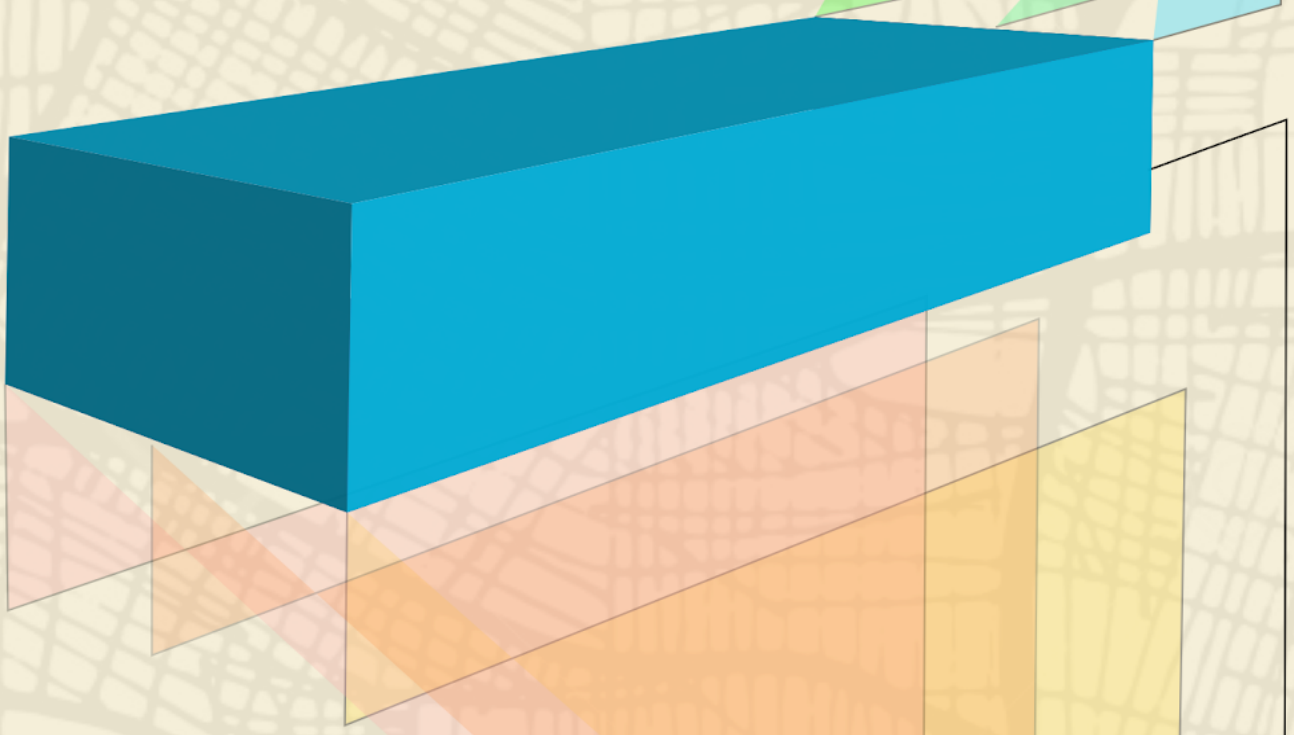
Applied Machine Learning: Predicting Future Rent Increases in Los Angeles Neighborhoods

MSBA - Mini 3 2019 ML1 Final Project



Steven M. Barnard

Ray Li

Tarak Talpade

Applied Machine Learning:
Predicting Future Rent Increases in Los Angeles Neighborhoods

Steven M. Barnard
Carnegie Mellon University - Tepper School of Business
MSBA Candidate
sbarnard@tepper.cmu.edu

Ray Li
Carnegie Mellon University - Tepper School of Business
MSBA Candidate
ruili2@tepper.cmu.edu

Tarak Talpade
Carnegie Mellon University - Tepper School of Business
MSBA Candidate
ttalpade@tepper.cmu.edu

Abstract

We present an application of linear regression, it's variations, and random forest models on time series data mainly consisting of crime, distance, and population statistics. Our application of these models ultimately aims to predict future median rent prices for individual zip codes within Los Angeles, CA. Our linear regression models build a basis for interpretability and intuition, while our random forest model would be more appropriate for production environments given the stronger results.

1 Introduction

While machine learning models and applications for predicting rent prices and real estate opportunities exist [1], creating models on time sensitive data for the years of 2010 to 2016 in a location as volatile an diverse as Los Angeles [7] presents interesting opportunities for continued application to other cities of interest.

We sought to use open source data (see data references) to confirm if past theories on topics such as Alonso's [5] theory on rent value versus distance from city center, age & income [3,4], and population [6] held true for time series prediction data in Los Angeles, CA.

