Rental Price Modeling

By: Andrew Whittenbarger

Agenda

- Problem Statement
- Related Work
- Proposed Work
  - Intro, Dataset, Tools, Main Tasks
- Evaluation Metrics
  - Experimental Setup
  - Results
- Conclusion
- Discussion/Lessons Learned

Problem Statement

- With rent being over 20% of monthly expenses for the average American, rental costs should be a key planning factor when considering annual adjustments to employee compensation as well as making competitive offers to new hires
- **How can employers use rental data available at the end of the fiscal year (September) to predict rent costs during peak months of the following year and adjust their compensation budget accordingly?**

Related Work

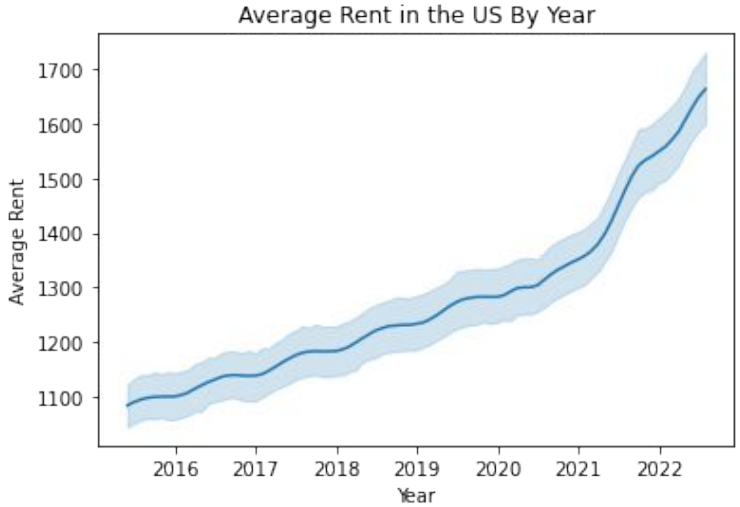
- Department of Defense (DOD) has published their methodology on how they project housing costs and set the Basic Housing Allowance (BAH) for Military Members
  - Collects data during the peak spring and summer months when the housing market is most active and uses these averages to determine the current market rent
  - Sets current market rent as the rate for the following calendar year (current model does not attempt to predict change)

Proposed Work: Intro

- The the goal of this project is to build upon previous work by evaluating if a linear model built at the end of the fiscal year (September) is effective to forecast rent prices during peak summer months of the following calendar year

Proposed Work: Dataset

- Will use the Zillow Observed Rental Index (ZORI)
  - Public dataset that aggregates data available from Zillow and the US Census Bureau
- At time of Download on November 15th, 2022 Data Set Contained :
  - Historical monthly rent for 507 Metro Areas
  - Dates ranged from March 2015 to October 2022
  - 47,141 total data points
  - 18,490 missing data points
    - Addressed in main task section of report



Proposed Work: Tools

- Python used for all programming
  - Pandas Library used for all steps of the data mining process
  - ScyPi library used for linear modeling
  - Seaborn Library used for plots

Proposed Work: Main Tasks

- Data Preprocessing/Cleaning
  - Unused columns removed
  - Metro areas missing more than 5% of the monthly rent data points removed
  - Metro areas missing 2+ consecutive months of data were removed
  - The remaining missing months were estimated as the average of the previous and next month
  - Upon Completion:
    - 215 Remaining Metro Areas
    - 18,705 Data Points
    - 215 Averaged Rent Values

Proposed Work: Main Tasks Cont.

- Data Warehousing
  - Original data format contained dates as column headers with metro areas as the row index
    - Not conducive for planned calculations
  - Transposed so dates are all in a single column

	Metro	Date	Rent
0	Chicago, IL	2015-05-31	1447.457061
1	Chicago, IL	2015-06-30	1471.120201
2	Chicago, IL	2015-07-31	1477.492074
3	Chicago, IL	2015-08-31	1483.756181
4	Chicago, IL	2015-09-30	1481.774141

Proposed Work: Main Tasks Cont.

