Geography Embeddings for Predicting Prices of Self-Storage Units

Bassem Akoush, Hashem Elezabi
{bakoush, hashem}@stanford.edu

Project Mentor: Scott Fleming

Problem Setup

- Our goal is to predict prices of self-storage units by incorporating *geography embeddings.*
- Baseline data: *unit features* (e.g. square footage, elevator access, climate-controlled)
- Main questions:
  1. How well can we predict prices based on just unit features?
  2. How much can we improve by adding additional location-based features, termed "geography embeddings"?
- Our sources for geography embeddings:
  1. Census features
  2. Satellite imagery

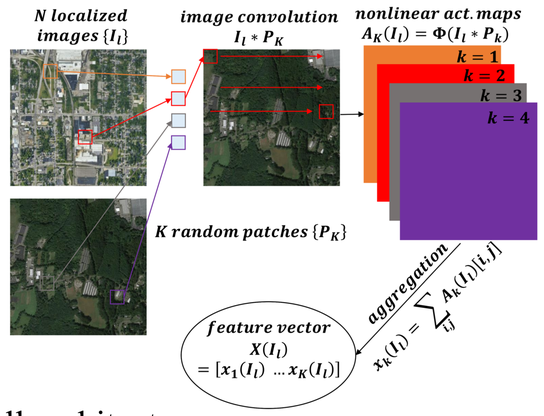
Dataset and Features

- Dataset I: 33,085 self-storage units across 2,271 facilities, each with 36 features including square ft, latitude and longitude, elevator access, etc.
- Dataset II: Dataset I augmented with census features about the facility's location. 1044 features about the county (household size, income, demographics).
- Satellite image-based features: MOSAIKS vector representations [1] derived from satellite images at all 2,271 facility locations. Satellite images retrieved via Google Maps Static API, each of which covers 1km x 1km area around the facility.

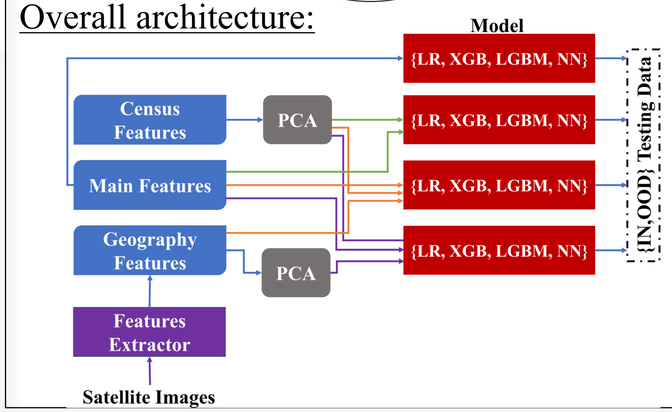
Methods

Learning algorithms: linear regression (baseline), k-nearest-neighbors, neural network, gradient-boosted decision tree (XGBoost, LightGBM).

MOSAIKS feature extraction:



Overall architecture:



Future Work

1. Explore feature extraction using street-level imagery.
2. Develop framework for data fusion to integrate multiple data sources effectively.
3. Studying impact of nearby competitors.

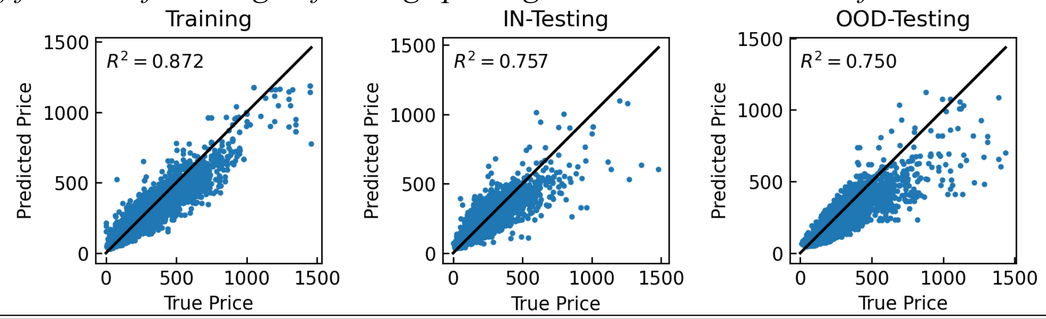
Results

Data split into Training and Testing-IN ("in-distribution") sets from two companies, and Testing-OOD ("out-of-distribution") from a third company. # {Training, Testing-IN, Testing-OOD} = {22039, 5429, 5412} examples.

