**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 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 index 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 index variations for the same area using and comparing different machine learning algorithms.

This analysis is structured as follows: After a brief summary of 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 objectives on this topic.

**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. For decades many of these studies were based on hedonic regression models (Jafari & 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 et al., 2012) and (Montero et al., 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.

Machine learning methods have been employed to model housing prices since the early 2000 and as Park and Kwon Bae indicate, multiple studies have tested their performance against traditional hedonic models. The authors themselves proceed to comparing various classifiers and conclude that these can significantly contribute towards accurate house price predictions (Park & Kwon Bae, 2015).

Similarly, but more recently, (Phan, 2019) compare different algorithms, among them, 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.

Multiple studies have examined the relation between environmental and socio-economic factors and the real-estate market. (Barreca et al., 2018) showed the spatial correlation between housing and social vulnerability indicators and house prices while (Montero et al., 2018) 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. A paper by Boyle and Kiel, is a comprehensive review of hedonic model studies that considered the environmental factors of air quality, water quality and distance from toxic sites (Boyle & Kiel, 2001).

This study will as well compare different machine learning methods in combination with dimensionality reduction, but instead of focusing on a specific city, it will be conducted on the whole of California using the spatial unit of the U.S Census tracts. Moreover, 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 which will also enable to shift the focus to the influence of environmental and socioeconomic factors rather than the properties’ attributes.

**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 as 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 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

Both datasets contain a considerable number of entries with missing values. After the 2 datasets have been joined these entries 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 a baseline model, for the evaluation of the performance of the more complicated supervised machine learning models implemented.

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. 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.

Finally, it 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.

For more details, please refer to the Analysis section.

**Analysis**

Add table of variables?

This section contains the code and documentation of the analysis.

6.1 Data Cleaning

First, the necessary libraries need to be imported

Reading in the data

6.1.1 House Price Index data

Check for missing values:

After the first inspection of the data the key things to note are:

\* There are 1,048,575 entries in total, containing information for the whole State of California

\* There are multiple NA values which will need to be addressed.

\* hpi1990 has more than 50% of NA values

\* The data covers the years 1975-2021

\* There are no categorical variables

The dataset will be filtered and sub-setted to:

\* Extract census tracts of California only.

\* Keep only 'hpi1990' as the dependent variable, so that all tracts have the same baseline year.

\* Year: keep the most recent year with the least NA values

There is a significant number of missing values even for the most recent years.

As described above, the depended variable will be created as the difference between 2017 and 2019

5.1.2 CalEnviro Data

Load the dataset that will be used for the independent variables.

While this dataset does have missing values, they only form a very small percentage.

5.1.3 summary statistics

The KDE plots above show that most of the variables are not normally distributed. Furthermore, as the last row of the scatterplots shows. none of the independent variables displays a linear relationship with the dependent variable (hpi). Therefore, a multivariate linear regression model will not be appropriate for explaining this relationship.

The boxplots indicate the existence of outliers in all variables with a few of them recording extreme outliers: "Pesticides","Tox.Release", "Traffic"

The correlation matrix shows that some multicollinearity exists between variables with the highest being between poverty and education. However, the VIF with a threshold of 5 does not drop any of the variables.

6.3 In-depth Analysis

Linear regresssion

As expected, the R2 is very low, and the p-values indicate that many of the variables are statistically insignificant. Furthermore, the dense Residuals versus fit plot does not convey a constant variance of the residuals indicating that the assumptions of the linear regression model are not met.

Random Forest

The [Random Forest Regressor is an ensemble method, fitting multiple Decision Tree Classifiers on subsets of the dataset. The output is the average of these sub-trees.

While the R2 on the training data is very high the rather low R2 on the testing data indicates overfitting.

\*\*HYPERPARAMETER TUNING\*\*

two methods will be compared:

\* Holdout validation (grid search)

\* Cross validation (grid search)

There are of course more hyperparameters that can be tuned, but we will focus on:

\* ```n\_estimators ```(default=100): The number of decision trees.

