Final Report: DS 320 Term Project

Housing Prices Prediction Model

Team:
Daniel Chadourne (dac5928@psu.edu),
Jess Strait (jls7571@psu.edu),
Cayla Pun (csp19@psu.edu),
Josiah Kim (juk483@psu.edu),
Julia Wurzel (jkw5638@psu.edu)

Objective

This project will focus on utilizing several federal and commercial data sources and features to predict house prices. The team will use the information from various data sources while focusing on ensuring the integrity and consistency of the data used to predict the housing prices. With the rise of the COVID-19 pandemic, affordable housing has become highly sought after, yet difficult for consumers to identify. Therefore, the team is interested in applying various data models, integration, and analysis to predict housing prices based on socioeconomic and property-specific factors.

Data Sources

The data sources of interest include two federal and one commercial dataset. Our primary source, Zillow, is an online real estate marketplace founded in 2006. In addition to listing and managing properties, Zillow Group seeks to provide “timely and accurate” data house valuation data (“Zillow Research”). This data will primarily focus on house attributes and will be integrated with federally aggregated data from the United States Census Bureau and Federal Emergency Management Agency in order to include more economic and social factors. Combining these datasets together through data integration, we will begin to provide unique house-price predictions that include a broad variety of cost factors.

The primary dataset including the target variable of house price will be derived from Zillow Research, made freely available on the data-sharing platform Kaggle (Mooney). Secondary datasets to build a stronger model for the prediction of house prices in a given geographical region will include socioeconomic factors- specifically, federal data pertaining to census attributes such as racial diversity, income, and poverty rates, and Federal Emergency Management Agency public assistance grant funding. Public assistance grants are federally obligated funding that can be requested by county and state governments to assuage the aftereffects of disasters affecting a locality. Such disasters can include natural disasters like hurricanes or tornadoes, but public assistance funding can cover any emergency that significantly damages the locality. The funding can be used for things like debris removal, reconstruction of roads and public buildings, and emergency protective measures.

Exploratory Data Analysis (EDA)

The EDA for the two federal datasets was conducted following the integration tasks (outlined below). Key findings from the federal dataset EDA was identifying which states had the most public assistance funding since 2000, and which had the most funding per capita. New York and Louisiana had the highest overall amount of funding; given that the selected time period included disasters such as the 9/11 attacks in New York City in 2001 and the devastation of Hurricane Katrina in Louisiana in 2008, these findings are consistent with domain knowledge. Louisiana also had the highest amount of per capita funding, possibly explained by increased rates of hurricanes and tropical storms in Louisiana in recent years with the rise of climate change-related disasters. Louisiana was followed in the per capita rankings by Texas, Kansas, and Mississippi, again consistent with domain knowledge on where the United States experiences hurricanes and tropical storms, tornadoes, and other disasters.

One additional interesting finding from the EDA for the federal datasets was through the development of correlation matrices and choropleth maps about census and FEMA data through geographically-based visualizations. In this report, we take California as a case study example from the EDA. By creating county choropleth maps (visible in the federal dataset notebook on this website) of public assistance funding per county, diversity, and median income per county, the team observed visually inverted color schemes between public assistance funding and the diversity/income maps. Developing correlation heatmaps suggested a negative relationship between public assistance funding and these two variables; but interestingly, subsequent correlation testing showed that diversity was a stronger predictor of less public assistance funding than income. Though outside the scope of this project, this relationship suggests that need for further research on if counties with higher concentrations of racially diverse populations experience disproportionately low amounts of federally obligated grant funding in the wake of disasters.

Data Integration

The first data integration task is to clean and combine the federal datasets from the United States Census Bureau and the Federal Emergency Management Agency (FEMA). Two distinct datasets from FEMA are used and simply integrated by joining on a key value; the primary dataset is public assistance grant funding data, and the secondary dataset is the records of the disasters occurring in a designated area (i.e. a county). The joined FEMA dataset is cleaned to limit the time series data to the year 2000 and later. Mutations of interest for the FEMA dataset include counting the unique number of disaster declarations for a county and frequency per capita and calculating the total obligated public assistance funding for a county along with the per capita total obligated. The census dataset already includes most variables of interest cleaned, but one additional mutation is run to generate a “diversity” variable which sums the percent of self-reported racially diverse population categories in the county. Notably, even though both the census and FEMA datasets are generated and managed by federal government entities, schema heterogeneity poses a challenge; similarly, county and state name syntax differences pose a challenge for a simple join. The census and FEMA datasets can be integrated quite easily using the county and state names with help from a uniquely identifying number known as a FIPS code, or Federal Information Processing System Code.

The original planned methodology for this project involved concatenating county and state names into one variable, and leveraging a string matching technique to appropriately join the two federal datasets even with the challenge of a heterogeneous syntax across the values. However, through research and developing a domain knowledge of emergency management, the team discovered FIPS codes and a fips() function in R that leverages an additional robust database that matches county and state names to their unique 5-digit FIPS code. A simple for loop generated FIPS codes for every instance and not only made integration effective for the federal datasets, but also opened up the possibility for mapping as part of the exploratory data analysis. One drawback to the FIPS code approach is that the FIPS code database in R is limited only to counties in the United States; thus, the state of Louisiana, which uses parishes instead of counties, was unable to be mapped. After weighing the options, the team determined that excluding Louisiana from state-specific analysis and mapping during the EDA was acceptable as long as public assistance and census data was still available from the state for the integration with the commercial dataset.

