Predicting the Zillow Rental Index

Partly Parrots x 7Park

Agenda

1. Project Scope
2. Public Data
3. Feature Selection
4. Base Model
5. Looking at Other Features
6. Next Steps

Project Scope

What are we working on?

Data Collection

Independent Predictors

Data	Granularity	Focus
Public Datasets • *ACS* • *IRS*	Zip code	Top 10 Metro Areas **tested on national level*

Data cleaning

Modelling: Multiple Linear Regression

Predicting ZRI

Train / Test Split	*Additional Features*
Train: 2016 - 2018 Test: 2019	*Forest Fire Effect* • *Air Quality* • *Fire Frequency* • *Acres Burned*

Public Data

American Community Survey (ACS)

Raw Data

Population
- Gender & Age
- Ethnicity
- Marital Status
- Education Level

Commuting
- Commuters by time
- Time

Housing
- Building Type
- Building Age
- Occupancy

Cleaning

Dropping Income + Rent Data
Using solely ZRI and IRS data

High Missingness in Columns
Data collected in 2014-2015, but not in later years

Low Missingness in Certain Zipcodes
Occurs in 0.4% of zipcodes

Aggregating Features
- Redefining age groups

Standardization

Total of **82 features** to consider

Internal Revenue Service (IRS)

Raw Data

Zipcode Granularity

Number of Tax Returns

- With taxable pensions & annuities
- etc.

Amount

- Contributions amount

Cleaning

Keeping Common Features

- For 2014-2018 data

Feature Generation

- High / low / average income

Normalization

Standardization

Correlation Drop

- Removing features with less than 20% correlation with ZRI

Total of **89 features** to consider

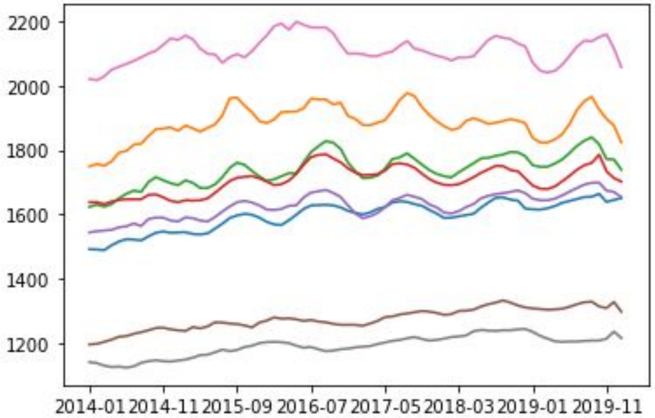
Lag period of historical ZRI : 12 months

Correlation: as lag time increases correlation decreases, however not significantly

Data availability: ZRI data taken directly from Zillow is available at end of every month, although the availability of the commercial data set is currently unknown

Seasonality: from data analysis we notice a seasonal trend of ZRI in many zip codes

Lag	ZRI
1 Month	1.000
2 Month	0.999
3 Month	0.998
6 Month	0.995
12 Month	0.989



Base Model
Feature Selection

Feature Selection by Lasso Cross Validation

Data Set

- Focus on top 10 Metro Areas
- ACS data includes 2,192 zipcodes
- IRS data includes 2,826 zipcodes

Lasso Cross Validation

- **Typical Method**: search for best lambda based on least MSE cross test set
- **Partly Parrots Method:** select the best model based on a **smaller number of features,** with similar R^2 and MSE as best "typical" method model
- Use selected features in the base model for ZRI predictions

Feature Selection Results

IRS Data

Lasso CV	Typical	Partly Parrots
R^2	0.985	0.982
MSE	0.0016	0.0019
# of Features	89	10

Selected Features

- Paid preparation
- Taxable interest amount
- Returns with:
  - Ordinary dividends
  - State local tax
  - Qualifying dividends
- Income
  - High
  - Adjusted gross
  - Average
  - Total