Many prior models aim to provide the best prediction possible, while this is key and we do attempt to reduce the loss of our models, we take a dual approach of providing interpretability through linear regression models and performance through the use of random forest models.

2 Exploratory Data Analysis

After accounting for null/infinite values, our initial review of the data focused on analyzing crime data. Over the entirety of the dataset, the ratio of arrests for misdemeanors to felonies was approximately 2.2 misdemeanors per felony arrest (SB.4), this ratio was incorporated into the dataframe. Misdemeanor and felony rates were also created by taking the count of misdemeanors and felonies and dividing by the population per zip code to get two separate crime ratios, one for misdemeanors and one for felonies (SB.5, TT.7-9). These crime ratios were then plotted against median rent (see Figure 1).

** References to code blocks from authors is indicated by (Initials.Input Cell) i.e. (SB.12) = Steven Barnard cell 12 in the notebook.

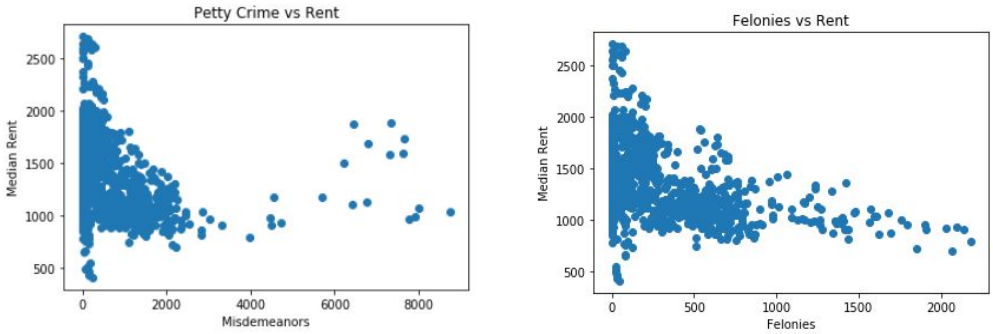


Figure 1: Scatter plots of rent versus crimes. (TT.7-9)

Misdemeanor and felony count were also broken down per year (SB.12-13). Crimes appear to be stagnant from 2010-2014 with spikes in crime rate in 2015 and 2016 for both misdemeanors and felonies (See Figure 2).

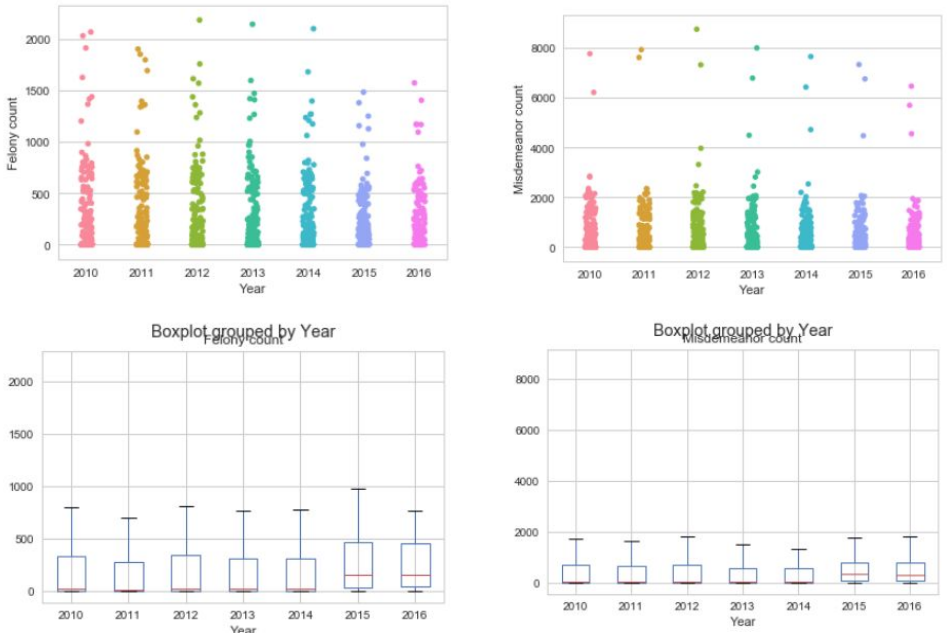


Figure 2: Scatter and Box plots for felony (left) and misdemeanor (right) broken down per year.(SB.12-13)

Additional reviews on crime were completed and collinearity was assessed (RL.12-13). High collinearity was observed over the crime data, we attributed this to the fact that the data representing arrests and crimes were not guaranteed to be mutually exclusive [A]. For this fact, we chose to aggregate crimes when running our models (RL.16).

