Term Project CSCI 4360, Housing Market Forecasting

Ayush Kumar, Faisal Hossain, Brandon Amirouche

April 27, 2021

Contents

1 Problem Statement: Forecasting Housing Market **2**

2 The Datasets **2**

3 Preprocessing Techniques **2**
3.1 Zillow Data 2
3.2 HMDA Data 2

4 Exploratory Data Analysis **3**

5 Modeling **4**
5.1 Auto-Regressive Model 4
5.2 Seasonal Auto-Regressive Integrated Moving Average 4
5.3 Gated Recurrent Unit RNN 4
5.4 Long Short-Term Memory 4
5.5 Results 5

6 Results & Conclusion **5**

7 Recommendations of Study **5**

1 Problem Statement: Forecasting Housing Market

The goal of our project was to determine if future housing market pricing can be explained by previous housing market data. We also wanted analyze the usage of exogenous variables in our time-series modeling, and whether it makes a difference. This task is distinct from housing market prediction, which relies on the features of the market, and the house. Using time-series data presented our group a new and novel challenge.

2 The Datasets

We identified two datasets for potential use in our projects.

1. Quandl Data from Zillow
2. USA Home Mortgage Disclosure Act Data

Each dataset has very different forms, and presented different challenges. The Housing Mortgage Disclosure Act was passed by congress in 1975, and requires multiple federal agencies to keep track of all mortgage loans filed for and denied. The datasets that we are using from the HMDA contain all mortgages filed for in America for a given year. This data is very large, and we decided to use only a subset of the the available data from 2014-2017. These 4 datasets combined were around 40GB of data, and so we required a different approach for data analysis. We got around the large nature of the data using an SQLite3 database and querying it for EDA using the corresponding sqlite3 python library.

The Zillow data comes from Quandl, a web api for many types of time-series data. This data has over 70,000 regions and corresponds to years of aggregated monthly home sales. For the purposes of this project we downloaded 43 different regional data, mostly from the state of Washington. These different regions will be used for testing our models, and we will try to compare them to one another. A key drawback of this approach is that some regions have data that goes further back than other regions, as long as the future housing values aren't affected by values from further in the past than 10+ years the results should be directly comparable.

We will decide on which datasets to use for our modeling after taking into account the preprocessing that we will need, and doing some exploratory data analysis.

3 Preprocessing Techniques

3.1 Zillow Data

Different models will require different data manipulation techniques, but the Quandl curates clean easy to use datasets. For this reason all our group had to do was take out the required columns, and cast them to the correct types. The columns we used for our analysis were:

1. date - the date of the observation, always the 1st of the month
2. value - the average home sell value in that particular location

We simply cast the date to a numpy date-time object and left the values as were. For the RNN models we did rescale the data using a MinMaxScaler, but we will cover that more in depth in the modeling section of the report.

3.2 HMDA Data

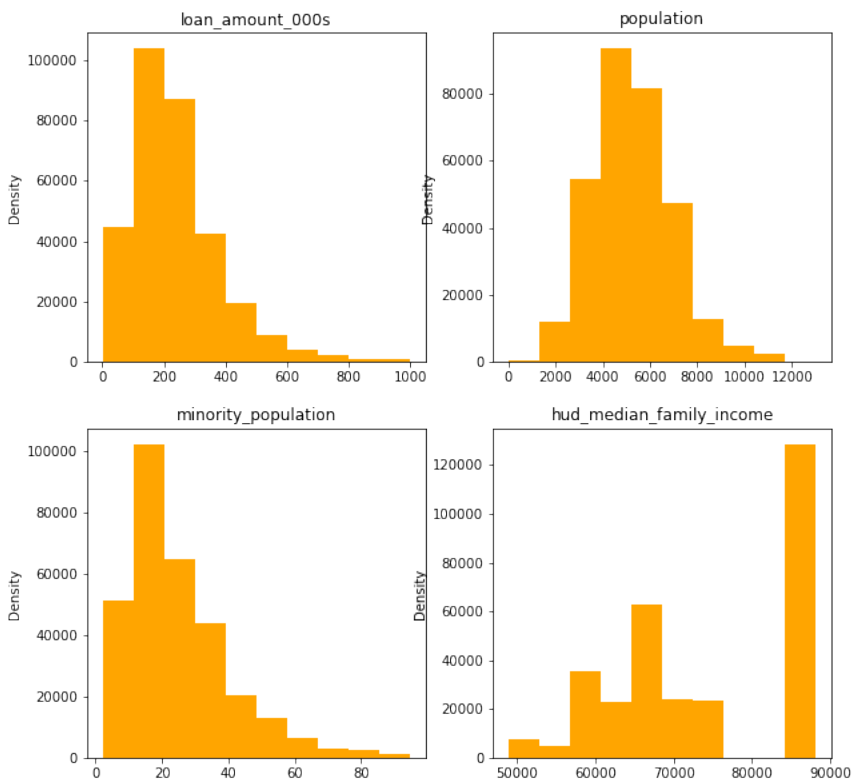
The HMDA data is much dirtier and much more complicated to use. We applied the following filters before undergoing any sort of EDA.

