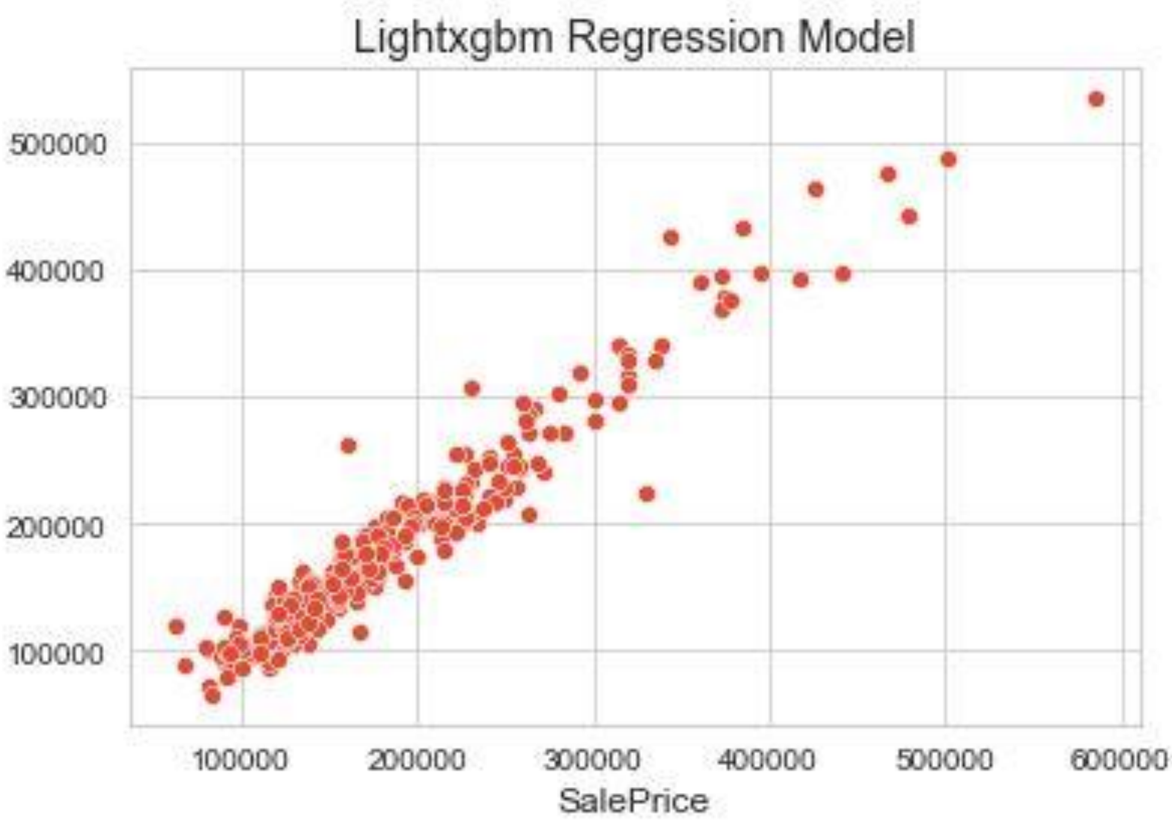
Ames, Iowa in Focus: Using Pertinent Economic Data to Predict Housing Sales Prices

Sonny Desai, Northwestern University

Contact: sjdesai87@gmail.com

Abstract

Making predictions of future realities based upon historical data is one of the main tools used in machine learning. This paper presents the findings of a project revolving around the Ames, Iowa dataset describing the sale of individual residential property in Ames, Iowa from 2006 to 2010. Using additional economic data, the project sets to predict future sales prices for these homes. This original dataset consisting of 2,930 observations and a large number of explanatory variables (23 nominal, 23 ordinal, 14 discrete, and 20 continuous), was expanded to include pertinent economic data for the years between 2006 to 2019. The prediction models will use the same economic features, but for the years 2011-2019, to predict 2020 housing sale prices in Ames, Iowa. This paper will show that lighgbm boost regressor performed the best of all prediction models and that accurate predictions can be found by expanding the original dataset to include pertinent economic features.