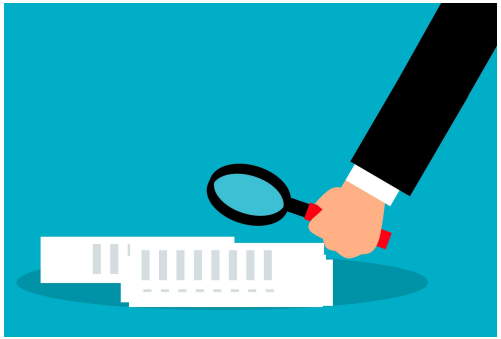
- Data Modeling
  - Linear models will be built for each metro area with date as the predictor variable.
  - 2 Types of linear models:
    - Rolling linear model (lm1)
      - Used all historical data up to the cut off date
    - Annualized Linear Model (lm2)
      - Uses data from previous 12 months only

General Linear Equation:

$y = \beta_0 + \beta_1 * x$

Evaluation

- Models will be compared to the actual market average
  - Percent Difference with confidence interval main evaluation metric
    - Percent DIfference=100*(Predicted-Actual)/Actual
  - P-Values and R^2 values analyzed for linear models
- Due to the size of the data being used, efficiency is not a concern



Evaluation Metrics: Experimental Setup

- calculations were conducted in an iterative process as seen in the table
- Rent models built using dates up to the “Max Input Date” and used to estimate rent on “Prediction Date”
- 4,515 total models (645 per year)

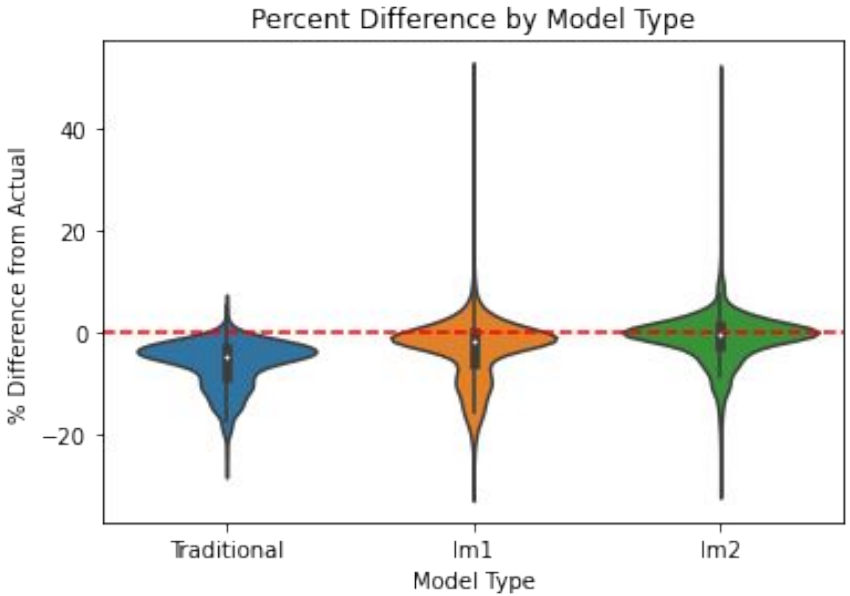
(Note: annualized and rolling linear models are mathematically identical for the first year)

Table 1: Experimental Data Setup

Iteration	Max Input Date	Prediction Date
1	AUG 2015	JUL 2016
2	AUG 2016	JUL 2017
3	AUG 2017	JUL 2018
4	AUG 2018	JUL 2019
5	AUG 2019	JUL 2020
6	AUG 2020	JUL 2021
7	AUG 2021	JUL 2022