Other data exploration was completed via summary statistics, feature distributions (RL.7,9-10), rent distribution (SB.19, TT.11), distance from downtown (SB.18), median age per year(SB.17), household size per year (SB.16), population per year (SB.14), and households per year (SB.15).

With these additional plots we noticed that over time population decreased slightly, median age remained stagnant, distance from city center decreased, household size decreased while number of households remained stagnant, and rent prices increased.

We speculate these changes are indicative of families moving out of LA and individuals moving in, likely towards the city center where rent prices would be higher based on our data and prior established theories [5].

** References to code blocks from authors is indicated by (Initials.Input Cell) i.e. (SB.12) = Steven Barnard cell 12 in the notebook.

3 Methods

We cast this problem initially as a linear regression to use our set of features to predict a real value of median rent in USD, with the intention of delving into more complicated models with the intent to maximize results for more realistic production applications. Our data, being time series data, required special handling of the train, validation, and test splits; we applied backtesting in order to account for the temporal nature of the data [2]. Data from 2010 to 2014 was included in the training set, the validation set was 2015, and 2016 was the test set.

Each of the three authors conducted models consisting of linear models (SB and TT) and random forest (RL).

The models were implemented using the scikit learn python packages for each model respectively. Feature selection was done manually in both initial linear models from SB and TT. Feature selection was also attempted via ridge and lasso from TT. Alphas were also selected for both high and low values during implementation of the ridge and lasso linear regression models (TT 14,17). Coefficients were then ranked, and metrics on the models were recorded for mean squared error (MSE), root mean squared error (RMSE), and R-squared values (SB.29,31)(TT.12-18).

Data for the random forest was standardized via pre-processing, a baseline was then run, followed by hyper-parameter tuning, and determination of feature importance as well as reporting of RMSE and R-squared metrics (RL.18-26).

4 Results

The results of the metrics output from the models are listed below in Table 1 and followed by the visualization of the the feature importance output from the random forest model; additional visualizations from on the linear regressions can be viewed in the notebooks (TT.12-18). Model coefficients for linear regression models can be seen in notebooks as well (SB.25-31)(TT.12-18).

Table 1:

Model #	Model Type - Data (Author)	MSE	RMSE	R-Squared
1	Linear Reg - Validation(SB)	133437	365.29	0.407
2	Linear Reg - Test(SB)	124611	353.00	**0.480**
3	Linear Reg - Test (TT)	**106791**	**326.79**	0.461
4	Linear Reg Ridge - Test (TT)	106831	326.85	0.461
5	Linear Reg Lasso - Test (TT)	110269	332.06	0.444
6	RF Baseline - Test (RL)	Null	140.28	Null
7	RF Tuned - Test (RL)	Null	**133.21**	**0.914**

*blue highlights indicate best in model class, black bold indicate best overall.

** References to code blocks from authors is indicated by (Initials.Input Cell) i.e. (SB.12) = Steven Barnard cell 12 in the notebook.

	importance
Median Age	0.242031
distance_from_towncenter_mi	0.149067
Average Household Size	0.125321
Zip Code	0.101261
CrimeCount	0.062211
Total Males	0.058450
AverageHealthScore	0.057814
Total Population	0.046954
Total Females	0.046469
HealthViolationCount	0.045926
Total Households	0.040238
Year	0.024256

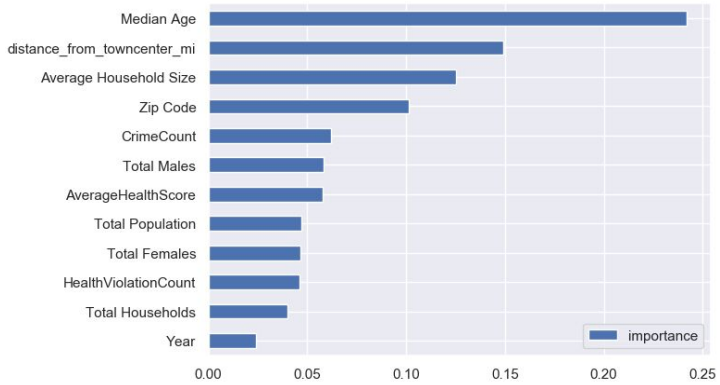


Figure 3: The table and graphical representation of feature importance output by the random forest model (RL24,26)

5 Discussion

Our models for linear regression, while not as powerful in terms of prediction as our random forest model, are easily interpretable and can be used as tools to assist in building intuition when it comes to pricing Los Angeles rent in future years. The random forest model was by far the best in terms of prediction with an R-squared value that explained ~91% of the fit.

The features that explained high levels of fit across model types where median age, year, distance from town center, felonies per capita(crime), and average household size. These features confirmed our speculation on why areas would have higher or lower rent.