Introduction

Ames, Iowa is a largely suburban city in the middle of Iowa. The famous dataset created about its housing properties from 2006 to 2010 covers 1,500 home and their sales price. The objective of this project is to predict 2020 sales prices of the 1,500 Ames, Iowa homes using the original dataset coupled with select economic factors. The original Ames, Iowa data set has both continuous and non-continuous variables with 82 explanatory variables. The dataset contains 23 ordinal, 23 nominal, 22 continuous, and 14 discrete variables. The numerous continuous variables are data describing the metrics surrounding the description of the housing properties: including square feet, living area, number of rooms, baths, and kitchens. The non-continuous variables discuss more details about the housing and its area including neighborhoods, garage type, and sale type. This original dataset was expanded to include economic data for the years 2006-2019, beginning with the starting year of 2006 from the original dataset. Economic data based on their effect on housing prices for Ames were chosen based on general economic conditions as well as housing related economic variables. This expanded dataset was first used for the years 2006-2010 to create the initial dataset and test the models. These years were dropped, and the years 2011-2019 were add to create the dataset for the predictor variables. 2020 was select as the predictor year despite there being more current data, the Covid-19 Pandemic would have a large impact on bias, that it would be unviable to create the model with it.

Given this large amount of historical data, this project aimed to create models to help predict sales price of properties based on the expanded dataset. The models created were linear regression, lasso, ridge, elasticnet, support vector regression, light gradient boost regressor, xg gradient boost, as well as a scaled down version of the dataset with Principal Components Analysis (PCA). Regression models were chosen instead of classification, as the project was still using the same original dataset, but instead making predictions based on the updated economic data. PCA was also included in this project with the goal to produce better performing models and both version of the datasets were tested: the expanded dataset, and the PCA, component reduced version. The initial findings using the 2006-2010 expanded economic data were conclusive showing that lighgbm performed the best with a 88% accuracy rate. Both gradient boosting models performed better than the linear regression models suggesting the added weights had a positive impact on accuracy. The PCA reduced dataset was interestingly lower in result score for the highest performers without PCA, despite the testing with different component reductions.

Literature Review

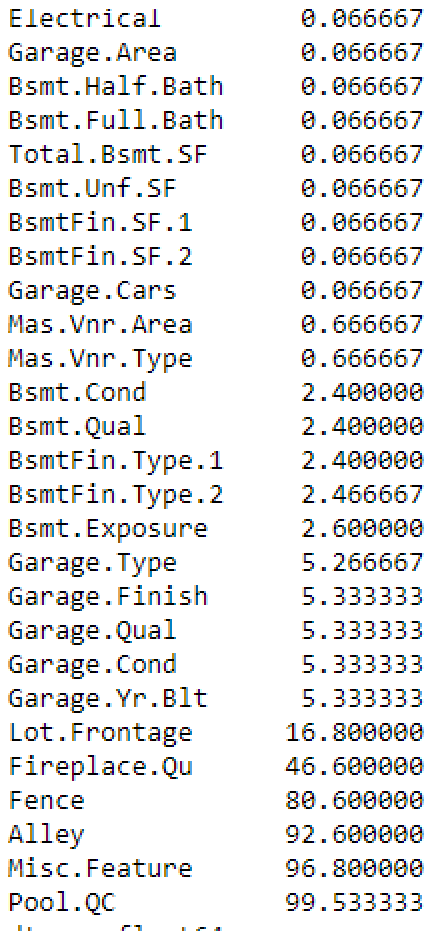
The Ames, Iowa Dataset as a standalone is full of useful insights, interesting correlations, and plenty of null and missing values to make any researcher wonder "What if?'. It is a mix of continuous and categorical variables all related about descriptive statistics around the homes and their sales price. When doing correlation findings, data transformation, and feature selection (which will be discussed more in depth in the below sections) it is clear there is a lot to work with from the original dataset. As noted by De Cock (2011) on the Ames, Iowa dataset, the ordinal variables present specific difficulties. He writes, "Almost all of these variables are quality related, with the expectation that higher categories should yield a coefficient at or above the previous category. In some of my initial modeling, I found that the estimated coefficients for a number of these categories did not follow this rule, likely due to interrelations with other variables within the model". Despite these concerns with the ordinal variables, De Cock would later note "I believe the data set has unlimited potential". The issues with the ordinal and discrete variables could cloud the results of any predictive model, but if the data was transformed properly, and the best features selected, the potential of the dataset was still open. One idea to add more conclusiveness to the dataset is to add additional data, which leads to the natural question "but what additional data?". As noted by Nagarajan, Ogwal, Yellajosyula, Jiang, Xu, & Choi (2020) on House Prices in Ames, Iowa, there are several features that could be created by adding together separate features from the dataset to create new features with higher correlation. Their results built on an OLS

