Floor Area Estimates

Table *of* Contents

Executive Summary

Measurabl's Floor Area Estimates (FAE) is a data service that provides the real estate sector with the ability to estimate gross floor area for commercial and multi-unit residential properties when actual floor areas are not available. Currently, FAE is available as part of Measurabl's Asset Level Data and Listed Data Products, where it is utilized to fill in missing floor areas prior to estimating energy usage and carbon emissions.

Methodology

To provide gross floor area estimates, Measurabl leverages machine learning (ML) and a combination of anonymized proprietary, purchased, and public data for over 150,000 properties spanning across more than 90 unique property types located in over 85 countries (Fig 1).

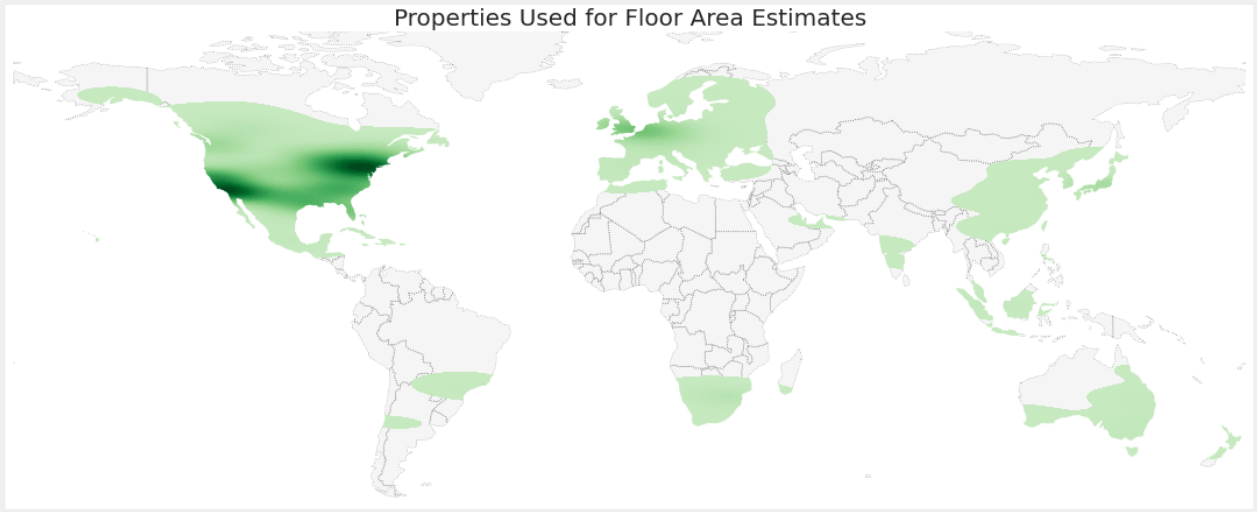

Fig 1. Heatmap visualization of all properties used to train the FAE ML model. Darker green areas indicate higher density of properties.

The ML model is trained to find patterns between floor area and other property features - a process called "training." Once trained, the model can utilize available features from new properties to produce estimated floor areas (Fig 2).