1. Property Type - limited to one-to-four family dwellings
2. Approved Mortgages - Houses were actually sold
3. loan purpose - only home purchase
4. state - Washington State Only because of Zillow regions

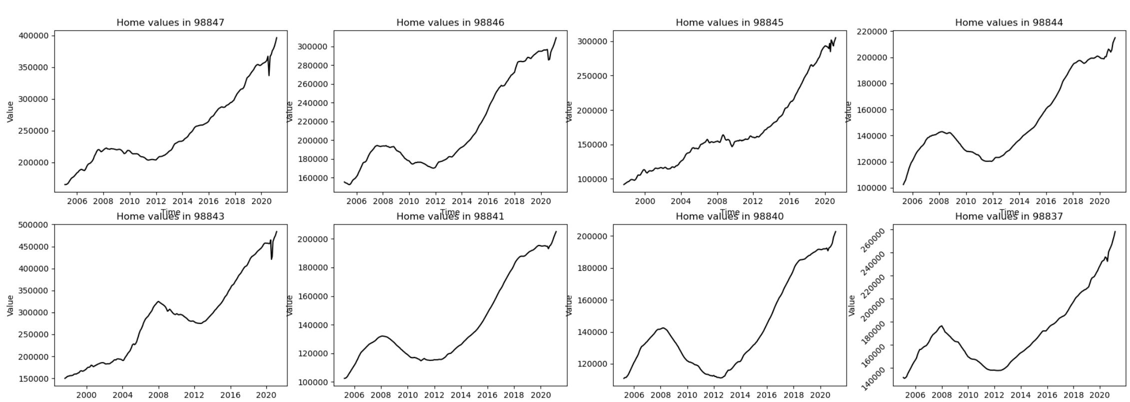
Since we will be focusing on aggregated data we will simply be ignoring data that was missing. We aggregated the following columns for further analysis, as they were based on characteristics of the region overall rather than specific to a particular mortgage application.

1. Loan Amount (Thousands)
2. Population
3. Minority Population (Expressed as a percentage)
4. HUD Median Family Income

4 Exploratory Data Analysis



The histograms are all very similarly shaped, and may be highly correlated. In particular the loan amount, population, and minority populations seem to have very similar skewed right distributions. However, the data we have access to is yearly in nature, and the variations in the Zillow data are monthly. For this reason we will only be using the Zillow data to for our time-series modeling.



As we can see from this sample of

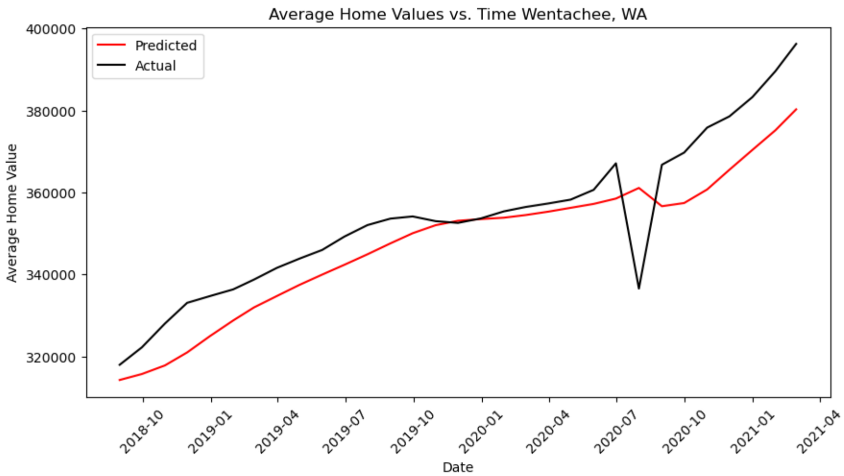
5 Modeling

5.1 Auto-Regressive Model

5.2 Seasonal Auto-Regressive Integrated Moving Average

5.3 Gated Recurrent Unit RNN

The Gated Recurrent unit model requires inputs to be reshaped in a 3 dimensional input, where a particular value of at time t_k will be backed by m values ranging from t_{k-m} to t_{k-1} for every observation in the dataset. We used a GRU with 64 recurrent units and the data was reshaped to use a look-back of 8 months as we found that was ideal.



