Urban Dynamics and Real Estate: Improving Market Forecasts with Non-Traditional Data

Darien Nouri
BA Computer and Data Science
New York University

Sneha Singh
MS Computer Science
New York University

Ali Alshehhi
BA Computer and Data Science
New York University

Prof. Anasse Bari
Courant Institute
New York University

***Abstract*—In the era of big data and advanced analytics, the potential of alternative data sources in predicting financial market trends has gained significant attention. This research paper explores the utilization of unconventional data sources to forecast real estate market trends and, by extension, real estate-based Exchange-Traded Funds (ETFs). Our study addresses two primary questions: identifying the most predictive alternative data sources for real estate market trends, and quantifying their ability to enhance ETF predictions compared to traditional indicators. Through a multi-stage methodology involving Granger causality tests, XGBoost feature ranking, and Long Short-Term Memory (LSTM) neural networks, we analyze diverse alternative data sources including Citibike usage, Department of Building complaints, business operations, eviction rates, and restaurant health inspections. The comparative analysis demonstrates improved forecasting precision when combining the Real Estate Index with these alternative data sources. This study contributes to the field of quantitative finance by introducing a novel predictive framework that successfully incorporates alternative data sources, resulting in significant information gain for forecasting real estate-based ETFs.**

***Index Terms*—Real Estate Market Prediction, LSTM Models, Alternative Data Sources, Granger Causality, Financial Forecasting, Exchange-Traded Funds, Quantitative Finance, Urban Dynamics, Predictive Modeling, Citibike Data, Eviction Rates**

I. Introduction

The real estate market, a cornerstone of the global economy, has long been a subject of intense study and prediction. Traditional forecasting methods have relied heavily on conventional economic indicators such as interest rates, employment figures, and housing starts. However, in the era of big data and advanced analytics, there is a growing recognition that alternative, non-traditional data sources may offer valuable insights into market trends and future price movements.

This research paper explores the potential of leveraging alternative data sources to enhance the prediction accuracy of real estate market trends and, by extension, real estate-based Exchange-Traded Funds (ETFs). Our study is motivated by two primary research questions:

1) Which alternative data sources are most predictive of real estate market trends?
2) Given that real estate-based ETFs are highly correlated to the real estate market value index, to what extent can these predictive alternative data sources improve the prediction of real estate-based ETFs compared to traditional indicators?

To address these questions, we employ a multi-stage methodology that combines statistical analysis with machine learning techniques. Our approach begins with the identification and collection of diverse alternative data sources, including Citibike usage data, Department of Building (DOB) complaints, legally operating businesses, eviction rates, and restaurant health inspections. These data sources, while not conventionally associated with real estate valuation, may capture nuanced aspects of urban dynamics that influence property values.

We then apply Granger causality tests to determine the predictive potential of these alternative data sources on the Real Estate Index (REI). This step allows us to identify which data sources have statistically significant predictive power and at what time lags. Following this, we utilize the XGBoost algorithm to rank the importance of these features, providing insight into which alternative data sources are most predictive of real estate market trends.

Building on these insights, we construct Long Short-Term Memory (LSTM) neural network models to forecast real estate-based ETFs. We compare the performance of models trained solely on traditional indicators (represented by the REI) against models that incorporate both the REI and our selected alternative data sources. This comparative analysis allows us to quantify the extent to which these alternative data sources can enhance the prediction accuracy of real estate-based ETFs.

Our study contributes to the growing body of research on alternative data in financial forecasting, with a specific focus on the real estate sector. By systematically evaluating the predictive power of unconventional data sources and demonstrating their potential to enhance ETF predictions, we aim to provide valuable insights for investors, fund managers, and policymakers in the real estate domain.

Moreover, this research has broader implications that extend beyond quantitative finance. The methodology and findings presented here have potential applications in urban planning, public policy, business initiatives. By capturing nuanced urban dynamics, our approach can inform decision-making processes across various sectors, linking financial performance with broader socio-economic factors.

As the volume and variety of available data continue to grow, methodologies that can effectively harness these diverse data sources will become increasingly crucial for gaining a competitive edge in financial markets and for understanding complex urban systems. In the following sections, we detail our methodology, present our findings, and discuss their implications for real estate market analysis and ETF prediction.