Regression predictor model showed a R Squared score of .94, and fond that the new features (especially those around size) were highly influential to the success of the model. This approach would help resolve the issues of the confusing data, but the new data features would still revolve around the characteristics of the homes themselves. This project wanted to see if there were any external data sources, especially economic ones, that could be added to bring about better prediction results.

The idea of for external data being added to the Ames, Iowa dataset was broached in a report by Anthony Bellotti. Recent work by Bellotti (2018) on Reliable Region Predictions for Automated Valuation Models showed that using London Housing Price data, including economic data with the London Housing Price index resulted in a successful prediction model. The model in that project returned a 90% accuracy score, and although it was not as accurate as the original project (to be expected based on different geographic regions), the model was still highly accurate. Their model used a random forest regressor on the expanded dataset first created in their original project on the London Data, to find the results. Given the success of expanded datasets, the project in this paper used this idea to show a high scoring predictor model on the expanded dataset with similar economic features.

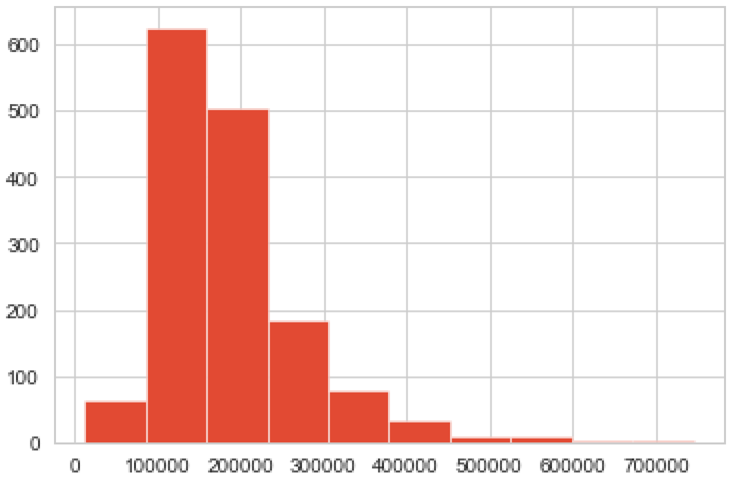
Data

The original Ames, Iowa dataset contains 83 columns of variables and records for 1,500 local homes. In performing exploratory data analysis (EDA) for this project, it became readily apparent that there were numerous variables with missing data. The list below shows the percent of null value data for the variables with null values:

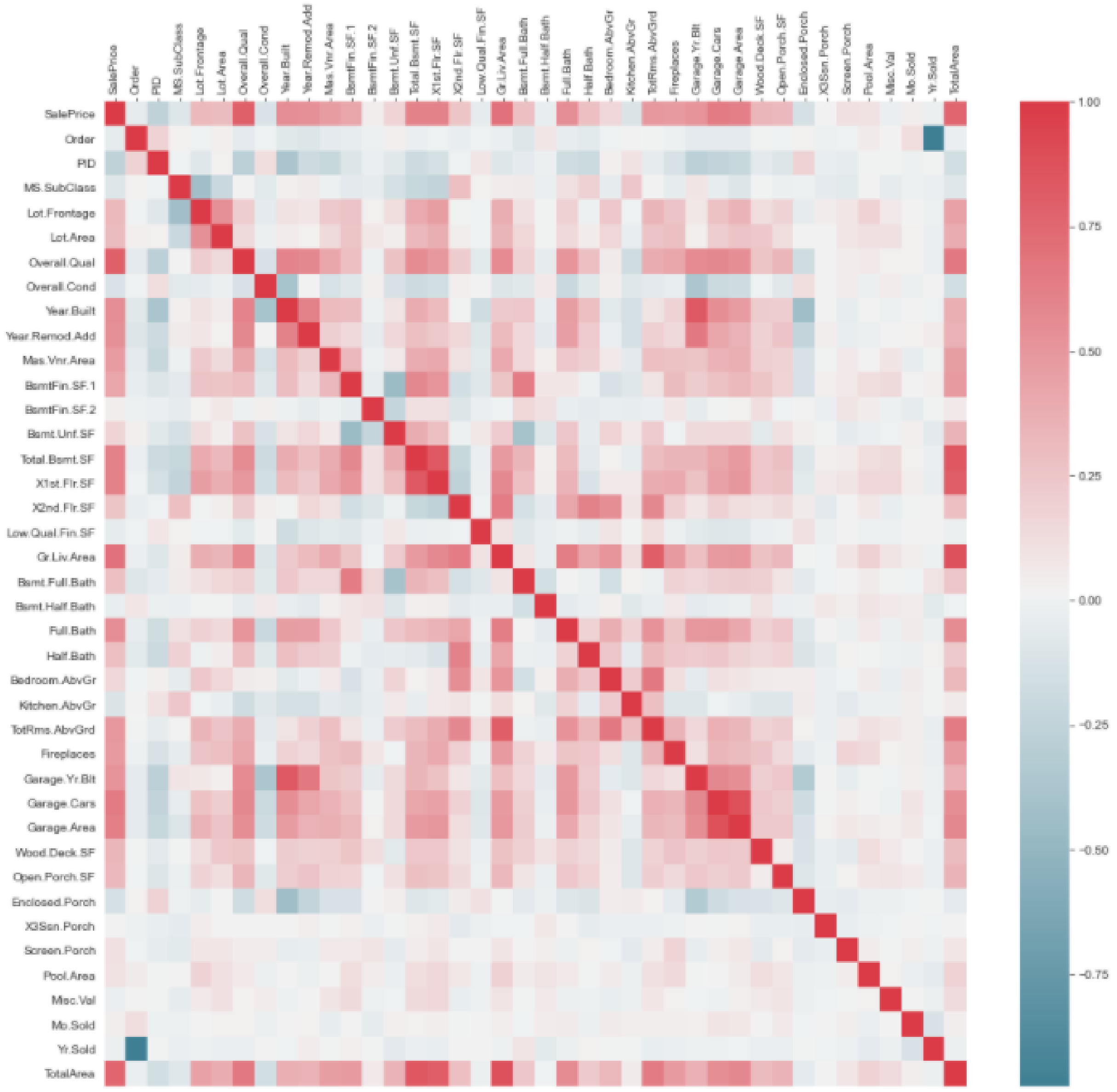


```
Electrical          0.066667
Garage.Area         0.066667
Bsmt.Half.Bath      0.066667
Bsmt.Full.Bath      0.066667
Total.Bsmt.SF       0.066667
Bsmt.Unf.SF         0.066667
BsmtFin.SF.1        0.066667
BsmtFin.SF.2        0.066667
Garage.Cars         0.066667
Mas.Vnr.Area        0.666667
Mas.Vnr.Type        0.666667
Bsmt.Cond           2.400000
Bsmt.Qual           2.400000
BsmtFin.Type.1      2.400000
BsmtFin.Type.2      2.466667
Bsmt.Exposure       2.600000
Garage.Type         5.266667
Garage.Finish       5.333333
Garage.Qual         5.333333
Garage.Cond         5.333333
Garage.Yr.Blt       5.333333
Lot.Frontage       16.800000
Fireplace.Qu       46.600000
Fence              80.600000
Alley              92.600000
Misc.Feature       96.800000
Pool.QC            99.533333
```