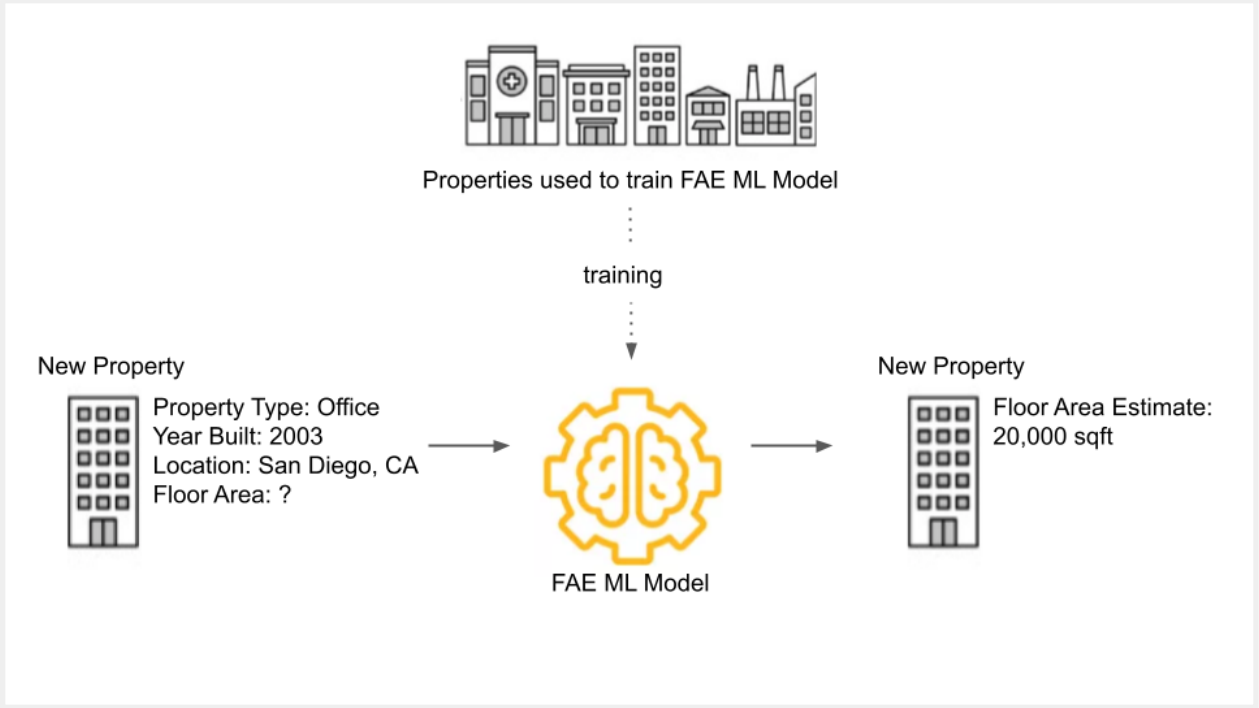

Fig 2. Diagram showcasing: (1) the process of training the FAE ML model and (2) using the FAE ML model to make floor area estimates for a new unseen property.

The set of features utilized are universally available across all property types and locations, and at the same time are helpful in explaining property floor area variance. The goal of the ML model is to provide accurate estimates for properties across a wide variety of property use types, locations, sizes, etc.

For further details on the FAE data and ML model see Appendix A.

Performance

In order to evaluate the ML model's performance, Measurabl splits the dataset into two parts: a training dataset and test dataset. The training set is used to train the model, while the test set is used exclusively to evaluate model performance. In this way, test set performance results can be extrapolated to new properties that are also unseen by the model. Measurabl uses R^2 (coefficient of determination) as an indicator of how much of the floor area variance in the test set can be explained by the model. To reduce the impact of outliers, the model learns to predict logarithmically scaled

floor area. The model achieves a test set $R^2 = 0.53$ in logarithmic space. This means that the model is able to capture 53% of the variance of the floor areas being estimated.

For further details on ML model performance see Appendix B.

Impact on Energy Usage Estimates

FAE provides value through enabling Measurabl's data products to estimate energy usage for properties where no floor area data is available. Thus, the main performance metrics of interest are based on the impact to the quality of Measurabl's energy usage estimations when using estimated floor areas. In order to assess this, Measurabl compares energy usage estimation performance for the Whole Building Estimates test dataset using actual floor areas versus estimated floor areas.

While there is a 32% degradation in energy usage estimation performance - in terms of R^2 - when using estimated floor areas, the energy estimates remain coherent. Moreover, there is no significant degradation in energy usage estimation performance based on energy usage intensity (energy usage per sq. ft.) metrics when using estimated floor areas.