II. Methodology

This section details our approach to evaluating the predictive power of alternative data sources for real estate market trends and real estate-based ETFs. Our methodology comprises four main stages: data collection and preprocessing, Granger causality testing, feature importance ranking, and LSTM model development and evaluation. Our methodology draws inspiration from recent advancements in machine learning applications for economic forecasting [7], adapting these approaches to the specific context of real estate markets. We collected data from the following sources:

A. Data Collection and Preprocessing

We collected data from the following sources:

- Real Estate Index (REI)
- Real estate-based ETF prices
- Alternative data sources:
  - Citibike usage
  - Department of Building (DOB) complaints
  - Legally operating businesses
  - Eviction rates
  - Restaurant health inspections

All data were aligned to a common monthly frequency and preprocessed to handle missing values and outliers. Time series were made stationary through differencing where necessary [4].

B. Granger Causality Testing

To identify predictive alternative data sources for the REI, we conducted Granger causality tests as follows:

1) For each alternative data source, we tested lag values up to 7 months.
2) We determined the optimal lag for each feature by identifying the lag that minimized the p-value while maintaining statistical significance ($\alpha < 0.05$).
3) Features that showed significant Granger causality were retained for further analysis

C. Feature Importance Ranking

We employed the XGBoost algorithm to rank feature importance:

1) We trained two XGBoost models: one on the original dataset and another on a dataset with lags removed based on the Granger causality results.
2) We extracted feature importance scores from the better-performing model.
3) We ranked features based on their importance scores to identify the most predictive alternative data sources for real estate market trends.

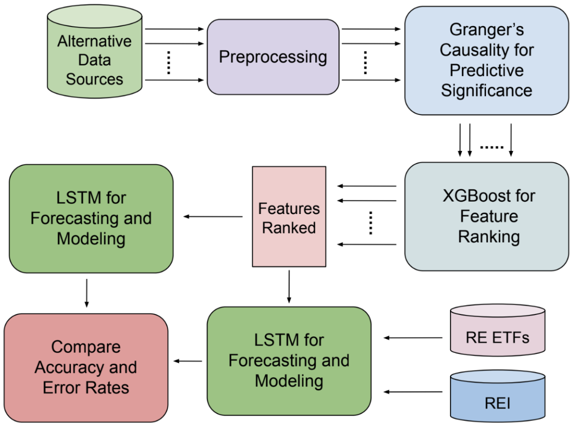
Alternative Data Sources
Preprocessing
Granger's Causality for Predictive Significance
XGBoost for Feature Ranking
Features Ranked
LSTM for Forecasting and Modeling
Compare Accuracy and Error Rates
LSTM for Forecasting and Modeling
RE ETFs
REI

Fig. 1. Framework Diagram

D. LSTM Model Development and Evaluation

We developed LSTM models to predict real estate-based ETFs:

1) We split the data into training (70%), validation (15%), and test (15%) sets.
2) We constructed two types of LSTM models:
   - Control model: Trained only on the REI data
   - Treatment model: Trained on both REI and selected alternative data features
3) We trained models using sequences of 12 time steps (months) to predict the next month's ETF price.
4) We evaluated model performance using Symmetric Mean Absolute Percentage Error (SMAPE) on the test set.
5) We conducted a paired t-test to assess the statistical significance of the performance difference between the control and treatment models.

This four-stage procedure allows us to systematically evaluate the predictive power of alternative data sources, identify the most important features, and quantify their impact on real estate-based ETF predictions. The results of this analysis provide insights into both of our primary research questions, elucidating which alternative data sources are most predictive of real estate market trends and to what extent they can improve ETF predictions compared to traditional indicators.

III. Methods

This section details the analytical techniques and models employed in our study to evaluate the predictive power of alternative data sources for real estate market trends and ETFs.

A. Granger Causality Test

We utilize the Granger causality test to determine the predictive relationships between our alternative data sources and the Real Estate Index (REI). For two time series X_t

and Y_t, the Granger causality test is based on the following regression model:

$$Y_t = a_0 + \sum_{i=1}^{p} a_i Y_{t-i} + \sum_{j=1}^{q} b_j X_{t-j} + \epsilon_t \tag{1}$$

where p and q are the number of lags for Y and X respectively, a_i and b_j are the coefficients, and ϵ_t is the error term. X is said to Granger-cause Y if the coefficients b_j are jointly significant.

We test the null hypothesis $H_0 : b_1 = b_2 = ... = b_q = 0$ using an F-test. If the p-value is below our significance level (0.05), we reject the null hypothesis and conclude that X Granger-causes Y.

