Capstone Project

Predicting Future US Multifamily Rental Values

By: Aditya Jayasuri, Ryan Burakowski, & Gabby Klein

Presentation Layout

Objectives
Three main objectives

Data
Sources & Cleaning

Methods
Feature Engineering, PCA, & Modeling

Conclusions
Wrapping up Objectives

Future Work

Objectives

1. **Predict** the Multifamily Zillow Rental Index (ZRI) by ZIP code, three years into the future, using publicly available data sources.
2. **Determine** features that can predict future rent values in the absence of current rent prices.
3. **Explore** the importance of current rent prices on predicting future rent prices.

DATA

Sources & Cleaning

Data Sources:

- **US Census American Community Survey** (ACS) by ZIP code.[Link]
  - 33,000 ZIP codes x 250 features
- **US Census Bureau**, Businesses by ZIP Code. [Link]
  - 8 API calls yielding 303,000 rows
- **Homeland Infrastructure Foundation-Level** location data for Universities, State Govt Buildings, Hospitals, and Transit Terminals. [Link]
  - 19,500 Locations x 30 data fields
- **Multifamily ZRI** by ZIP code. [Link]
  - 1,860 ZIP codes x 113 months

Data Cleaning:

Multifamily ZRI: Significant missingness issues.

- Dropped zip codes with 50%+ missing rental values. (364/1860 ZIP codes)
- Linear interpolation for interior missing values.
- Filled leading/trailing missing values with year-ago / year-ahead data and assumed 4% annual rental growth rate.

ACS Census Data: Significant missingness issues.

- Dropped columns with over 20% missing values (7/252 features)
- Filled missing values for ZIP code attributes with city average of the same year.

Methodology

Feature Engineering, PCA & Modeling

Feature Engineering:

Modified ACS data to generate useful features with minimal duplicated information.

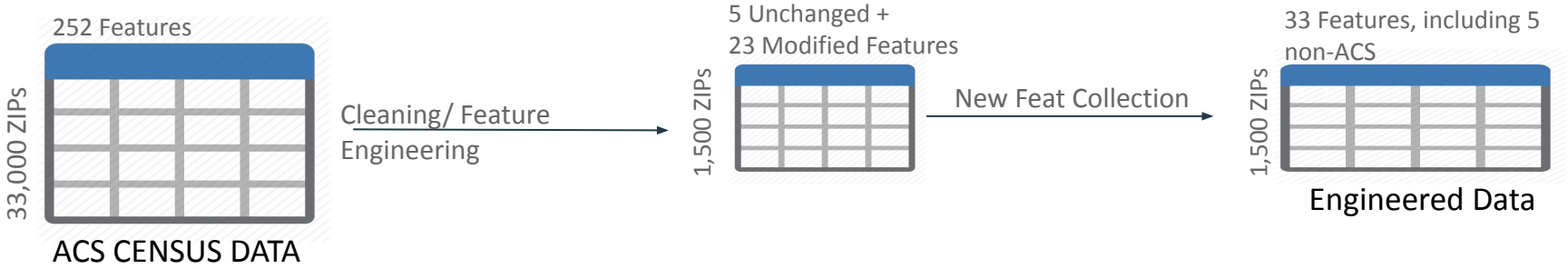
- Combined correlated features to make fewer with stronger predictive power
- Transformed features to be percentages (of total population, labor force, housing units etc) to avoid multicollinearity

Collected new data we hypothesized to be predictive at the ZIP code level (business count per ZIP code, university count, etc).

Transformed engineered features using PCA to help control multicollinearity

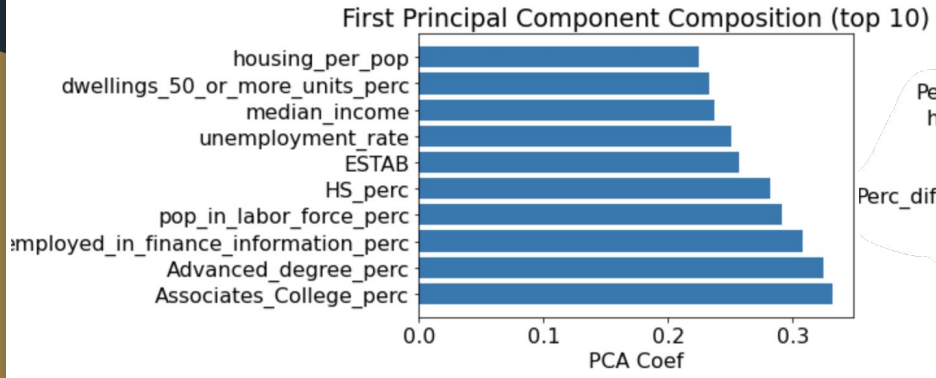
Feature Engineering Examples:

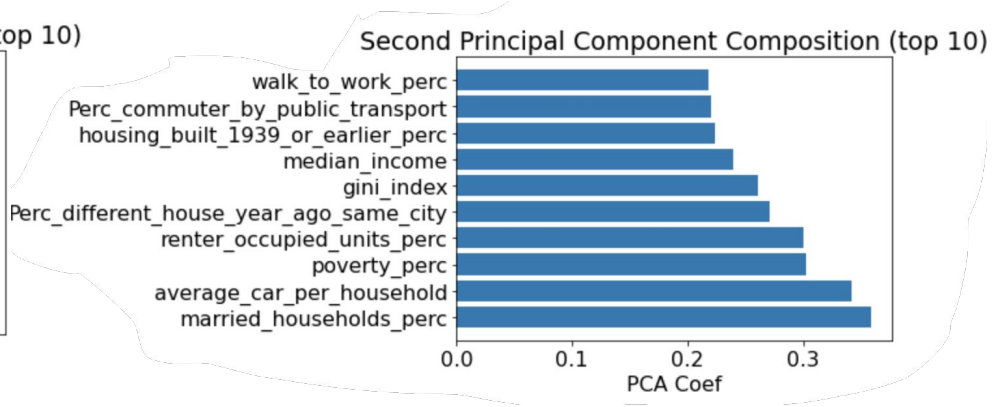
- Number of business establishments located in the ZIP code
- Number of transit terminals located in the zip code (Homeland Infrastructure data)
- Housing units per capita
- Percent of commuters with 45m+ commutes
- Percent of housing units occupied by renters



Principal Component Analysis:

Initially ran PCA on data for dimensionality reduction and to ensure no multicollinearity. Ended up keeping all principal components as models performed and had no multicollinearity





Models

- Stepwise MLR
- Lasso-Penalized MLR
- Random Forest
- Gradient Boost

Model Results: Stepwise MLR

Data Used	R^2	RMSE	Features Used	Important Features	Notes
Engineered Features	.75	348.9	22 Features	Renter occupied units, median income, walk to work percent, Percent vacant housing	Performs significantly better than the full acs data without our feature engineering (R2 ~50%)
PCA, no current rent	0.75	342.7	25 PCs		No multicollinearity concerns.
Only Current Rent	.94	248.6	1 Feature	Current Rent	Current rent is a dominant predictor

Model Results: Lasso-Penalized MLR

Data Used	R^2	RMSE	Features Used	Important Features	Notes
Engineered Features	0.74	341.2	30 Features	% units renter occupied, median income, cars / household, housing units per capita	Worse than rent-only model. R^2 of 0.55 before external sources / feature engineering.
PCA, no current rent	.74	342.2	33 PCs		
Only Current Rent	0.90	212.5	1 Feature	Current Rent	Null-model for models including Current Rent