Specifically for the random forest model, median age has a high level of importance and we have found literature that agrees that higher median age can be synonymous with higher income [4] and lower crime rates[3] both associated with higher median rent. Higher median age can also account for a reduction in household size, children heading out to college but families staying in the area. The level of importance for distance from city center can be supported with literature from William Alonso's bid-rent theory which discusses rent decreasing in relation to distance from a city's center [5]. Year's level of importance appears to mimic approx 50% of the 1.26% increase in prices due to inflation; we surmise the remaining 50% would be attributed to the increases in rent due to inflation compounded over prior years considering not all leases would be for 1 year. Further iterations of the model should take into account the use of feature impution and consider the effects that impution can have when used across large proportions of data. Due to the small number of data instances and feature imputation for 2 of the 10 features listed highest in feature importance, health violation count and average health score, the model likely tends to over fit currently; however, with additional time series data and features we maintain that this model would be preferable for production.

Further research is needed to fully explain the prediction of rent in Los Angeles and have the predictions applied to other cities across the United States. We anticipate that with a much larger dataset our results would improve, the additional data would be available from future years and the model would improve over time. Additional features that would potentially increase predictive power of our models would be features such as: income per zip; credit ratings per zip; population changes(immigration/emigration); distance from police/fire dept.; emergency service response time per zip; number of whole food stores per zip; check cashing stores per zip; and school rankings.

Overall, our results provide confirmation of other theories and support common knowledge that rent generally increase in safer areas the closer you get to a city's center; however, the applied process of being able to predict rent prices, in a city as diverse and volatile as Los Angeles, CA, for the following year given a set of predictions can have great practical applications for those seeking to move to Los Angeles or develop/buy real estate.

** References to code blocks from authors is indicated by (Initials.Input Cell) i.e. (SB.12) = Steven Barnard cell 12 in the notebook.

Acknowledgments

We acknowledge the teaching and support of Professor Zachary C. Lipton.

References

[1] Baldominos, Alejandro, and Iván Blanco. “Identifying Real Estate Opportunities Using Machine Learning.” *Applied Sciences*, 21 Nov. 2018, arxiv.org/pdf/1809.04933.pdf.

[2] Brownlee, Jason. “How To Backtest Machine Learning Models for Time Series Forecasting.” 16 Dec. 2016, machinelearningmastery.com/backtest-machine-learning-models-time-series-forecasting/.

[3] Cornelius, Caitlin, et al. “AGING OUT OF CRIME: EXPLORING THE RELATIONSHIP BETWEEN AGE AND CRIME WITH AGENT BASED MODELING.” Apr. 2017, p. 7., doi:http://scs.org/wp-content/uploads/2017/06/6_Final_Manuscript.pdf.

[4] Fontenot, Kayla, et al. “Income and Poverty in the United States: 2017 Current Po.” *Current Population Reports*, Sept. 2018, p. 6., doi:https://www.census.gov/content/dam/Census/library/publications/2018/demo/p60-263.pdf.

[5] Alonso, W. (1960). A Theory of the Urban Land Market. Papers in Regional Science 6(1): 149-157

[6] Salcedo, Alejandrina, et al. “Families as Roommates: Changes in U.S. Household Size From 1850 to 2000.” *National Bureau of Economic Research*, Nov. 2009, p. 7., doi:https://www.nber.org/papers/w15477.pdf.

[7] Warden, Pete. “LA Map Image.” *Pete Warden's Blog*, 8 Oct. 2013, petewarden.files.wordpress.com/2012/04/436ca-6a00d83454428269e20168e9bd8766970c-800wi.jpg?w=550.

Data

[A]https://www.kaggle.com/cityofLA/los-angeles-crime-arrest-data

[B]https://www.kaggle.com/cityofLA/los-angeles-census-data#2010-census-populations-by-zip-code.csv

[C]https://www.kaggle.com/cityofLA/la-restaurant-market-health-data#restaurant-and-market-health-violations.csv

[D]https://usc.data.socrata.com/Los-Angeles/Rent-Price-LA-/4a97-v5tx

[E] http://www.in2013dollars.com/2015-dollars-in-2016

*** https://en.wikipedia.org/wiki/Haversine_formula -- reference to formula used to calculate geo-distance from city center

**** CSV data set attached to email along with ipy.notebooks.

Cover page created by Steve Barnard with transparent overlay of hand drawn LA map from - https://previews.123rf.com/images/booblgum/booblgum1803/booblgum180300211/98112601-los-angeles-california-usa-city-map-in-retro-style-black-and-white-color-outline-map-vector-illustra.jpg

** References to code blocks from authors is indicated by (Initials.Input Cell) i.e. (SB.12) = Steven Barnard cell 12 in the notebook.