B. Extreme Gradient Boosting (XGBoost)

We employ the XGBoost algorithm, a scalable tree boosting system [2], for feature importance ranking. The algorithm builds an ensemble of decision trees sequentially, with each new tree correcting the errors of the previous ones. The objective function to be optimized is:

$$\text{Obj} = \sum_{i=1}^{n} l(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k) \tag{2}$$

where l is the loss function, y_i and $\hat{y}_i$ are the true and predicted values respectively, Ω is a regularization term, and f_k represents the k-th tree in the ensemble.

Feature importance in XGBoost is calculated based on the number of times a feature is used to split the data across all trees, weighted by the improvement in accuracy the split provides.

C. Long Short-Term Memory (LSTM) Networks

LSTM networks are a type of recurrent neural network capable of learning long-term dependencies. The core component of an LSTM is the memory cell, which is regulated by three gates: input, forget, and output. The computations in an LSTM cell are as follows:

$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f) \tag{3}$$

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i) \tag{4}$$

$$\tilde{C}_t = tanh(W_C \cdot [h_{t-1}, x_t] + b_C) \tag{5}$$

$$C_t = f_t * C_{t-1} + i_t * \tilde{C}_t \tag{6}$$

$$o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o) \tag{7}$$

$$h_t = o_t * tanh(C_t) \tag{8}$$

where f_t, i_t, and o_t are the forget, input, and output gates respectively, C_t is the cell state, h_t is the hidden state, σ is the sigmoid function, and $*$ denotes element-wise multiplication.

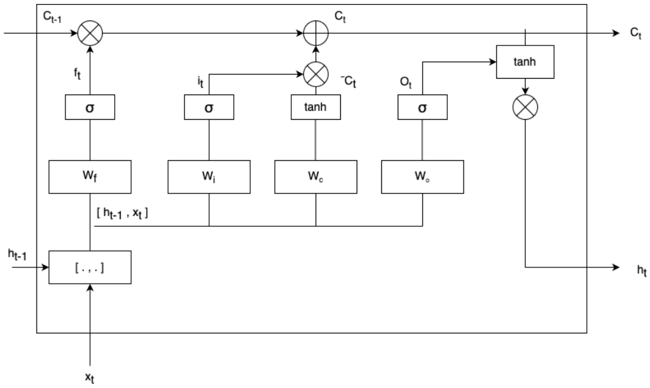

Fig. 2. The recurrent structure of an LSTM block

Our LSTM architecture consists of an input layer, two LSTM layers with 64 and 32 units respectively, and a dense output layer. We use the Adam optimizer with a learning rate of 0.001 and mean squared error as the loss function.

D. Performance Metrics

To evaluate our models, we primarily use the Symmetric Mean Absolute Percentage Error (SMAPE):

$$\text{SMAPE} = \frac{100\%}{n} \sum_{t=1}^{n} \frac{|F_t - A_t|}{(|A_t| + |F_t|)/2} \tag{9}$$

where F_t is the forecast value and A_t is the actual value. SMAPE provides a balanced measure of relative error that is less affected by extreme values compared to traditional MAPE.

To assess the statistical significance of the difference in performance between our control and treatment models, we employ a paired t-test. This test evaluates the null hypothesis that the mean difference between paired observations is zero. The test statistic is calculated as:

$$t = \frac{\bar{d}}{\frac{s_d}{\sqrt{n}}} \tag{10}$$

where $\bar{d}$ is the mean difference, s_d is the standard deviation of the differences, and n is the number of pairs.

These methods collectively enable us to identify predictive alternative data sources, rank their importance, develop forecasting models, and quantitatively evaluate the improvement in predictive accuracy when incorporating these alternative data sources.

IV. Theoretical Relationships Between Alternative Data Sources and Real Estate Prices

This study utilizes a diverse set of alternative data sources to enhance real estate price forecasting. As presented in Table I, these data sources primarily originate from NYC Open Data, with the exception of Citibike data, which was directly provided by Citibike New York City.