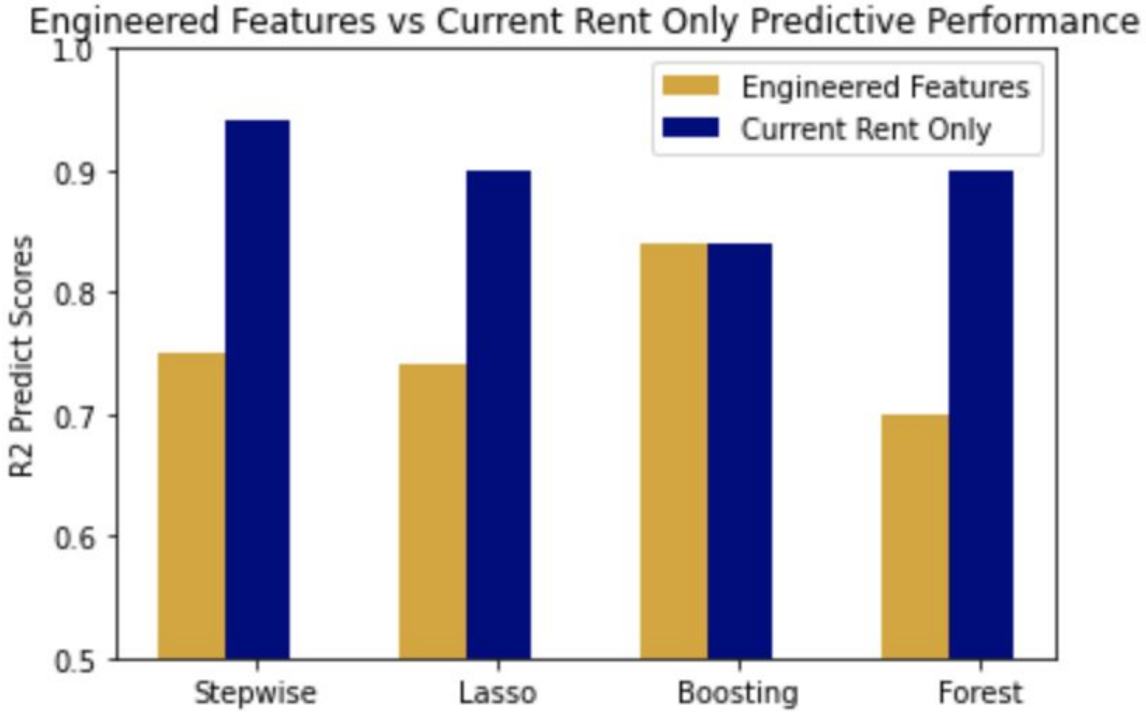
Model Results: Gradient Boost

Data Used	R^2	RMSE	Features Used	Important Features	Notes
Engineered Features	.84	258.4	33 Features	Median Income, Percent commuters over 45 minutes, Commuters by public transport, average cars per household	Best performance of models tried. Explains significantly more of future rent price.
PCA, no current rent	.81	289.3	32 PCs		Comparable results to engineered feats
Only Current Rent	.84	262.5	1 Feature	Current Rent	

Model Results: Random Forest

Data Used	R^2	RMSE	Features Used	Important Features	Notes
Engineered Features	.70	356.27	33 Features	% commuters over 45 mins, % commuters by public trans, Median income, avg car per household	Doesn't perform as well as linear models, potentially no significant non-linear patterns
PCA, no current rent	.65	386.55	33 PCs		
Only Current Rent	.90	208.77	1 Feature	Current Rent	

Model Summary



Conclusions

Wrapping up Objectives

1. Predicting Multifamily ZRI

- **Gradient Boosting Model** has the highest performance out of all the models we tried.
- Still underperforms any model that compares current rent price to future rent price.
- Combining our engineered features with current rent prices only shows increased predictive power over current rent price alone in the Random Forest model (and the improvement is only slight).

Random Forest Model Data Used	R^2	RMSE	Features Used
Only Current Rent	.90	208.77	1 Feature
Features + Current Rent	.93	179.44	34 Features

2. Predictive Features

What are the most important features in predicting future rent price?

Linear Feature Ranking	Coefficient
Percent Renter Occupied Units	436
Median Income	406
Housing per Capita (inverse)	-300
Percent Commuters Walk to Work	259

Non-Linear Feature Importance	
Features	**Feature Importance**
Percent commuters over 45 mins	11.5%
Percent commuters by public transportation	11.1%
Median income	10.7%
Average car per house	8.5%
Percent employed in finance/information	5.5%

2b. Features Lacking Predictive Power

What features are surprisingly unimportant?

- Unemployment Rate
- Poverty Percentage
- Armed forces Percentage
- Dwellings built before ‘39

3. How Important is Current Rent Price?

Just how important is current rent price?

Extremely. Explains more of the future rent price than all of our other features combined.

The rest of our features have limited predictive power when combined with current rent, only showing improvement in the Random Forest model.

Future Work

Models for **top 20 regional markets**.

Model to **predict the percent change in ZRI** over the next 3 years, more directly what we are looking for.

Features that represent the change in values over time for features found to be important by our current models.

Find rental index information that covers a **wider historical time range** (through an entire housing market cycle).

Questions?