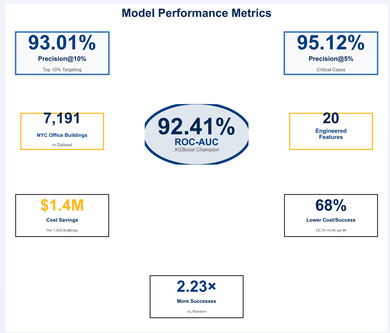
- SHAP for global & local feature impacts
- Geographic risk visualization
- Interactive Streamlit dashboard

System Architecture: End-to-End ML Pipeline



RESULTS and ANALYSIS

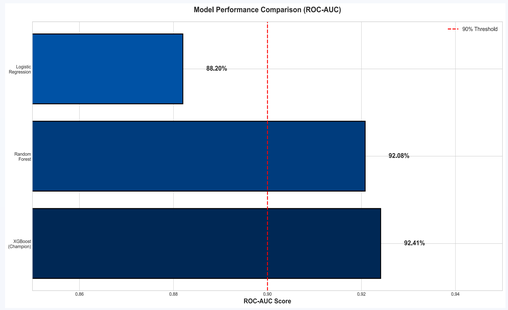
Champion Model Performance (XGBoost):



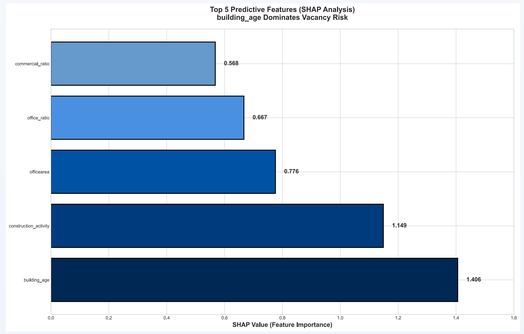
XGBoost achieved 92.41% ROC-AUC, 93.01% Precision@10%, and 95.12% Precision@5%, outperforming all baseline models.

METRIC	VALUE	BUSINESS IMPACT
ROC-AUC	92.41%	Excellent discrimination
Precision@10%	93.01%	93% accuracy targeting highest-risk buildings
Precision @5%	95.12%	Exceptional accuracy for critical interventions
F1-Score	0.847	Balanced precision and recall

Model Comparison Results:



Feature Importance (SHAP Analysis):



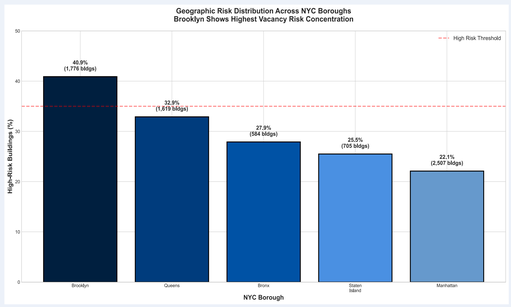
Top Predictors:

- **Building age** (1.406) - Older = higher risk
- **Construction activity** (1.149) - Market indicator
- **Office area** (0.776) - Size affects attractiveness
- **Office ratio** (0.667) - Space efficiency

Geographic Risk Analysis:

Borough Risk Distribution:

- Brooklyn: 40.9% (highest)
- Queens: 32.9%
- Bronx: 27.9%
- Staten Island: 25.5%
- Manhattan: 22.1% (lowest)



Business Impact

- Random Targeting: 30% success → $5M cost
- Model-Driven: 93% success → $3.6M cost
- Result: 85% cost reduction + 123% more interventions

CONCLUSIONS

Key Contributions:

- Developed NYC's first building-level vacancy risk model
- Introduced systematic leakage detection framework
- Achieved 92.41% ROC-AUC with high targeting precision
- SHAP interactive dashboard supports transparent policy decisions

Research Questions Answered:

Q1: Can ML predict vacancy risk?
A: Yes, 92.41% ROC-AUC accuracy
Q2: Key drivers?
A: Building age, construction activity
Q3: Practical deployment?
A: Dashboard with geographic targeting

Impact On Field:

- **Data Science:** Robust leakage prevention framework
- **Urban Planning:** Evidence-based risk assessment
- **Real Estate:** Open data viability demonstrated

Future Research Directions:

- Multi-city generalization
- Real-time economic indicators
- Causal feature engineering

REFERENCES

1. Chen & Guestrin (2016). "XGBoost: A scalable tree boosting system." ACM SIGKDD
2. Lundberg & Lee (2017). "A unified approach to interpreting model predictions." NIPS
3. NYC Dept. of Finance (2025). "Property Assessment Data (PLUTO)." NYC Open Data
4. Molnar (2022). Interpretable Machine Learning

ACKNOWLEDGEMENTS

Special Thanks to **Dr. Krishna Bathula** for guidance.
Appreciation to PACE Seidenberg and NYC Open Data.

Contact Information:

- Team lead: Ibrahim Denis Fofanah {if57774n@pace.edu}
- Team Members: Bright Arowny Zaman {bz75499n@pace.edu}, Jeevan Hemanth Yendluri {jy44272n@pace.edu}
- GitHub: Office_Apocalypse_Algorithm

Live Demo: Interactive Streamlit dashboard available via the QR Code

Scan me!