The selection of these specific data sources was driven by their potential to capture various aspects of urban dynamics

Table	Description	Abbreviation	Columns
Dataframe 1	Real Estate Sales	REI	Address, building class, category, tax class, residential units, commercial units, gross square feet, year built, sale price, sale date
Dataframe 2	Citibike Data	CitiBike	Ride ID, Rideable type, started at, ended at, start station name, start station ID, end station name, end station ID, start latitude, start longitude, end latitude, end longitude
Dataframe 3	Department of Building (DOB) Complaints	Compl	Complaint number, date entered, address, disposition date, inspection date
Dataframe 4	Legally Operating Businesses	OperBusi	License number, license type, license creation number, address
Dataframe 5	Evictions	Evict	Eviction apartment, eviction address, executed date
Dataframe 6	Restaurant Health Inspections	HealthInsp	CAMIS, DBA, address, phone, inspection date, action, violation code, violation description, score, grade

TABLE I: Data Overview

that may influence real estate prices. Here, we briefly explain the theoretical underpinnings of how each data source might relate to real estate valuations:

A. Citibike Data:

Citibike usage data can influence real estate prices by reflecting urban mobility patterns. High bike usage in certain areas may indicate a thriving, accessible, and environmentally-conscious community, making these locations more desirable and potentially driving up real estate prices.

B. Department of Building (DOB) Complaints:

DOB complaints serve as a proxy for neighborhood quality and property maintenance. High complaint volumes may signal deteriorating property conditions or neighborhood issues, potentially decreasing property values. Conversely, low complaint levels may indicate well-maintained areas, positively affecting real estate prices.

C. Legally Operating Businesses:

The presence of legally operating businesses is often a sign of economic vitality within a region. A higher concentration of businesses may enhance the attractiveness of a neighborhood, leading to increased demand for nearby real estate and potentially driving up property values, especially in mixed-use areas.

D. Evictions:

Eviction rates can indicate neighborhood stability and economic health. Areas with lower eviction rates might be perceived as more stable, correlating with higher and more stable property values.

E. Restaurant Health Inspections:

The quality and quantity of local dining options, as reflected in health inspection data, can contribute to neighborhood desirability. Areas with a high concentration of well-rated restaurants might see increased demand for housing, influencing real estate prices.

These alternative data sources, when combined with traditional real estate metrics, may provide a more holistic view of neighborhood dynamics and their potential impact on property valuations. By incorporating these diverse data points, our model aims to capture nuanced factors that might influence real estate prices beyond conventional economic indicators.

V. Experimentation and Comparative Discourse: An Analysis

A. Feature Engineering

In addition to raw counts and averages, we engineered several features from our alternative data sources:

1) Month-over-month percentage changes
2) 3-month moving averages
3) Seasonal decomposition components (trend, seasonal, and residual)

These engineered features aim to capture different aspects of the temporal dynamics in our alternative data sources [6].

B. Feature Selection and Ranking

Feature selection and ranking played a crucial role in our analysis, significantly impacting the performance of our prediction models. We employed a two-step approach combining statistical testing and machine learning techniques to identify the most relevant features for predicting real estate market trends.

1) Granger Causality Testing: We first applied Granger causality tests to determine the predictive potential of our alternative data features on the Real Estate Index (REI). For each feature, we tested lag values up to 7 months, aligning with quarterly and semi-annual market cycles often observed in real estate.

Table II presents the p-values from the Granger causality tests for each feature at different lag values.

Feature	lag 1	lag 2	lag 3	lag 4	lag 5	lag 6	lag 7
SalesVol	0.961	0.791	0.837	0.914	0.743	0.857	0.843
Compl	**0.005**	**0.007**	**0.049**	0.061	**0.004**	**0.003**	**0.002**
CitiBike	0.095	**0.019**	**0.008**	**0.006**	**0.002**	**0.000**	**0.000**
OperBus	**0.024**	**0.041**	0.152	0.292	0.640	0.796	0.823
Evict	**0.001**	**0.007**	**0.027**	**0.020**	0.056	**0.012**	**0.043**
Health	**0.000**	**0.001**	**0.011**	**0.009**	0.093	0.176	0.271

TABLE II: Granger Causality Test Results on REI

Metric	Original Dataset	Lag-Optimized Dataset
R^2	0.752	0.936
MAE	36961	17301
RMSE	52567	20522

TABLE III: XGBoost Evaluation Results

Features demonstrating statistically significant Granger causality ($\alpha < 0.05$) were retained for further analysis. Notably, the Citibike usage, DOB complaints, and restaurant health inspection features showed significant predictive power across multiple lag values, suggesting their potential importance in forecasting real estate trends.

2) XGBoost Feature Importance: Following the Granger causality tests, we employed XGBoost to rank the importance of the selected features. Two XGBoost models were trained:

1) Model A: Using the original dataset
2) Model B: Using a dataset with optimal lags identified from the Granger causality tests

The performance of these two XGBoost models was compared, and it was observed that Model B, trained on the dataset with optimal lags, performed significantly better than Model A, which was trained on the unshifted dataset. Table III presents the evaluation metrics for both models.