Usually a higher number of trees improves the model's performance. However it is also slowing down the training process.

\* ```max\_depth``` (default=None): The depth (splits) of each tree

Usually the deeper the tree the more information it captures about the data. However allowing maximum depth might result in overfitting.

##### \*\*Holdout validation\*\*

Holdout validation will be used here, by specifying the 'cv' parameter (the cross validation splitting strategy) as an iterable yielding (train, test) as arrays of indices.

The Random Forest Regressor Model tuned with Holdout Validation has slightly worsened the performance, indicating that the default values were perhaps more appropriate.

The cross-validation return the same value for n\_estimators but 13 instead of 18 for the value for max-depth compared to the holdout validation method.

Since Cross validation is generally considered as more robust, we will keep these hyperparameters

The Random Forest Regressor Model tuned with Cross Validation, also returns slightly worse results. (In the default parameters max\_depth is defined as None which might explain the difference here). Since Cross-validation is generally considered more robust we will keep the later model.

RF final estimate of performance

Using the test set (that has not been used at any stage during the model training and validation process) the final estimate of the RF model can be reported.

While a \*\*validation curve\*\* is not used to tune a model, because it might lead to the model being biased and not a good estimate of the generalization of the model, it is useful for evaluating the existing model based on hyper-parameters.

Ideally the validation and training curve look similar as it is this case and the plot shows that the training and development scores do not change significantly with the number of trees the score seems to stabilise after the value of approximately 150 trees for the validation set.

Tree depth seems to affect the score values more, with the training score to improve when the tree is deeper as expected. However, the validation score seems to only slightly vary between the values of 10-18. Here the two curves are rather different with the training curve reaching high scores quickly which indicates overfitting.

Indicative Tree Visualisation

Since Random Forest is an assemble method, it does not really make sense to visualise the trees. However, visualising a few trees can give an indicative image and help at interpreting the results. We will plot just 2 trees.

The plots above show the split process of the 2 decision tree examples, where "education" is used in the first level split. However, to get a better picture of the importance of each feature we will use the Permutation Feature Importance below.

Interpretation: permutation feature importance

Using the rfpimp package, the importance of each variable can be computed. For this the testing set needs to be used (not used for model training) to avoid biased results.

XGB Boost

A GBDT (Gradient Boosting Decision Tree) model will be used and compared to the previous models.

The main difference to The random forest, is that in this ensemble learning the models are trained sequentially using the results of previous models as an input to the next (by assigning lower weight to correctly predicted outcomes and higher weight to the wrongly predicted outcomes)

The [XGBoost] package will be used for this

The XGB model seems to perform better than the previous models on the testing set, but this uncalibrated model does not perform well on the testing data

\*\*HYPERPARAMETER TUNING\*\*

Similarly to the above, two methods will be compared:

\* Holdout validation (grid search)

\* Cross validation (grid search)

The hyperparameters we will tune are:

\* ```n\_estimators ```(default=100): The number of decision trees.

Usually a higher number of trees improves the model's performance. However it is also slowing down the training process.

\* ```max\_depth``` (default=6): The depth (splits) of each tree

Usually the deeper the tree the more information it captures about the data. However allowing maximum depth might result in overfitting.

\* ```learning\_rate``` (default=0.3):

\* ```colsample\_bytree``` (default=1):

Holdout validation

Holdout validation will be used here, by specifying the 'cv' parameter (the cross validation splitting strategy) as an iterable yielding (train, test) as arrays of indices.

This time the hyperparameter tuning has slightly improved the model's performance on both the training and development set. However, the very high R2 on the training data indicates that the model is overfitting.

The two validation methods, pick the same values for all hyperparameters except n\_estimators (150 in cross-validation, 200 in holdout validation). The best score result is also almost exactly the same

We create the final model with the tuned hyperparameters:

The XGBoost Model tuned with Cross Validation, also returns slightly worse results, but the difference between training and development has been reduced which indicates there is less overfitting, and the model might have better generalization. We will therefore keep this model.