The second task to prepare for integrating the various data sources is to clean and preprocess the Zillow Dataset. Zillow offers many data files all containing information separated into either city or state. For our paper, we decided to use the city data files because it also gives us state information. Each file focused on a different metric, either sale prices for homes in each city, the median rental price for a 3 bedroom home, the Zillow Home Value Index for all homes in the city, and more. We mainly analyzed the sale prices for homes in each city and the Zillow Home Value Index data files. The first step in cleaning this were to change the formatting of the date column names. After changing the column names, we concatenated the two files (Zillow Home Value Index for all homes, and Sale Prices) and filtered it to only return the date values from after 1999 so that we can analyze the decades from 2000 to 2020. We then ensured that the state values were all in the same format by mapping the state to its respective abbreviation. We then filled any null values with the mean of each row (county).

The third data integration task was to integrate the cleaned and combined federal dataset (FEMA-Census) with the Zillow data on house prices and related attributes. The integration of these two datasets was fairly straightforward. There was first some further cleaning necessary on the Zillow dataset. There were a few cases where the county feature contained non-county (city) names, so those needed to be removed. Also, all the county names had “County” appended to all of the values which were removed. The Zillow dataset was finally able to be left-joined onto the FEMA-Census dataset which resulted in the final, model-ready dataset.

Modeling

Using our now integrated dataset and the individually merged datasets, we were finally in a position to begin testing our research question; Would predictions on an integrated data set be more accurate than those made using individualized data sets? This would be answered through the use of machine learning models and performance metrics.

In the case of our project, we as a group, decided that the first step to creating our model would be to decide upon the accuracy metric. We initially thought to utilize accuracy, root-mean-squared-error (RMSE), and precision. However, as our project was not focused on classification but rather predicting values, we chose RMSE to serve as our evaluation metric. Additionally, since our project was not to predict classifications, we were limited by the type of machine learning models that could be utilized.

As a result of this limitation, we chose to put most of our focus on linear regression. We did discuss utilizing an xgboost model, we decided against the idea as we were more focused on creating generalized predictions and not discovering the optimal value for each prediction. Choosing to discard xgboost also allowed us to reduce the time and resources that training said model would have consumed. After narrowing down the metric and model to be used, we then needed to create our test and train sets, as well as a validation set to be used to evaluate our predictions. As we had many “cost” variables that could be modeled, we chose to focus on the most general “price” variable in our integrated data set, “City_Zhvi_AllHomes.” This generalized “price” measure would become the primary feature that we would be modeling.

With the full dataset in hand, our first step was to standardize. After standardizing our dataset, we split our data into test, train, and validation sets using a 80/10/10 split which is an industry standard.

Summary & Results

After our modeling, we were prepared to answer our research question; would the integrated data set yield better predictions than the individual datasets? Based on our RMSE scores for each dataset, we believe the answer to be “yes”.

When taking the RMSE of each datasets’ predictions, we were essentially measuring the level of variance contained in each. In machine learning, an optimal model is one that encapsulates the highest amount of variance in the given data. While this is not evident when using RMSE as an evaluation metric, it is the basis of why we are able to answer our research question in such a way.

As outlined in Figure 1, the RMSE for the individual datasets (in dollars) was $1113 and $202344 for the Zillow and FEMA-Census datasets respectively. Understanding that our predictions are based on prices in the hundreds-of-thousands, it would appear that the FEMA-Census data does not have any strong internal relations or correlations, and that the vice versa is true for the Zillow data set. In comparison, our integrated model scored an RMSE of $359. This is incredibly significant, especially when compared to size of the values present in the original datasets. Taking one step backward, our mean squared error for the integrated dataset was approximately $128,881. When a simple percentage is calculated, an RMSE of $359 represents less than 3% of the mean squared error. This extremely low error suggests that our integrated data set and the features contained within, were able to capture more variance than those in the individual datasets.

Data Set	RMSE
Zillow	1113
FEMA-Census	202344
Integrated Dataset	359

Figure 1: The RMSE of predictions created using the integrated and individual datasets

It is clear that when data integration and fusion is used to create a merged dataset, it is more useful for research purposes, and for generating predictions. As more features are added, machine learning models become capable of “learning” more variance. In this case it appears that our integrated data set gathered together features that were highly correlated/highly connected, thus increasing the amount of variance that could be learned. As stated previously, we believe that the answer to our research question is most certainly “yes.”

Discussion

The data integration and model prediction used throughout this project provided better predictions of housing prices. However, the project can be improved even further. One of the main improvements is integrating additional data into our model. By integrating additional data that provides different features to the Zillow dataset, would work to further advance the accuracy of the model. This is due to the model having more features to be trained on. Another improvement is expanding our dataset and model to include various countries. Currently our model is built for the United States. Therefore, including data from other countries will expand the usability of the model. This may also help future homeowners identify cheaper options throughout various countries. An additional suggestion for future work is to focus on building a more robust data model. One way to accomplish this is removing any outliers from the data. For example, housing price outliers would be removed prior to the model being trained. This would result in the model producing more robust predictions. The project explained throughout this report clearly creates a powerful foundation that allows future research, as well as changes, to be conducted.

Works Cited

“Census Bureau Data.” United States Census Bureau. Accessed 26 Oct 2021. http://data.census.gov

Mooney, Paul. “Zillow House Price Data.” Published 7 Dec 2020. Accessed 26 Oct 2021. https://www.kaggle.com/paultimothymooney/zillow-house-price-data

“Public Assistance Funded Project Details.” Federal Emergency Management Agency. Accessed 26 Oct 2021.
www.fema.gov/openfema-data-page/public-assistance-funded-projects-details-v1
“Zillow Research.” Zillow, Inc. Accessed 26 Oct 2021.
https://www.zillow.com/research/