CSCI 4360 Data Science II Term Project Proposal
Project Team I

Ayush Kumar, Faisal Hossain, Brandon Amirouche

April 04, 2021

For our term project we would like to focus on using housing data to perform predictive analysis on future housing prices. We have two main data sources for this:

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

We plan on using this and more to predict housing prices in the United States using the following predictive models:

1. Traditional Regression Modeling
2. Transformed Regression Modeling
3. Feed Forward Neural Networks
4. Time Series Modeling

We will also be using time series models, and panel data and comparing and contrasting with traditional models. The test dataset will be the HMDA data from 2018 as it is the latest dataset and provides a testing ground for a model. We also plan on doing rolling modeling to show which data can be used to make accurate predictions over time. The HMDA datasets have over 10 million observations each and pose one of the largest data challenges that any of us have previously faced.

We will also be using batch processing, and learning techniques associated with big data because we simply cannot hold all the data in memory during the training sequence. We cannot hold all the data in memory simply because we are using 4 HMDA Datasets each of which when unzipped is 9-10 GB in size. This would total to around 40GB of RAM used when using Pandas or Base R statistical tools, and none of us have this much memory. To get around this we will be using a variety of tools.

We will be using the following tools for our project:

1. SQLLite3 - for holding all the HMDA Data in a convenient database that can be used to stream data given our memory restrictions
2. R - for basic regression analysis and variable selection
3. Keras - for neural network modeling
4. Apache Spark - for making the most efficient use of our systems & practice for the real world