**TITLE: XXX**

**-running time**

**-student info**

**-word count**

**- github repo(make public?) and change the data paths? Or add them in the upload file**

**Introduction 270**

According to the *Understanding Inequalities Project* *''the housing market has a key role in "sorting" poorer households into areas with the worst pollution, schools, crime, and employment.* ''(*Housing | Understanding Inequalities*, no date). While housing is a basic human need, for some homeownership is an unachievable goal and for others another great investment opportunity. Housing affordability has long been a very pressing matter for cities around the world and although housing discrimination and practices such as Redlining in the USA(X) have been condemned decades ago, the housing crisis in states such as California is still present (Chew and Muñoz Flegal, 2020).

This study will investigate whether house price variations are affected by environmental and socioeconomic factors and test the premise that pollution, unemployment, and health statistics all contribute to the overall house price trends. For this, California is chosen as a case study, and a variety of environmental, health and socioeconomic indicators for its census tracts are used to predict house price variations for the same area.

This analysis is structured as follows: After a summary of the relevant literature in section 2, section 3 poses the research question of this paper. The presentation of the datasets in section 4 is then followed by section 5 outlining the methodology used throughout the analysis section 6. Section 7 presents the results that are discussed in more detail in section 8. Finally, section 9 concludes and proposes further research.

**Literature review (min 3 sources) 300**

House prices, like the economy, are in constant flux and are of course affected and driven by many factors, which is why house price estimation and the study of the drivers behind has been and still is, a broad field of study. Many of these studies are based on hedonic regression models (Jafari and Akhavian, 2019) that quantify the influence of the various factors on the good (i.e. price), estimating the influence attributes such as the number of bedrooms or the location have on the demand or price of the house (Hedonic Regression Definition, 2022).

(Hanink, Cromley and Ebenstein, 2012) and (Montero, Fernández-Avilés and Mínguez, 2018) however criticise these models as they do not incorporate the spatial parameter inherent in property data. They argue that spatial heterogeneity and spatial autocorrelation need to be accounted for and therefore proceed their analysis based on Geographically Weighted Regression (GWR) and spatial hedonic models such as spatial error models (SEM) and global spatial autoregressive (SAR) models. (Montero, Fernández-Avilés and Mínguez, 2018) in particular, focused on the ­­impact of the environment, concluding that house prices are strongly affected by pollution and odours. However, their measure of environmental factors was based on the resident’s perception, which is a somewhat subjective measure and arguably difficult to be generalised.

More recent studies make use of modern machine learning methods in their attempt to model house prices such as (Phan, 2019), comparing different methods such as Regression Tree and Support Vector Machine (SVM) in combination with dimensionality reduction methods in order to predict house prices in Melbourne from the properties’ features.

This study will be conducted not for a specific city but on the whole of California using the spatial unit of the U.S Census tracts. Differently from previous studies and partially driven by the unavailability of data in the public domain, this research will not focus on individual house prices but the change in the House Price Index for each census tract.

**Research question 30**

The research question is split into two sub-questions formulated as follows:

*R.Q.1: Is there a relationship between the variation of the House Price Index and environmental, health, and socioeconomic indexes of census tracts in California?*

*R.Q.2: Can the change in House Price Index for a census tract be predicted from environmental, health, and socioeconomic indexes?*

**Presentation of data 350**

The data used for this analysis were obtained from the following sources:

\* [CalEnviroScreen 4.0, downloaded 06 April 2021:

A detailed dataset containing a variety of environmental and health indicators for California’s 8000 census tracts from which the CalEnviroScores are calculated, a measure of pollution and potential vulnerability of a population to the effects of pollution. The results are summarised in a report and featured in an online mapping tool, The California Communities Environmental Health Screening Tool that aims to ‘*provide a clear picture of cumulative pollution burdens and vulnerabilities in communities throughout the state’* (August et al., 2021).This study will be using the latest version of this dataset, published in October 2021 by the Office of Environmental Health Hazard Assessment (OEHHA).

\* [House Price Index Census Tracts] downloaded 04 April 2021:

An extensive dataset containing House Price Index values on census tract level across USA for each year since 1975. The House Price Index (HPI) is a broad measure of price changes of single-family houses, measuring repeat sales on the same properties. (*FHFA House Price Index | Federal Housing Finance Agency*, no date). The HPI is published by the Feral Housing Finance Agency (FHFA) and serves an indicator of house price change and can be used to estimate housing affordability. This study will be using the most recent dataset available online, including information up to 2021.