For further details on computing energy usage estimation performance see the Whole Building Estimates Whitepaper - Appendix E.

Appendix A - Machine Learning Model & Data

Data for FAE

Measurabl applies a data cleaning process to maximize integrity and quality of the underlying dataset used for FAE. After cleaning, the breakdown of property counts across the most common property types are as follows:

Property Use Type	# of Properties Used for ML Model Training
All Properties	~159,000
Retail	~36,000
Multi-Unit Residential	~31,000
Warehouse/Storage	~30,000
Office	~27,000
Manufacturing/Industrial	~8,000
Hotel	~4,700
Medical Office	~4,100
Healthcare	~2,500
Food Sales & Service	~2,200
Leisure	~1,100
Data Center	~700

Table 1. Approximate counts for properties used to train the FAE ML model. Counts are shown for "All Properties" as well as for the most common high level property types.

ML Model Features for FAE

The following property information is used to derive the set of ML model features used for FAE:

- Property type
- Location
- Year built (where available)

Property type information is based on Measurabl's standardized property types. When creating ML features based on property type information, Measurabl uses both high level type information (e.g. "Retail", "Residential", "Office", etc.) as well as granular subtype information (e.g. "Residential: Multifamily Housing", "Residential: Senior Care Facility", etc.) in order to capture common trends between subtypes as well as effects specific to individual subtypes.

ML Model Framework for FAE

To estimate floor area Measurabl uses a gradient boosted decision tree framework, which is an ML ensemble method combining hundreds of classification and regression trees (CART). Each CART can be thought of as a decision tree that determines what floor area value should be expected based on multiple yes/no questions (Fig. 3).

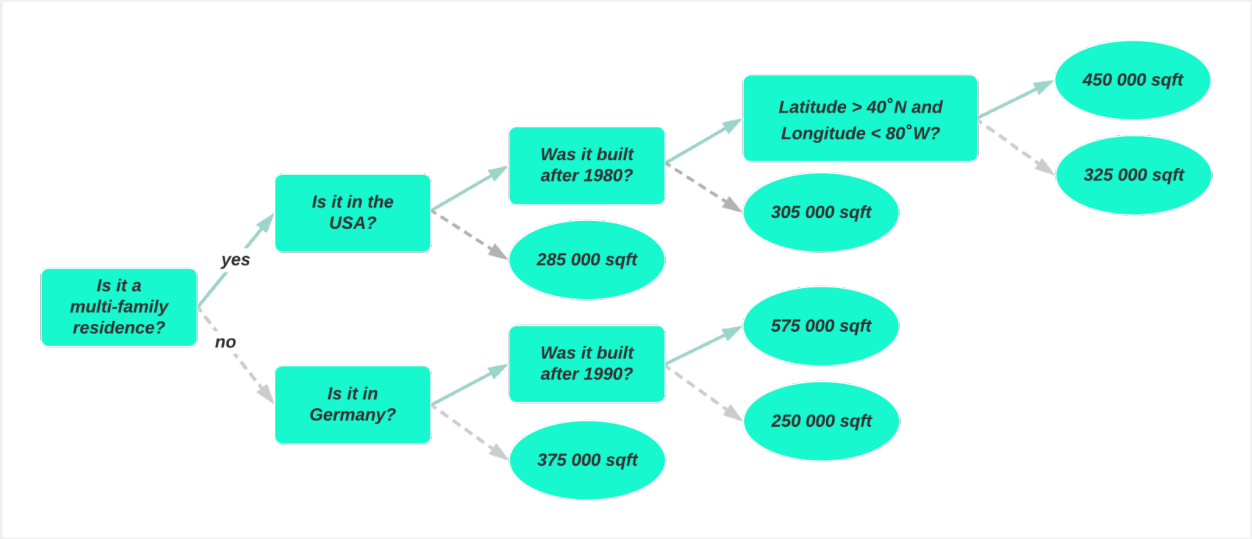

Fig 3. An example of a single classification and regression tree. The FAE ML model combines hundreds of trees to produce floor area estimates.