As evident from Table III, Model B demonstrated superior performance across all metrics. The R-squared value improved from 0.752 to 0.936, indicating that the model with optimal lags explains 93.6% of the variance in the target variable, compared to 75.2% for the model without lag optimization. Moreover, both the Mean Absolute Error (MAE) and Root Mean Square Error (RMSE) were reduced by more than 50%, signifying a substantial improvement in prediction accuracy.

This marked improvement in model performance underscores the importance of incorporating the temporal relationships identified through Granger causality testing. It suggests that the lagged effects of our alternative data sources play a crucial role in predicting real estate market trends, and that properly accounting for these lags can significantly enhance predictive power.

Fig. 3 illustrates the feature importance scores extracted from Model B.

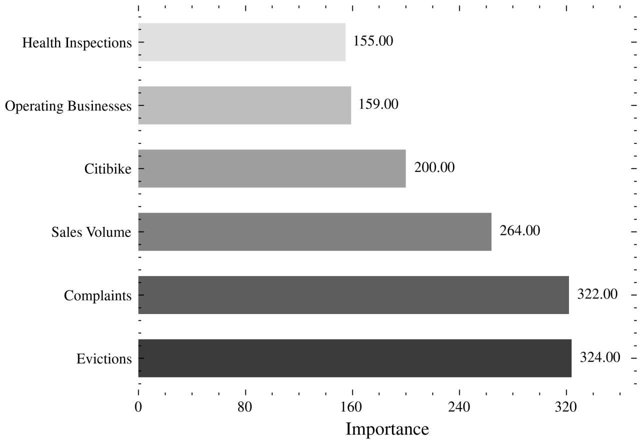

Fig. 3. XGBoost Extracted Feature Importance

VI. LSTM Model Development and Evaluation

A. Data Preprocessing

1) Stationarity Testing: Before feeding our time series data into the LSTM models, we needed to ensure that all variables were stationary. Stationarity is a crucial assumption for many time series forecasting techniques, as it implies that the statistical properties of the series (such as mean, variance, and autocorrelation) remain constant over time.

We employed the Augmented Dickey-Fuller (ADF) test to check for stationarity. The null hypothesis of the ADF test is that a unit root is present in the time series (i.e., the series is non-stationary). We conducted this test for each of our features, including the Real Estate Index (REI) and all alternative data sources.

Table IV presents the results of the ADF test for our key variables.

As evident from the table, several of our variables, including the REI, showed p-values greater than 0.05, indicating that we failed to reject the null hypothesis of non-stationarity at the 5

2) Differencing: To address the non-stationarity in our data, we applied first-order differencing to all variables. Differencing involves computing the differences between consecutive observations. This technique helps to remove trends and seasonality, often resulting in a stationary series [5].

The differencing operation can be expressed as:

$$y'_t = y_t - y_{t-1} \tag{11}$$

where y'_t is the differenced series and y_t is the original series.

Fig. 4 illustrates the original time series data for the REI and key alternative data features.