Last year's ZRI included

ACS Data

Lasso CV	Typical	Partly Parrots
R^2	0.988	0.987
MSE	0.0014	0.0015
# of Features	81	12

Selected Features

- No Car
- Bachelor's degree or higher (25 to 64 y/o)
- Only Bachelor's degree
- Total white population
- Owner occupied housing units at median value
- Management Arts occupation
- Median year structure built
- Number of 2-unit dwellings
- Aggregate travel time to work
- Renter occupied housing units
- Total number of housing units

Base Model
Multiple Linear Regression

Multiple Linear Regression Model Structure

Features

ZRI *12 month lag*	Selected ACS + IRS	Month dummified	Metro dummified

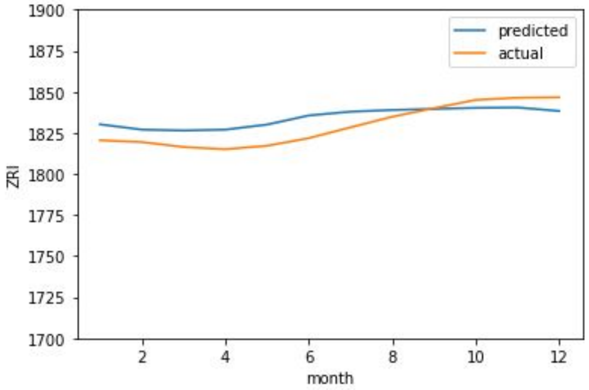
Focus Zipcodes

- Top 10 Metro areas (2,192)
- National (11,362)
- ZRI Outliers remove

ZRI
following log transformation

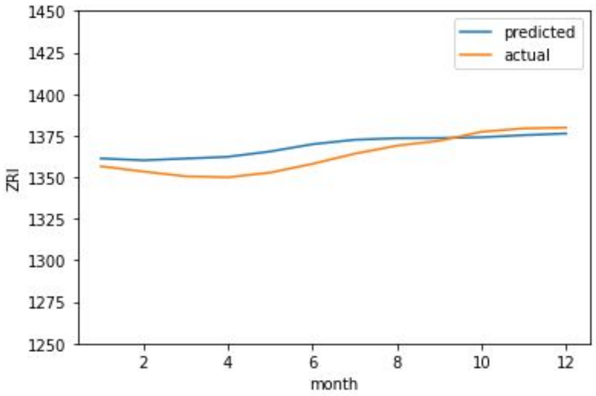
Model with Top 10 Metro Area zip codes

	Base Model	Model 2	Model 3
ZRI Previous Year	✓	✓	
Dummified month	✓	✓	✓
Dummified Metro	✓	✓	✓
ACS + IRS Features		✓	✓
Train R^2	0.989	0.989+	0.864
Test R^2	0.985	0.985+	0.872
RMSE	4.1%	4.0%	12.3%