Variables

Since the HPI is a measure of change, the value itself has little meaning. Therefore, the difference between two points in time is considered more appropriate for this analysis and the dependent variable is defined as the difference in the HPI value between 2017 and 2019. This timeframe corresponds to the CalEnviroScreen 4.0 data and does not consider 2020 and 2021, which due to the Covid19 pandemic, are not representative years and could have unusual trends. Finally, it should be mentioned that a HPI index with a common base line year was chosen so that the values are directly comparable across census tracts. After comparing the number of missing values, the HPI2000 was selected over the HPI1990.

All independent variables for this study derive from the CalEnviroScreen4.0 dataset. Indicators such as the CalEviroScore (CES), indicators averaging subset of indicators and the Percentile version of indicators were excluded from this study as it is not clear how these were calculated. While the difference between the indicator values from the same time period (2017-2019) might have been more accurate, the accompanying report states that the calculation method for multiple indicators has been improved in the 4.0 version (August et al., 2021). This means that the values between this and the CalEnviroScreen 3.0 dataset will not have been directly comparable. Since this study is focusing on HPI change across census tracts we will assume that the change in pollution indicators between two years as minimal and focus more on the differences across the census tracts.

More detailed information for each variable can be found in the Data Dictionary accompanying each dataset.

For ease of access, the data has been saved in this [GitHub repository]

ADD table?

Initial Limitations

The large number of missing values and the fact that all variables used are indicators and broad measures that cannot capture in detail the variation across the geographic space, which poses serious limitations to the generalisation of this study.

**Methodology 400**

The methodology of this report will be laid out in detail in the analysis section of this notebook. A summary however of the main important points is presented here:

Data cleaning and Initial data exploration

After the initial data cleaning, sub-setting and filtering, the basic summary statistics showed the existence of a large number of outliers in many of the independent variables as well as in the dependent variable itself. The analysis proceeds without their removal as each entry represents a census tract and therefore a valuable part of the geographic entity of California.

data preparation

A considerable amount of entries with missing values, which unfortunately are removed leaving the dataset with just over half of California's 8000 census tracts.  It is nevertheless considered a large enough dataset to be split into test, validation, and training sets in preparation for the subsequent analysis.

In-depth analysis

This study uses regression analysis methods to investigate the relationship between the dependent and independent variable and test whether the different models used can predict the change in HPI.

Even though the scatterplots between each of the independent variables and our dependent variable show that there is no linear relationship, a linear regression model is used as baseline model, for the evaluation of the performance of the more complicated supervised machine learning models used.

The coefficient of determination R2 will be used for model performance evaluation and comparison. The training of the models will be solely done with the training set, the validation set will be used for hyperparameter tuning and as is common practice the testing set will be used only for the final evaluation of each model.

A Random Forest Regression model and a XGBoost model (Gradient Boosting Decision Tree) are used followed by hyperparameter tuning by performing two different methods: holdout validation and cross-validation. For more details, please refer to the Analysis section.

It is then tested whether the best performing model can be improved further by applying Principal Component Analysis (a dimensionality reduction technique) beforehand on the set of independent variables.

Permutation feature importance is applied in all cases, aiming to identify the most important variables for the HPI change prediction and in an attempt to better interpret and understand the results and performance of the models.

**Analysis**

Add table of variables?

**Results 250**

-visualisations (labels, clear)

-interesting results and interpretation

This section summarises the results of the analysis performed

The validation prediction plots below illustrate and compare the performance of the calibrated models on the validation and test sets. As the shape of the data demonstrates none of the model performs well at predicting the actual values for the HPI change and the low R2 value confirms this with values below 40% in all cases. However, judging from the shape of the data XGBoost model without PCA seems to be the best performing model.

Chart, scatter chart

Description automatically generated

b

Table

Description automatically generated

As expected, the linear regression model, while performing similarly on the training and testing sets, has poor performance overall. In fact, comparing the R2 values, the models do not perform significantly better than the linear regression model on the testing sets.

The XGBoost Model records the highest R2 scores on the testing data despite performing worse than the RF model on the training set. The Random Forest model has a significant difference in the R2 scores between training and testing set, which indicates high bias and overfitting.