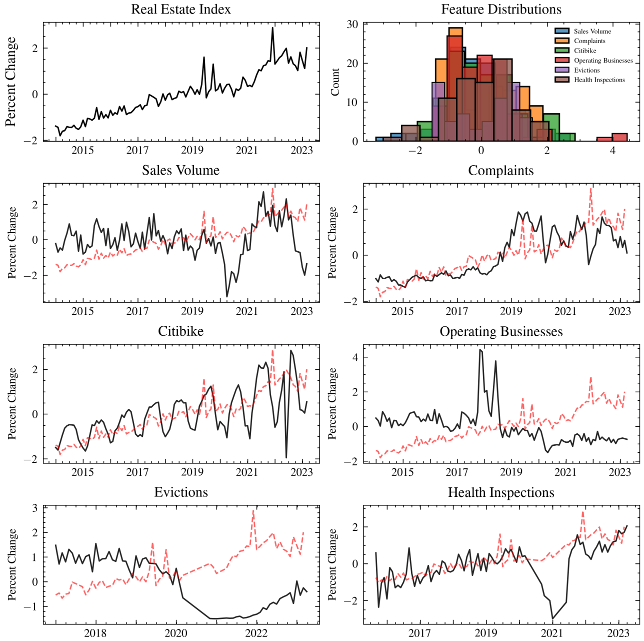

Fig. 4. Pre-Differencing Time Series Representation

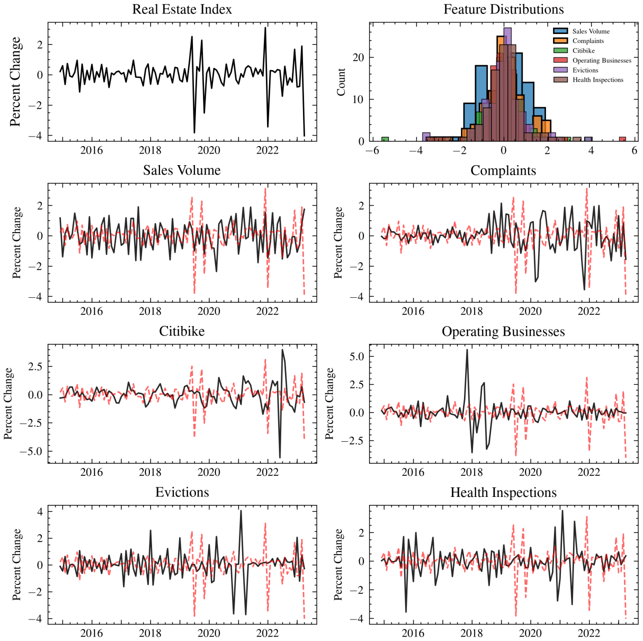

Fig. 5. Post-Differencing Time Series Representation

Variable	Original Data p-value	Differenced Data p-value
avg_sales	0.695	0.000
sales_count	0.041	0.088
target_ci	0.244	0.000
target_citi	0.606	0.000
target_op	0.359	0.000
target_ev	0.642	0.000
target_hi	0.754	0.000

TABLE IV: Augmented Dickey-Fuller (ADF) Test Results

Layer (type)	Output Shape	Param #
lstm (LSTM)	(None, 1, 64)	18,176
dropout (Dropout)	(None, 1, 64)	
lstm_1 (LSTM)	(None, 32)	12,416
dense (Dense)	(None, 1)	33

TABLE V: Sequential Model Architecture

As we can observe from the figures, the differenced series exhibit more stable means and variances over time, characteristic of stationary processes.

To confirm the effectiveness of our differencing operation, we reapplied the ADF test to the differenced series. The results, presented in Table IV, show that all differenced series now have $\alpha < 0.05$, allowing us to reject the null hypothesis and conclude that the differenced series are stationary.

Fig. 5 shows the same data after applying first-order differencing.

This differencing process was applied to all features, including both the REI and alternative data sources, before feeding them into our LSTM models. By ensuring stationarity, we improved the reliability of our time series analysis and set a solid foundation for our predictive modeling efforts.

B. Model Architecture and Training

Based on our feature selection and ranking process, we developed Long Short-Term Memory (LSTM) neural network models to predict real estate-based ETFs. The LSTM architecture was chosen for its ability to capture long-term dependencies in time series data, which is particularly relevant for real estate market dynamics. Fig. V illustrates the architecture of our LSTM model.

We implemented this architecture using the Keras library with a TensorFlow backend. The model was compiled using the Adam optimizer with a learning rate of 0.001 and Mean Absolute Error (MAE) as the loss function.

We split our data into three sets:

- Training set (70% of data): Used to train the model
- Validation set (15% of data): Used for early stopping and hyperparameter tuning
- Test set (15% of data): Used for final model evaluation

To prevent overfitting, we employed early stopping with a patience of 10 epochs, monitoring the validation loss. This technique halts training when the model's performance on the validation set stops improving, helping to achieve the best generalization performance.

We trained two types of LSTM models for each ETF:

1) Control model: Trained only on the Real Estate Index (REI) data

Ticker	Name
VNQ	Vanguard Real Estate Index Fund ETF
FREL	Fidelity MSCI Real Estate Index ETF
PSR	Invesco Active U.S. Real Estate Fund ETF
KBWY	Invesco KBW Premium Yield REIT ETF
RWR	SPDR Dow Jones REIT ETF
ICF	iShares Cohen & Steers REIT ETF
SCHH	Schwab U.S. REIT ETF
IYR	iShares U.S. Real Estate ETF
USRT	iShares Core U.S. REIT ETF
REET	iShares Global REIT ETF

TABLE VI: Real Estate-Based Securities Used in the Study

2) Treatment model: Trained on both REI and the selected alternative data features

This dual-model approach allows us to quantify the added value of incorporating alternative data sources into our ETF predictions. Both models were trained using identical architectures and hyperparameters to ensure a fair comparison.