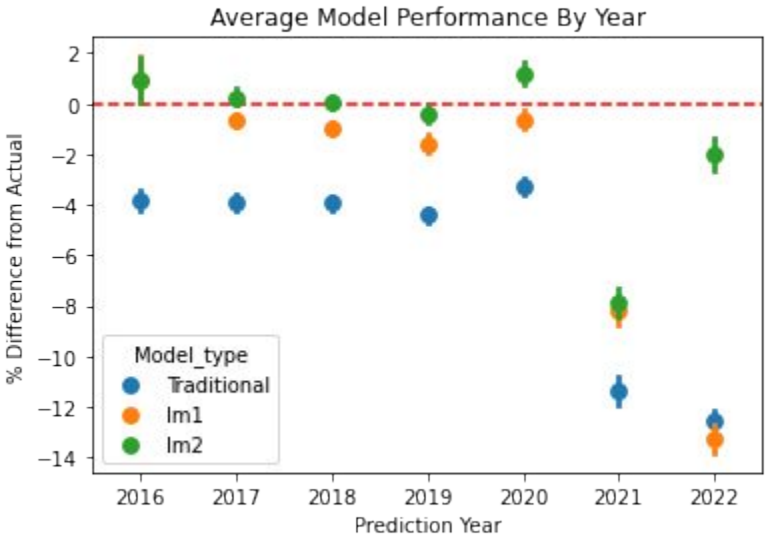
Evaluation Metrics: Results

- The annualized linear model, lm2, exhibited the smallest differences from market rent (AVG of -1.13% or $338 difference per year)
- Nearly all of the traditional averaging methods predictions fell below the actual rent for the prediction year (AVG of -6.18% or $1850 difference per year)
- Linear models tended to have higher percent differences for a small amount of models at the edges of the distribution



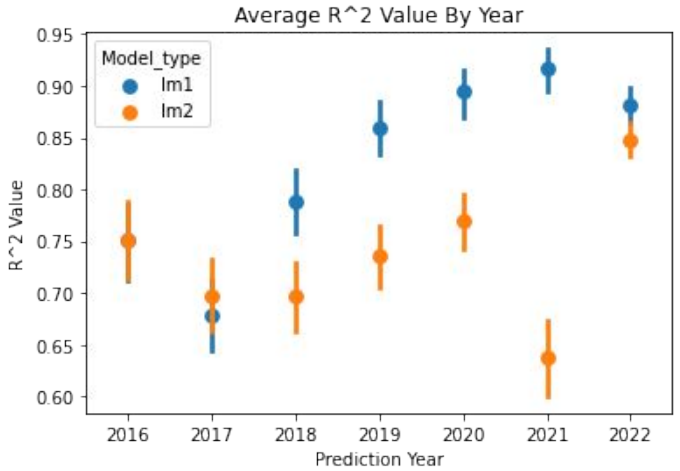
Evaluation Metrics: Results Cont.

- annualized linear model, lm2, outperformed the traditional average in every prediction year and generally outperformed the the rolling linear model
    - More sensitive to shifts in market patterns than other models
    - Potentially better suited for dynamic markets
- rolling linear model outperformed the traditional averaging method in every year except 2022



Evaluation Metrics: Results Cont.

- At an alpha=0.05 significance level 13% of the linear models were statistically insignificant
  - High percent in first year due to limited available data points
- P-values for the rolling linear model (lm1) seemed to level out, but the annualized linear model (lm2) p-values were more unpredictable
- R^2 values were between 0.62 and 0.9 suggesting a linear model is a relatively good fit to the data in most cases



% of Statistically Insignificant Models by Year

MD\YR	‘16	‘17	‘18	‘19	‘20	‘21	‘22	AVG
LM1	54	10	6	2	1	2	0.5	11
LM2	54	13	8	7	5	19	0.5	15

Conclusion

- Summary
  - Both annualized and rolling linear models show potential to significantly outperform traditional averaging, but more research is required before widespread adoption is possible
- Limitations
  - Dataset used based on data that is mostly available on the internet. It does not capture rental listings that are not advertised online and undisclosed to the Census Bureau.
  - Rents mostly increased over covered period; Results may be different in a market downturn
- Future Work
  - Retain practice of experimenting with a rolling and annualized linear model
  - Explore models that incorporate different predictors such as: inflation data, population trends, job growth, and income trends
  - Consider if the mean is the best metric to use? (Rent is a distribution, not a set value, guarantees some employees will be undercompensated)

Discussion/Lessons Learned

- Some data proved more difficult to find than others which led to a narrowing of the project scope
  - Starting with a broad areas of interest
  - If you start with a project scope that is too narrow, you might run into issues completing the project if you can't find the data you need
- Timeline shifted due limited internet access and more data preprocessing than expected which led to the project being submitted 1 week later than expected
  - Allow more time than what you think necessary to make up for unforeseen setbacks
- Gained a working understanding of several new libraries in Python

Link to Github repository with references, full report, and code

https://github.com/arwhit/rental_price_modeling