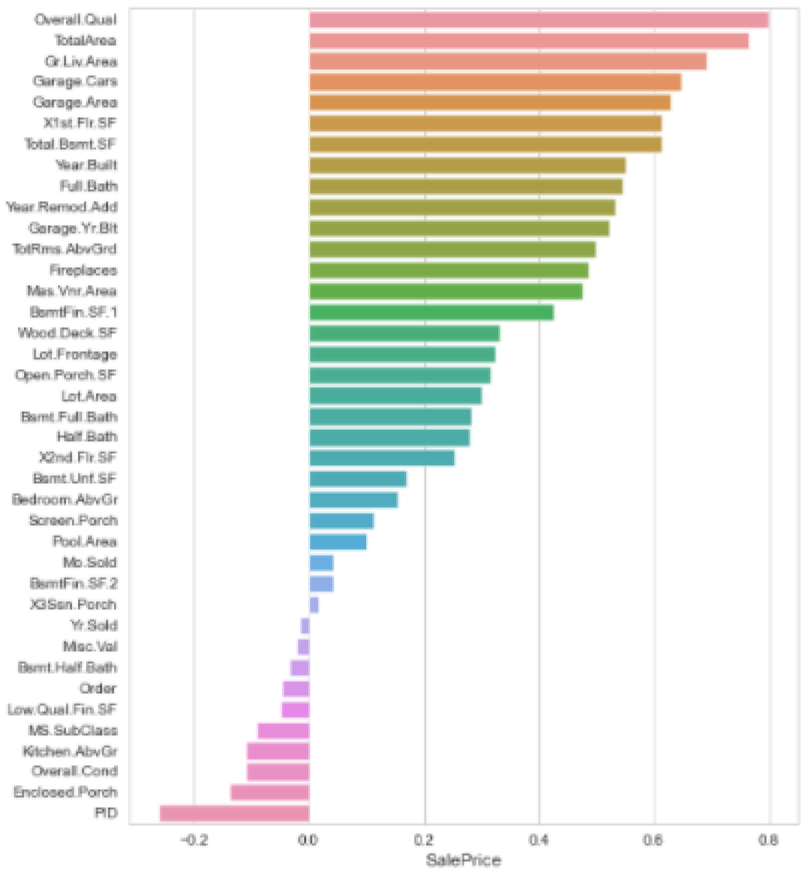
The issue became about what to do with the features with the missing data. For the features that had a percentage missing of greater than 40%, the variables were dropped. This was due to the fact that filling this missing data with a mean value would be too inaccurate. For the remaining data, the values were replaced with the mean value for the variable. The main feature to be predicted and the y variable in the machine learning models is sales price. The histogram below shows the amount of homes categorized by their sales price:

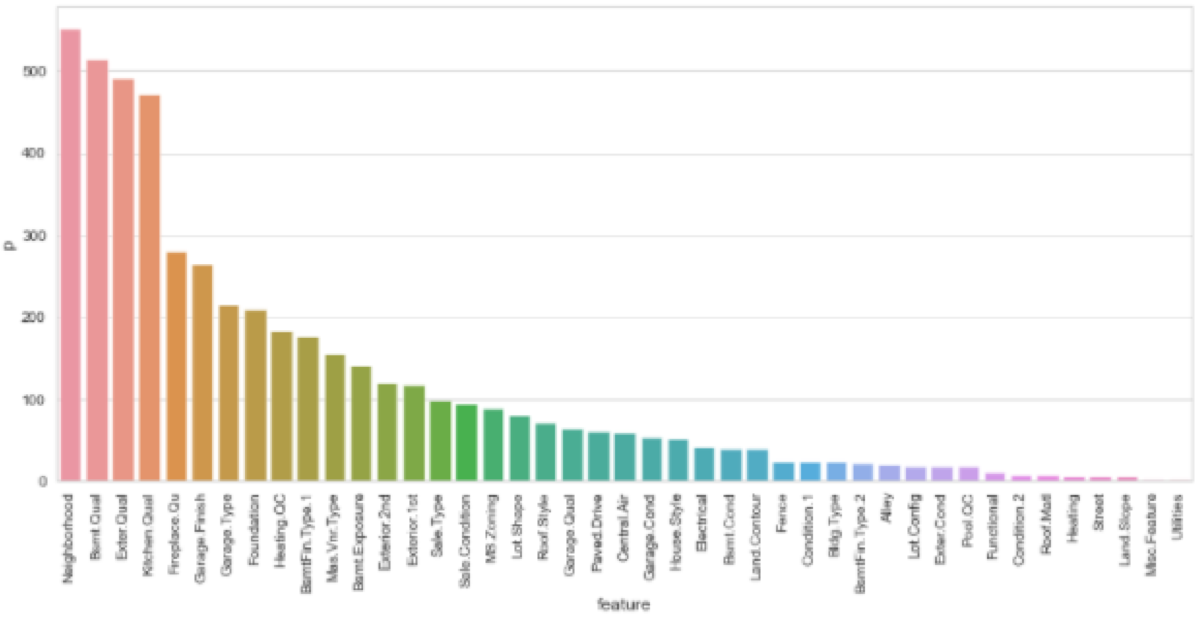


The plot shows that the majority of homes are in the 100 – 250k sales range, with a sharp drop off after 250k. To decide which features to select, a correlation matrix was created to view the relationships of the other variables in depth and in comparison to the response variable sales price:



To get more insights in the categorical and continuous variables, two separate lists were created with the features of each added into them separately. The following charts show the correlation first of the quantitative variables and then of the qualitative ones:


Overall.Qual
TotalArea
Gr.Liv.Area
Garage.Cars
Garage.Area
X1st.Flr.SF
Total.Bsmt.SF
Year.Built
Full.Bath
Year.Remod.Add
Garage.Yr.Blt
TotRms.AbvGrd
Fireplaces
Mas.Vnr.Area
BsmtFin.SF.1
Wood.Deck.SF
Lot.Frontage
Open.Porch.SF
Lot.Area
Bsmt.Full.Bath
Half.Bath
X2nd.Flr.SF
Bsmt.Unf.SF
Bedroom.AbvGr
Screen.Porch
Pool.Area
Mo.Sold
BsmtFin.SF.2
X3Ssn.Porch
Yr.Sold
Misc.Val
Bsmt.Half.Bath
Order
Low.Qual.Fin.SF
MS.SubClass
Kitchen.AbvGr
Overall.Cond
Enclosed.Porch
PID
-0.2
0.0
0.2
0.4
0.6
0.8
SalePrice


p
0
100
200
300
400
500
Neighborhood
Bsmt.Qual
Exter.Qual
Kitchen.Qual
Fireplace.Qu
Garage.Finish
Garage.Type
Foundation
Heating.QC
BsmtFin.Type.1
Mas.Vnr.Type
Bsmt.Exposure
Exterior.2nd
Exterior.1st
Sale.Type
Sale.Condition
MS.Zoning
Lot.Shape
Roof.Style
Garage.Qual
Paved.Drive
Central.Air
Garage.Cond
House.Style
Electrical
Bsmt.Cond
Land.Contour
Fence
Condition.1
Bldg.Type
BsmtFin.Type.2
Alley
Lot.Config
Exter.Cond
Pool.QC
Functional
Condition.2
Roof.Matl
Heating
Street
Land.Slope
Misc.Feature
Utilities
feature