In certain instances where the number of units for multi-unit residential properties are available, Measurabl uses a supplementary linear regression model, instead of the FAE ML model, to estimate gross floor area based on the number of units.

As a final note, FAE model framework is optimized to produce estimates for properties with a gross floor area up to 1 million sq ft. The reason for this optimization is that FAE was developed to supplement Measurabl's data products, which are currently capable of estimating energy usage and carbon emissions for properties up to 1 million sq ft.

Appendix B - Machine Learning Model Performance

This section is intended to provide additional details regarding the FAE ML model's performance.

Table 2 showcases test set performance across the most common property types. The performance metrics presented in Table 2 are: R^2 based on logarithmic space estimations ($\log_{10}$(sq. ft.)) and median absolute percent error (MdAPE) based on linear space estimations (sq. ft.).

Property Use Type	R^2	MdAPE
All Properties	0.53	48%
Retail	0.60	55%
Multi-Unit Residential	0.53	42%
Warehouse/Storage	0.29	43%
Office	0.25	51%
Manufacturing/Industrial	0.16	52%
Hotel	0.26	39%
Medical Office	0.16	49%
Healthcare	0.55	47%
Food Sales & Service	0.74	33%
Leisure	0.43	37%
Data Center	0.19	55%

Table 2. Test set performance metrics broken down across the most common high level property types. R^2 metrics based on logarithmic space predictions ($\log_{10}$(sq. ft.)). MdAPE metrics based on linear space predictions (sq. ft.).

Fig. 4 showcases the relationship between model error and property size. Analyzing Fig. 4, it is apparent that the model's relative error changes depending on property size. Additionally, when examining the data used to create Fig 4, it was observed that the model's relative error is lowest for properties that are between ~75,000 and ~150,000 sq. ft.

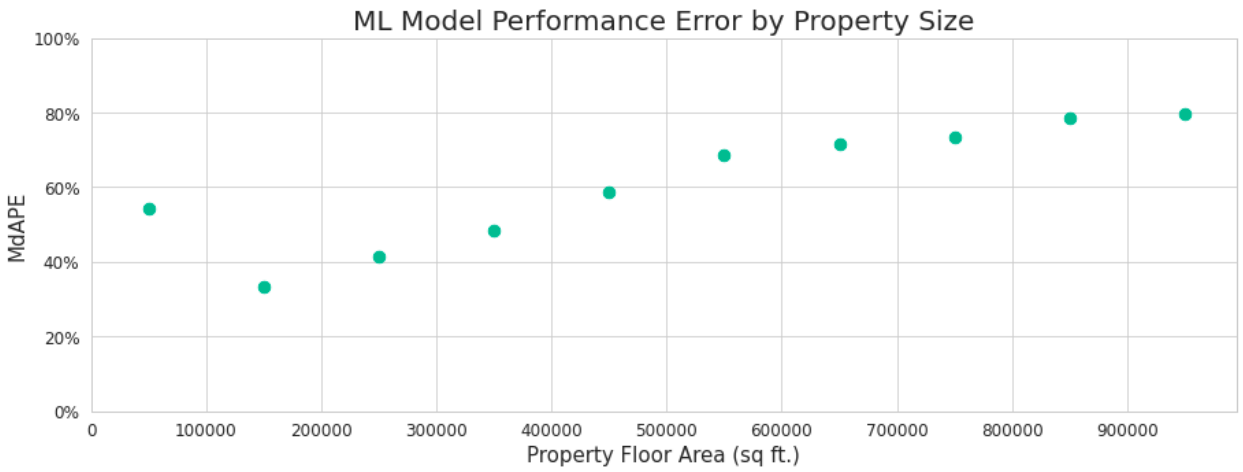

Figure 4. ML model error (MdAPE) plotted against actual property floor area for the test set.