After performing PCA on the dataset, the prediction score for both models and in all data sets has decreased. The drop however in the R2 difference indicates that these model, albeit less accurate perhaps have better generalisation. It has been argued that PCA is not always appropriate for predictions as it ‘pollutes’ the predictors (*PCA or Polluting your Clever Analysis | R-bloggers*, no date).

Chart, bar chart, histogram

Description automatically generated

Comparing the feature importance plots of the RF and XGB models, we identify Education, Asthma and Lead as the three most important ones in both models, one representative indicator for socioeconomic, health and environmental factors respectively. It should also be noted that in the XGB model feature importance seems more spread across the variables whilst in the RF model Education seems to dominate. The interpretation however of the above is rather difficult and since the model performance is weak the validity of this results is highly questionable.

**Discussion 200**

**Outliers: https://stats.stackexchange.com/questions/140215/why-boosting-method-is-sensitive-to-outliers**

-limitation and shortcomings of analysis

-critical reflection

-add some lit reference here as well

- these indicator should be additional to other, not investigated alone

-Not considered: demographics(race)

-outliers:

-Better tuning: to start with a randomized search to reduce the parameters space and then launch a grid search to select the optimal features within this space.(link)

-other models: ANN? SVG? Kernel PCA

spatial autocorrelation - spatial dimension [PAPER: Examining the spatial relationship between environmental health factors and

house prices: NO 2 problem?]

2020-2021 covid

spatial unit

ANN

[link](https://thinkingneuron.com/using-artificial-neural-networks-for-regression-in-python/)

This template can be used to fit the Deep Learning ANN regression model on any given dataset.

You can take the pre-processing steps of raw data from any of the case studies here.

Deep ANNs work great when you have a good amount of data available for learning. For small datasets with less than 50K records, I will recommend using the supervised ML models like Random Forests, Adaboosts, XGBoosts, etc.

The simple reason behind this is the high complexity and large computations of ANN. It is not worth it, if you can achieve the same accuracy with a faster and simpler model.

You look at deep learning ANNs only when you have a large amount of data available and the other algorithms are failing or do not fit for the task.

The compared methods and models do not perform well at predicting the change in HPI for the study area. Random Forest and XGBoost were chosen as they are considered relatively robust to outliers and can handle Non-linear data. However as mentioned, this dataset has extreme outliers that might have strongly influenced the model performance. Perhaps other models and methods such as ANN and SVM that have not been used here would yield better results.

Furthermore, the hyperparameter tuning performed in this analysis was certainly not exhaustive und could be optimised. It has been suggested to widen the hyperparameters values tested by starting with a randomized search first to reduce the parameter space and then use a grid search to select the best hyperparameters.(*Grid Search vs. Randomized Search -*, no date)

Another important limitation of this analysis is that the spatial dimension and spatial autocorrelation has not been considered. +++

The spatial unit of the census tract might also not be appropriate, and it would be interesting to repeat the analysis on different scales.

+++Add some literature here (NO2?)

Finally, this study will argue that while there seems to be relationship between the environmental, health and socioeconomic factors and the change in HPI, predicting the change solely from the above factors is not meaningful. It is suggested that these parameters should be added as additional parameters to traditional models looking at house price predictions from the properties’ attributes to improve the prediction accuracy.

**Conclusion 200**

Answer my RQ

This analysis has investigated the relationship between the HPI change and environmental, health and socioeconomic indicators for census tracts in California. In summary and returning to the research questions set at the beginning of the report while a relationship between the variables has been identified the constructed models do not suffice in accurately predicting the change HPI from the mentioned factors.

**Summary**

**reflection**

**Bibliography**

August, L. *et al.* (no date) *OEHHA Authors: OEHHA Editors: Acknowledgments*.

Chew with Chione Lucina Muñoz Flegal, A. *et al.* (no date) *Facing History, Uprooting Inequality: A Path to Housing Justice in California Thank you to Acknowledgments*. Available at: http://www.policylink.org.

*FHFA House Price Index | Federal Housing Finance Agency* (no date). Available at: https://www.fhfa.gov/DataTools/Downloads/Pages/House-Price-Index.aspx (Accessed: April 12, 2022).

*Grid Search vs. Randomized Search -* (no date). Available at: https://maelfabien.github.io/machinelearning/GridRand/#randomized-search (Accessed: April 14, 2022).