The qualitative chart surprisingly shows that most of the features to have some correlation to the feature variable. The highest performing ones were overall quality, total area, and the garage features. The quantitative chart, in turn, shows a lot of features with a with very low correlation to the data. The highest performing features were Neighborhood and quality of the room types in the home. Instead of removing these low performing features, PCA was applied to reduce the number of dimensions and components with the regression models being tested again on the PCA reduced dataset. This was done to put mor emphasis on the economic data, and to allow all features, even those with low correlation, to have impact on the model. To deal with the qualitative data, dummy variables were created that turned these categorical variables into continuous ones by transposing the data in multiple new variables with a 1 for present data, and 0 for without. This ended the transformation of the original dataset.

Multiple variables were reviewed as potential features to add as the new economic features to be added to the original dataset. Health variables were also explored but because of the quality of the economic variables, they were not included. The five variables chosen were Population, Unemployment Rate, Median Household Income, Housing Price index, and Property Tax valuations. Population was chosen for its affect on demand, with a larger population needing a larger housing supply. Unemployment rate and median household income were both selected for its economic value, with each variable suggesting the potential buying power of the residents. Housing price index was a natural feature because of its close link to supply, demand, and property valuations of the homes in Ames. The last value, was the only variable that data could not be found for Ames, Iowa specifically. Instead, property valuations for the entire state of Iowa were chosen because of its economic affect. Population and Median Household Income data was gathered from the Census Bureau database, Unemployment Rate was gathered from the U.S. Bureau of Labor and Statistics, Housing Price Index was gathered from the St. Louis Fed, and Property Tax Valuation was gathered from the Iowa Department of Revenue.

Lastly, two datasets were created, one fore the data with the years 2006-2010, and the years 2011-2019, and finally the data was split into a train-test split with test data being 20%.

Research Design and Modeling Method

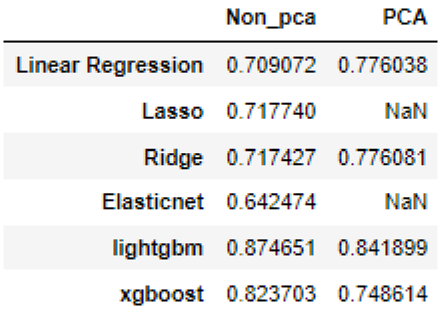
Multiple regression models were created and tested on the expanded dataset. PCA was applied and some of the regression models were again tested on the PCA reduced data. Regression models were chosen because the project is trying to approximate an unknown function. Classification on the other hand, would have required a randomized dataset using the Ames, Iowa data features. This, regression was chosen as the new economic features would the main emphasis on predicting the unknown values of future sales price. The first model tested was a Linear regression model without additions or specific weights. The next two models tested were Lasso and Ridge, which have the benefit of using the math of linear regression models but with techniques to reduce model complexity and prevent over-fitting. Due to the built-in techniques, no further weights were added. For the elasticnet model, a random_state was added and set to one, which took one variable and randomly updated. Only one was chosen for the random_stte after testing showed this produced the highest scoring result. Next was a support vector regression model which added a standard scaler to the data which removed the mean and scaled to unit variance, and important part of removing variance form the model. After much testing C, which is the regularization parameter was set to 999999, and epsilon, which specifies the epsilon-tube within which no penalty is associated in the training loss function with points predicted within a distance epsilon from the actual value, was set to 1. The next model was the highest scoring model, lighgbmboost. Here, the objective was set to regression, and four leaves were chosen. The lighgbm is based on decision tree algorithms, it splits the tree leaf wise with the best fit whereas other boosting algorithms split the tree depth wise or level wise rather than leaf-wise. So when growing on the same leaf in Light GBM, the leaf-wise algorithm can reduce more loss than the level-wise algorithm and hence results in much better accuracy which can rarely be achieved by any of the existing boosting algorithms. Learning rate which has to be less than 1, was set to .1 which was the advice given for higher accuracy. Number of estimators was set to 50,000; max_bin, the max number of bins to bucket the feature values was set to 2000000; bagging_fraction, specifies the fraction of data to be used for each iteration and is generally used to speed up the training and avoid overfitting, was set to .5555; baggin_freq was set to 1; baggin_seed was set to 5; feature_fraction, specifies the fraction of features to be taken for each iteration, was set to .99; feature_fraction_seed was set to 100; verbose was set to 5.