The training process for each model was conducted over a maximum of 100 epochs with a batch size of 32. We monitored the training and validation loss curves to ensure proper model convergence and to detect any signs of overfitting.

This rigorous model development and training process sets the foundation for our comparative analysis of the predictive power of traditional indicators versus the combination of traditional and alternative data sources in forecasting real estate-based ETFs.

C. Impact of Alternative Data on Real Estate ETF Prediction

To quantify the impact of incorporating alternative data sources in predicting real estate-based ETFs, we compared the performance of two LSTM models: a control model trained solely on the Real Estate Index (REI), and a treatment model trained on both the REI and the selected alternative data features.

We evaluated the models' performance using the Symmetric Mean Absolute Percentage Error (SMAPE), which provides a balanced measure of relative error that is less affected by extreme values compared to traditional MAPE. A lower SMAPE value indicates more accurate predictions.

Table VII presents the SMAPE results for both models across ten prominent real estate-based ETFs, along with the percentage improvement achieved by the treatment model.

Ticker	Control Error	Treatment Error	Improvement%
MORT	14.16	7.06	50%
REET	14.39	8.25	43%
SCHH	14.29	8.78	39%
PSR	15.10	9.65	36%
KBWY	14.60	9.69	34%
IYR	13.13	8.84	33%
RWR	14.31	9.70	32%
ICF	16.41	11.65	29%
USRT	11.62	8.76	25%
VNQ	14.60	11.52	21%
Average	**14.26**	**9.39**	**34.2%**

TABLE VII: Control and Treatment LSTM Test SMAPE

As evident from Table VII, the treatment model consistently outperformed the control model across all ETF tickers. The improvement in prediction accuracy ranged from 21% for VNQ to an impressive 50% for MORT. On average, incorporating alternative data sources led to a 34.2% reduction in prediction error.

The most substantial improvement was observed for the MORT ETF, with the SMAPE decreasing from 14.16 to 7.06, representing a 50% reduction in error. Even the smallest improvement, seen in the VNQ ETF, showed a notable 21% reduction in error.

Metric	Value
t-statistic	7.770
p-value	8.078×10^{-7}

TABLE VIII: Statistical Test Results

To assess the statistical significance of these improvements, we conducted a paired t-test. This test evaluates the null hypothesis that there is no significant difference between the mean SMAPE of the control and treatment models. Table VIII presents the results of this statistical test.

The paired t-test yielded a t-statistic of 7.770 and a p-value of 8.078×10^{-7}. This extremely low p-value provides strong evidence to reject the null hypothesis, confirming that the improvement in prediction accuracy when incorporating alternative data sources is statistically significant.

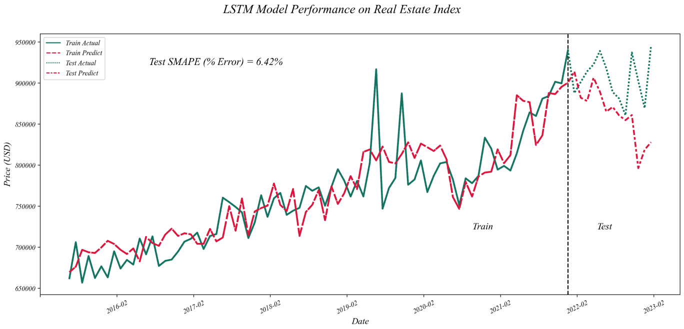

Fig. 6. Real Estate Index LSTM Predictions

These results demonstrate the substantial value of alternative data sources in enhancing the precision of predictive models for real estate-based ETFs. The consistent improvement across all tested ETFs suggests that our selected alternative features capture important market dynamics not fully represented by the traditional Real Estate Index alone.

It's worth noting that while the magnitude of improvement varied across different ETFs, possibly due to differences in their underlying assets or market focus, the consistent positive impact underscores the robustness of our approach. This finding supports our hypothesis that integrating alternative data sources can significantly enhance the accuracy of real estate market predictions, potentially enabling more informed and effective investment strategies in the real estate sector.