Hanink, D.M., Cromley, R.G. and Ebenstein, A.Y. (2012) “Spatial Variation in the Determinants of House Prices and Apartment Rents in China,” *Journal of Real Estate Finance and Economics*, 45(2), pp. 347–363. doi:10.1007/s11146-010-9262-3.

*Hedonic Regression Definition* (no date). Available at: https://www.investopedia.com/terms/h/hedonic-regression.asp (Accessed: April 12, 2022).

*Housing | Understanding Inequalities* (no date). Available at: https://www.understanding-inequalities.ac.uk/research-themes/housing (Accessed: April 12, 2022).

Jafari, A. and Akhavian, R. (2019) “Driving forces for the US residential housing price: a predictive analysis,” *Built Environment Project and Asset Management*, 9(4), pp. 515–529. doi:10.1108/BEPAM-07-2018-0100.

Montero, J.M., Fernández-Avilés, G. and Mínguez, R. (2018) “Estimating environment impacts on housing prices,” in *Environmetrics*. John Wiley and Sons Ltd. doi:10.1002/env.2453.

*PCA or Polluting your Clever Analysis | R-bloggers* (no date). Available at: https://www.r-bloggers.com/2012/08/pca-or-polluting-your-clever-analysis/ (Accessed: April 14, 2022).

Phan, T.D. (2019) “Housing price prediction using machine learning algorithms: The case of Melbourne city, Australia,” in *Proceedings - International Conference on Machine Learning and Data Engineering, iCMLDE 2018*. Institute of Electrical and Electronics Engineers Inc., pp. 8–13. doi:10.1109/iCMLDE.2018.00017.

**Appendix**

**NOTES:**

->publication standard

Methodology ideas

PCA (linear first, then non-linear to check my data is not ‘twisted’)

Result- input to regression or classification?

Try different models for machine learning

Hyperparameter finetuning

Model comparison

Final conclusions

(clustering?, ~~image classification~~)

This study will investigate if there is a relationship between a variety of indicators for environmental, health and socioeconomic features that compose a geographic unit and house price variations for the same area.

The case study chosen is the State of California (USA)

and the baseline assumption is that pollution, unemployment, and poor health statistics will be reflected in the overall house price trend.

*The potential increase in house value affects the quality of life of residents as well as the national economy.* (Jafari and Akhavian, 2019)

Lit review

-House price predictions usually: hedonic pricing method (Jafari and Akhavian, 2019)+ investopedia

* Most influential variables: size, location, number of bathrooms and bedrooms.
* Find that the house characteristics and location have a stronger influence on price than the neighbourhoods’ characteristics (proximity to open space)
* More information about neighbourhood characteristics might alter the results

(Hanink, Cromley and Ebenstein, 2012) suggest further research to incorporate environmental factors.

(Montero, Fernández-Avilés and Mínguez, 2018) using a vast array of different models… conclude that house prices are strongly affected by environmental factors. The measure for environmental factors was based on the ‘ residents’ perception of pollution and unpleasant odours’ which is a rather subjective measure and arguably difficult to be generalised.

Broad field of study: house price estimation and factors that are driving it

Hedonic regression models have become standard practice: focus on the house attributes

Studies such as X and Y incorporate the spatial factor, considering spatial autocorrelation and the like

More recent studies turning more to the environment and neighborhoud of the house, incorporating environmental factors such as y and z

This study focuses on the spatial unit of census tracts on the whole of California(not just one city ) and differently from previous studies and partially driven by data availability focuses not individual house prices but the change of the House price indicator for each tract.

Variables

All independent variables for this study derive from the CalEnviroScreen4.0 dataset. Indicators such as the CalEviroScore (CES), indicators averaging subset of indicators and the Percentile version of indicators were excluded from this study as it is not clear how these were calculated.

While the difference between the indicator values from 2018 and 2021 might be more correct, the accompanying report states that multiple indicators calculation method has been improved in the 4.0 version. This means that the values between the two datasets will not be directly comparable.

Since this study is focusing on HPI change across census tracts we will assume that the change in pollution indicators between one or two years as minimal and focus more on the differences across the tracts.

The HPI is a measure of change, and as such the value itself has little meaning. Therefore we will be comparing the change in HPI (whether it went up or down and by how much) between 2017-2019, which corresponds as a timeframe to the CalEnviroScreen 4.0 data and does not consider 2020 and 2021 which due to the Covid19 pandemic are not representative years and could have unusual trends.