The last model tested was th xgboost, a very similar model to the lightgbm model. Xgboost is slower, and considered to be less accurate than lightgbm. The parameters for this model were objective set to linear regression; colsample_bytree, the subsample ratio of columns when constructing each tree. Subsampling occurs once for every tree constructed and the parameter was set to .9; learning_rate which helps reduce overfitting by shrinking the feature weights to make the boosting process more conservative, was set to .9; max depth, the maximum depth of a tree, with more depth increasing the complexity of the model, was set to 500. This low amount paired with a high learning-rate were both used to check against overfitting. Alpha, regularization term on weights, was set to 100 to make the model more conservative. Lastly, n_estimators was set to 1000.

PCA and standard scaler were also applied to the dataset and again tested with some of the regression models. As mentioned previously, standard scaler removed the mean and scaled to unit variance. PCA reduces the amount of components and dimensions of the data to results in a smaller dataset.

Results

The results showed that the models with gradient boosting performed the best, while the linear regression models without weights or PCA performed the worst. The Linear regression model was the worst performing model, with the lasso and ridge performing better. The following table shows the performance:



	Non_pca	PCA
Linear Regression	0.709072	0.776038
Lasso	0.717740	NaN
Ridge	0.717427	0.776081
Elasticnet	0.642474	NaN
lightgbm	0.874651	0.841899
xgboost	0.823703	0.748614

PCA was also applied to the dataset and several models were run with the reduced dimension data. Interestingly, the models that performed poorly without PCA data, performed higher with PCA data, whilst the models that performed highly without PCA data performed poorly with PCA data. These were mainly the gradient booster models which had the highest scores without PCA but performed poorly with PCA. Linear regression, and Ridge performed better with the PCA data.

Analysis & Interpretation

Conclusions

Directions For Future Work

References

Geron, Aurelie. *Hands-On Machine Learning with Scikit-Learn, Keras & Tensorflow*. Canada: O'Reilly Media, Inc., 2019

Bellotti, Anthony. "Reliable Region Predictions for Automated Valuation Models" *Annals of Mathematics and Artificial Intelligence"* no. 81 (October 2017): 71-84. https://doi.org/10.1007/s10472-016-9534-6

De Cock, Dean. “Ames, Iowa: Alternative to the Boston Housing Data as an End of Semester Regression Project” Journal of Statistics Education 19, no.3 (2011). https://doi.org/10.1080/10691898.2011.11889627

Nagarajan, Ogwal, Yellajosyula, Jiang, Xu, & Steven Choi. “Grp12 Report – House Prices in Ames, Iowa”. . https://doi.org/10.13140/RG.2.2.23589.52967

U.S. Bureau of Labor Statistics. “Economy At A Glance”. Last modified August 26, 2021. https://www.bls.gov/eag/eag.ia_ames_msa.htm

Xu, Tu, Isabelle Guyon. “AutoML Meets Time Series Regression Design and Analysis of the AutoSeries Challenge”. arXiv:2107.13186 [cs.LG] (July 21, 2021). https://arxiv.org/pdf/2107.13186.pdf