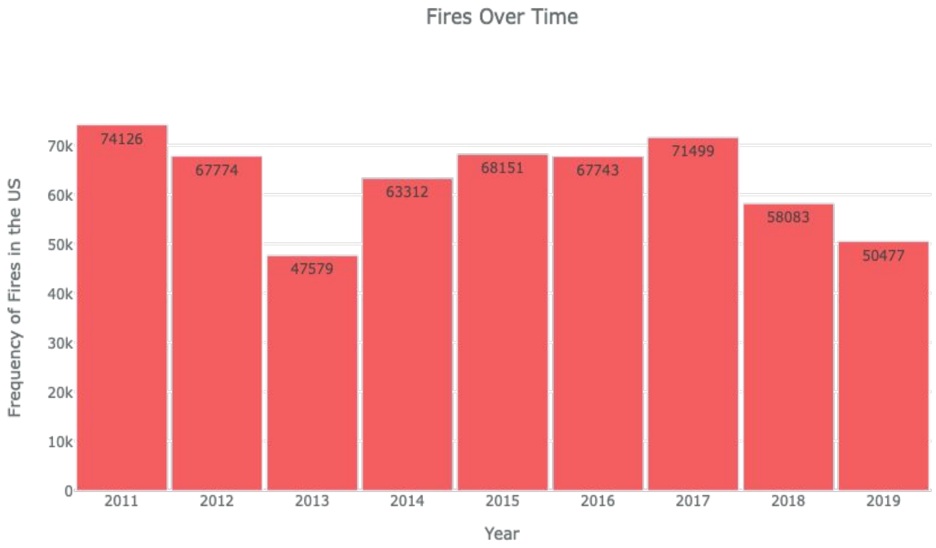
Expanding the model to national data

	Base Model	Model 2	Model 3
ZRI Previous Year	✓	✓	
Dummified month	✓	✓	✓
Dummified Metro	✓	✓	✓
ACS + IRS Features		✓	✓
Train R^2	0.984	0.984+	0.891
Test R^2	0.981	0.982	0.893
RMSE	4.9%	4.9%	12.1%



Looking at Other Features

Do forest fires affect rental prices?

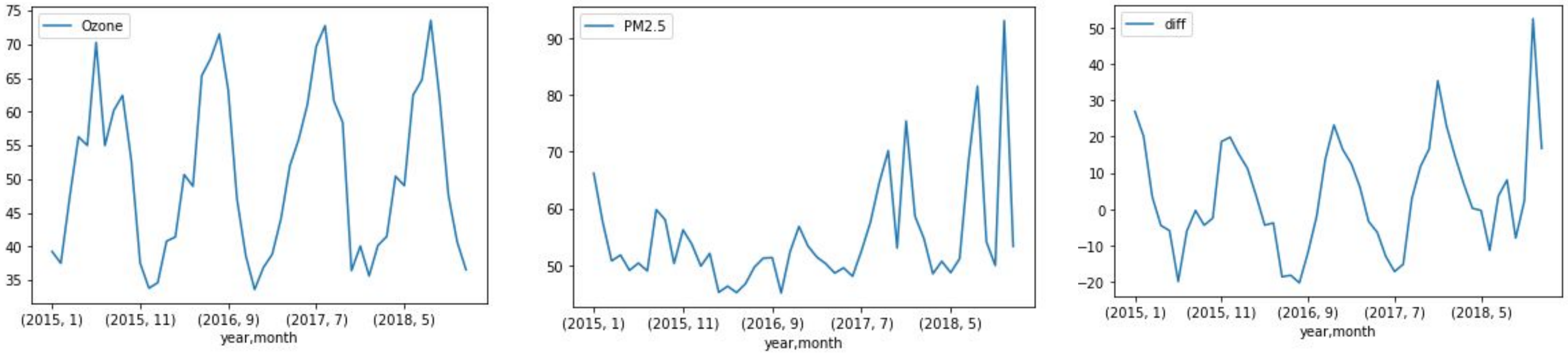


- Media attention to wildland fires increasing year over year
  - Might become an interesting facet for clients to look into
- Fires can affect:
  - Air quality
  - Home insurance
  - Rental prices?

Air Quality Index (AQI)

Pollution Sources:

- CO, NO2, SO2, **Ozone**, **PM2.5**, PM10
- Difference between PM2.5 and Ozone will also be considered as a predictor variable
- County level; Daily level → Monthly level



Next Steps

What we're focusing on after today

- Group metro areas & consider feature-feature interaction for base model
- Run model with AQI and see whether it affects rental prices
- Focus on refining the fire data
    - Link fires with physical locations
    - Create frequency and acres burned features
    - Evaluate the effect of these features on rental prices

Thank you!

Any questions?

Appendix

Data sets used in the model

data_split	ZRI_y	ZRI_x	IRS	ACS
test_data	2019	2018	2017	2017
train_data	2018	2017	2016	2016
train_data	2017	2016	2015	2015
train_data	2016	2015	2014	2014