Model	Training			Testing-IN		Testing-OOD	
	RMSE	R^2_{tr}	$\overline{R^2_{val}}$	RMSE	R^2	RMSE	R^2
Dataset I (Unit only)							
LR	80.211	0.633	0.618	87.248	0.585	108.852	0.58
KNN	65.134	0.759	0.714	71.160	0.70	96.363	0.681
XGB	57.176	0.814	0.755	67.868	0.749	88.676	0.721
LGBM	56.820	0.816	0.786	**66.035**	**0.762**	**88.387**	**0.723**
NN	63.021	0.774	--	72.424	0.714	90.933	0.707
Dataset II (Unit + Census)							
XGB	47.848	0.870	0.766	66.748	0.757	84.387	0.747
LGBM	47.636	0.871	0.772	**66.416**	**0.759**	**82.28**	**0.76**
NN	44.471	0.887	--	78.826	0.661	102.884	0.624
Dataset I + MOSAIKS (Unit + MOSAIKS)							
LR	72.748	0.699	0.659	81.693	0.636	103.354	0.621
XBG	25.338	0.837	0.743	70.928	0.733	**87.036**	**0.73**
LGBM	37.410	0.838	0.753	**68**	**0.742**	83.665	0.728
NN	54.211	0.833	--	83.583	0.619	89.36	0.717
Dataset II + MOSAIKS (Unit + Census + MOSAIKS)							
XBG	48.279	0.867	0.766	67.155	0.754	84.214	0.748
LGBM	47.365	0.872	0.771	**66.754**	**0.757**	**83.989**	**0.75**

- Green highlights best-performing model on each dataset, on Testing-OOD.
- **Baseline (LR) to best model (LGBM) improvement: 24.7% higher R^2.**
- **Adding census features yields ~5% higher R^2.**
- **Adding MOSAIKS features yields ~1% higher R^2.**
- Goodness of fit plots below are for LGBM on Dataset II + MOSAIKS.

Conclusion: With added location-based features and better learning algorithms, total improvement of 31% from baseline R^2=0.58 to R^2=0.76. But MOSAIKS didn't help much, likely because satellite image features aren't indicative enough of factors influencing self-storage pricing, not as much as census features.



References and Acknowledgments

[1] Rolf E, et al. A generalizable and accessible approach to machine learning with global satellite imagery. Nature Communications 2021.

We are grateful to Scott Fleming for providing datasets, and for his in-depth help and advice on this project. We would also like to thank the authors of [1] for their help on MOSAIKS.

MOSAIKS

Model	Training			Testing-IN		Testing-OOD	
	RMSE	R^2_{tr}	$\overline{R^2_{val}}$	RMSE	R^2	RMSE	R^2
Dataset I (Unit only)							
LR	80.211	0.633	0.618	87.248	0.585	108.852	0.58
KNN	65.134	0.759	0.714	71.160	0.70	96.363	0.681
XGB	57.176	0.814	0.755	67.868	0.749	88.676	0.721
LGBM	56.820	0.816	0.786	66.035	0.762	88.387	0.723
NN	63.021	0.774	--	72.424	0.714	90.933	0.707
Dataset II (Unit + Census)							
XGB	47.848	0.870	0.766	66.748	0.757	84.387	0.747
LGBM	47.636	0.871	0.772	66.416	0.759	82.28	0.76
NN	44.471	0.887	--	78.826	0.661	102.884	0.624
Dataset I + MOSAIKS (Unit + MOSAIKS)							
LR	72.748	0.699	0.659	81.693	0.636	103.354	0.621
XBG	25.338	0.837	0.743	70.928	0.733	87.036	0.73
LGBM	37.410	0.838	0.753	68	0.742	83.665	0.728
NN	54.211	0.833	--	83.583	0.619	89.36	0.717
Dataset II + MOSAIKS (Unit + Census + MOSAIKS)							
XBG	48.279	0.867	0.766	67.155	0.754	84.214	0.748
LGBM	47.365	0.872	0.771	66.754	0.757	83.989	0.75