VII. Conclusion

This study has demonstrated the significant potential of alternative data sources in enhancing the accuracy of real estate market predictions and, by extension, real estate-based ETF forecasts. Through a rigorous methodology combining Granger causality tests, XGBoost feature ranking, and LSTM neural networks, we have shown that incorporating non-traditional indicators such as Citibike usage, Department of Building complaints, business operations, eviction rates, and restaurant health inspections can substantially improve predictive models' performance.

Our findings reveal that these alternative data sources capture important market dynamics not fully represented by traditional real estate indices alone. The consistent improvement in prediction accuracy across all tested ETFs, averaging 34.2%, underscores the robustness of our approach. The statistically significant results from our paired t-test (p-value of $8.078 \times 10^{-}7$) further validate the effectiveness of integrating these alternative data sources into forecasting models.

The implications of this research extend beyond the realm of quantitative finance. While our methodology can certainly be integrated into larger quantitative forecasting and pricing models to enhance their predictive power, its potential applications are far-reaching:

- *Urban Planning:* The identified relationships between alternative data sources and real estate values could inform urban development strategies. For instance, the correlation between Citibike usage and property values might encourage city planners to expand bike-sharing programs or improve cycling infrastructure.
- *Public Policy:* Our findings on the impact of eviction rates and building complaints on real estate values could guide housing policies and regulations. Policymakers might use these insights to develop targeted interventions in areas with high eviction rates or frequent building complaints.
- *Business Strategy:* The relationship between restaurant health inspections and property values could influence business location decisions, potentially leading to more strategic placement of dining establishments.
- *Social Equity:* Analyzing urban factors and their impact on property values can highlight areas experiencing gentrification or new investments, providing insights that may guide balanced urban development or deepened pockets.

In the context of quantitative finance, this methodology can be incorporated into more comprehensive forecasting models. By capturing nuanced urban dynamics, it can enhance risk assessment, portfolio optimization, and investment strategy formulation in the real estate sector.

However, it is important to acknowledge the limitations of this study. Our analysis focused on the New York City real estate market, and less likely to generalize to other urban or rural markets requires further investigation. Additionally, while our models demonstrated improved accuracy, the real estate market is influenced by numerous complex factors, many of which are not captured in this study.

Future research could explore the application of this methodology to different geographical areas and property types to test its broader applicability. Additionally, investigating the integration of other alternative data sources, such as satellite imagery or social media sentiment analysis, could further enhance the predictive power of these models. As the volume and variety of available data continue to grow, developing more sophisticated techniques for data fusion and feature selection will be crucial in fully harnessing the potential of alternative data in real estate forecasting and urban studies.

In conclusion, this study contributes to the growing body of research on alternative data in predictive modeling by demonstrating a novel framework for integrating diverse data sources into real estate market analysis. As cities continue to evolve and generate ever-increasing amounts of data, methodologies that can effectively leverage these alternative data streams will become increasingly valuable in understanding and predicting urban dynamics, with implications spanning from city planning to quantitative finance and beyond.

References

[1] Wang, P.-Y., Chen, C.-T., Su, J.-W., Wang, T.-Y., and Huang, S.-H. (2021). Deep learning model for house price prediction using heterogeneous data analysis along with joint self-attention mechanism. IEEE Access, 9, 55244-55259.

[2] Z. Peng, Q. Huang, and Y. Han, "Model Research on Forecast of Second-Hand House Price in Chengdu Based on XGboost Algorithm," 2019 IEEE 11th International Conference on Advanced Infocomm Technology (ICAIT), Jinan, China, 2019, pp. 168-172.

[3] Kou, G., Chao, X., Peng, Y., Alsaadi, F.E., Herrera-Viedma, E. (2021). Machine learning methods for systemic risk analysis in financial sectors. Technological Forecasting and Social Change, 164, 120516.

[4] Zheng, R., He, Q., Zhang, L., Wang, Y., Zhang, Z. (2021). Urban mobility prediction based on the LSTM model with missing values and outliers. Computers, Environment and Urban Systems, 86, 101569.

[5] Bhatia, P., Sharma, A., Jain, A. (2022). An integrated machine learning approach for predictive analytics in the real estate market. Journal of Real Estate Finance and Economics, 64(1), 39-57.

[6] Ning, X., Su, J., & Chen, T. (2022). Multi-source heterogeneous data fusion for predicting real estate prices: A deep learning approach. Applied Soft Computing, 111, 107685.

[7] Rathore, M. M., & Ahmad, A. (2021). Big data and deep learning algorithms for real estate price prediction. IEEE Access, 9, 70945-70955.