Let’s also plot the \*\*validation curve\*\* (also based on cross validation) for each parameter

In both above plots, the Training and Validation curves are quite different with the training curve increasing abruptly while the validation curve is lagging behind, a sign of model overfitting.

Using the test set (that has not been used at any stage during the model training andn validation process) the final estimate of the XGBoost model can be reported.

While the RF performs best in the training data, the XGB scores better in the test data and has a smaller R2 difference indicating that this model is less subject to variance.

However, the results are not good overall and also not significantly better than the baseline model which we know is not a good fit.

PCA

This section will test whether using Principal Component Analysis (PCA) will improve the prediction results. This dimensionality reduction method will create a set of Principal Components (new features) from out independent variables based on the highest variance.

Before proceeding, the data needs scaling so that no feature dominates the results.

Below we can print the makeup of each component. Each component is a linear combination of the previous independent variables.

This heatmap and the colour bar basically represent the correlation between the various feature and the principal component itself.

One of the uses of PCA is to visualise high-dimensional data. Here we can visualise the first two components, using the HPI as the colour.

There is revere overlap in the plots above and the interpretability is very low, which does not allow the validation of the components

Selecting the number of components

There are different methods for selecting the number of PCA components.

1. 2 or 3 PCs for visualisation

2. PC with eigenvalues > 1

3. Scree plot

While using 2 or 3 components will allow better visualisation, we will check the components with Eigenvalue larger than 1 and the Scree plot.

The plot above does not indicate one definite point as a transition point from a steep change to a gradual flattening. However, we could argue that the right values are 3-6 components: the plot decreases quickly before 3 and starts to flatten out after 6.

We can therefore go for the two ends PC=3 or PC=7

Random Forest after PCA

Using the test set(that has not been used at any stage during the model training and validation process) the final estimate of the RF model can be reported.

**Results/ Discussion**

-visualisations (labels, clear)

-interesting results and interpretation

This section summarises and interprets 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, the XGBoost model without PCA seems to be the slightly better 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 models, 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) which might explain why the performance has dropped.

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.

The compared methods and models do not perform well at predicting the change in HPI for the study area. Random Forest and XGBoost Regressors were chosen as they are considered relatively robust to outliers and can handle non-linear data, however there is a plethora of different algorithms and methods that have not been tested here and which might yield better results. Furthermore, the hyperparameter tuning performed in this analysis was certainly not exhaustive und could be optimised. It has been suggested that a better approach would be 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).

The poor performance of the models might also be traced back to the datasets themselves. The limitations already mentioned such as the significant number of missing values, the extreme outliers and even the number and nature of variables used, all might have strongly influenced the models’ performance.

Moreover, just as the traditional hedonic models, this analysis has not considered the spatial dimension and spatial autocorrelation which poses a significant limitation. Different spatial units might also lead to different results and need always to be considered in such studies.

**Conclusion 200**

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.

An extension of this work would investigate a different approach: restating the analysis as a classification methodology that will attempt the prediction of defined categories of HPI range, instead of individual values.

Finally, the prediction of the change in HPI solely from the above indicators can be scrutinised and considered incomplete since the housing market is most notably influenced by a range of factors such as the housing supply, speculative demand, interest rates and economic growth in general. It is therefore suggested that further research incorporates these and tests whether the inclusion of environmental, health and socioeconomic indicators improves the prediction performance.

**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.

Park, B. and Kwon Bae, J. (2015) “Using machine learning algorithms for housing price prediction: The case of Fairfax County, Virginia housing data,” *Expert Systems with Applications*, 42(6), pp. 2928–2934. doi:10.1016/j.eswa.2014.11.040.

*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.

for example, (Park and Kwon Bae, 2015) compared the performance of various classifiers and found that these methods can contribute significantly towards accurate house price predictions.

~~More recent studies make use of modern machine learning methods in their attempt to model house prices such as~~