Here were the results of the model:

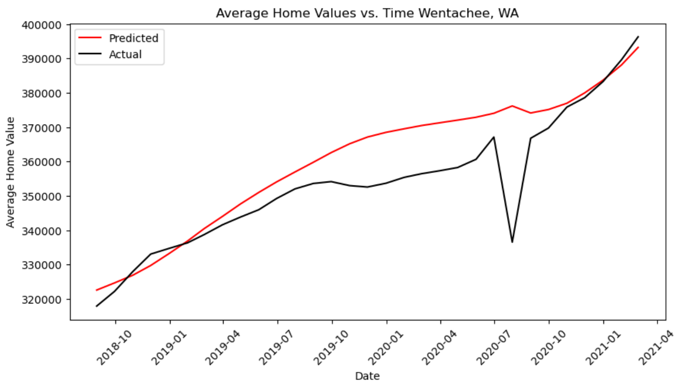
- R^2 - 0.69
- $RMSE$ - 9449.19
- $MAPE$ - 0.022

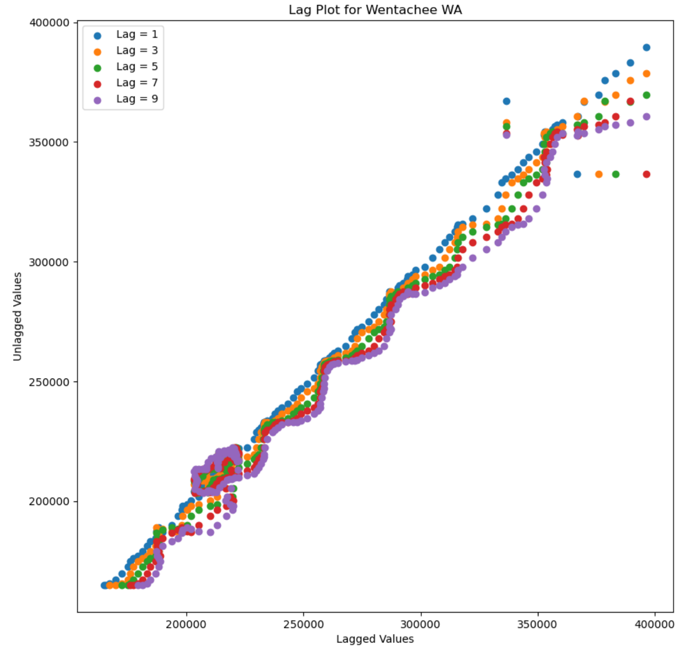
The GRU overall performed very well, and if wasn't for the black swan event that was the pandemic, the model would have performed even better overall. We can see that the model is affected by the pandemic observation months after it happens because of the nature of the time series. We could get around this problem by using a shorter look-back, but this may make the model too sensitive to observations that come right before.

5.4 Long Short-Term Memory

LSTM uses silimar shaped data as GRU, and maintains a control flow similar as an RNN. LSTM processes the data continuously as it grows forward, while the data analyst is allowed to keep or forget the previous code. The analyst may choose to use the previous data through operations that can allow the LSTM to keep or forget information as an instruction.

Graphs





5.5 Results

The LSTM gives us an accurate prediction that is slightly better than others. The R-Squared is the highest given of all three models (not including the autoregression). While the GRU model tended to underestimate, or rather undershoot the actual, the LSTM actually overshot it. The LSTM predicted higher numbers than the GRU, however both could not account for the dip in the housing market that we believe was caused by the coronavirus pandemic. This is a reasonable assumption since these models cannot predict real world issues and events, only how a model should perform under the same previous circumstances. The lag plot given was also similar to the GRU, and gave farely accurate measurements as well.

6 Results & Conclusion

Based on the results of our models, we believe that the best models in this case are the very simple auto-regressive models simply because of the high performance and the simplistic nature of the data. We were able to accurately predict the future housing prices solely on the basis of the previous values, and exogenous variables were not necessary.