HW3 Report

1. Build a Scoring function to rank houses by "desirability"

Analytical Hierarchical Processing

In this question, we used Analytic Hierarchy Process (AHP) to compute a score of the desirability of the parcels. In AHP the problem is broken into a hierarchy of easy-to-comprehend sub-problems. In this model we used the following metrics:

1. Actual Area: Total area of the house including the area of the floors
2. Pool Count: Number of pools in the house
3. Latitude & Longitude: Coordinates of the house
4. Unit Count: Number of units in the house
5. Lot Size: Area of the land
6. Bedroom and Bathroom count: Number of bathrooms and bedrooms
7. Hot Tub or Spa Present: If a hot tub or spa is present or not
8. Fireplace present: If a fireplace is present
9. Tax Value: The total tax value of the parcel
10. Building Type
11. Garage count: Number of cars that can fit in the garage
12. Age: Age of the parcel
13. Tax Amount: Total property tax assessed for that year
14. Air Conditioning Type: If the parcel is air conditioned or not
15. Region Id County: County of the parcel

In AHP, we give a relative importance to each of the above mentioned parameters as compared to the other parameters. Once the hierarchy is built, we evaluate its elements by comparing them to each other two elements at a time. For example, consider the example of Actual Area vs Pool Count. We think that Actual Area is 7 times more important than the pool count i.e. according to us having a bigger actual area of the property is more important than having a pool in the property.
The full list of comparisons is available in the distance_metric.csv file committed to the repository. After this table is ready with policy scores and attribute weights, we can apply Simple Additive Weighting (SAW) and Weighted Product Model (WPM) to get to a decision.

Simple Additive Weighting (SAW)

In general, suppose that a given MCDA problem is defined on m alternatives and n decision criteria. Furthermore, let us assume that all the criteria are benefit criteria, that is, the higher the values are, the better it is. Next suppose that wj denotes the relative weight of importance of the criterion Cj and aij is the performance value of alternative Ai when it is evaluated in terms of criterion Cj. Then, the total (i.e., when all the criteria are considered simultaneously) importance of alternative Ai, denoted as Ai WSM-score, is defined as follows:

SAW decision values:
P1 = 6.019
P2 = 5.3
P3 = 5.78

Final Decision: Using SAW, we found P1 to be the best Policy to be opted

When we applied this method to find the best and the worst parcel, we found the results to be in conformance with our expectation. For example, we consider the actual area to be one of the most important factor. In the results also we found that the area of the best house is 5 times more than the worst house.Similarly, we expected that the more tax amount would decrease the desirability of the house and this is exactly what we observed in our results. The tax amount of the most desirable house was almost half of that of the least desirable house.

2. Pairwise Distance Function

Important Features

Before we dive into calculating the distance between two properties, we need to first find out what parameters we think are important in order to find out which properties are more similar to each other.

After evaluation of the data available and considering real life logic, we have selected the following columns to find out the
distance between two properties:

1. actual_area
2. poolcnt
3. latitude
4. longitude
5. unitcnt
6. lotsizesquarefeet
7. bedroomcnt
8. calculatedbathnbr
9. hashottuborspa
10. fireplacecnt
11. taxvaluedollarcnt
12. buildingqualitytypeid
13. garagecarcnt
14. age
15. taxamount
16. airconditioningtypeid
17. regionidcounty

Relative importance of features

While we consider the above columns to calculate the distance between two values, it is only logical to assume that not all these columns are equally important while calculating the distance between the two properties. For example, whether the two properties belong to the same county is more important of a factor compared to a difference in the number of fireplaces available in the house.

We use Analytic Hierarchy Process (AHP) to calculate the weight vector associated with different features. The process to do so is described in an earlier part of this document.

We will be using a **weighted Euclidean distance** as a distance metric to calculate the distance between two properties. We calculate the distance over 20000 such properties to evaluate the performance of our distance metric

```
actual_area              6.600000e+02
poolcnt                  1.000000e+00
latitude                 3.404580e+07
longitude               -1.182610e+08
unitcnt                  1.000000e+00
lotsizesquarefeet        7.481800e+04
bedroomcnt               0.000000e+00
calculatedbathnbr        1.000000e+00
hashottuborspa           0.000000e+00
fireplacecnt             0.000000e+00
taxvaluedollarcnt        4.420000e+05
buildingqualitytypeid    7.000000e+00
garagecarcnt             2.000000e+00
age                      1.100000e+01
taxamount                5.407730e+03
airconditioningtypeid    1.000000e+00
regionidcounty           3.101000e+03
Name: 3136, dtype: float64
```

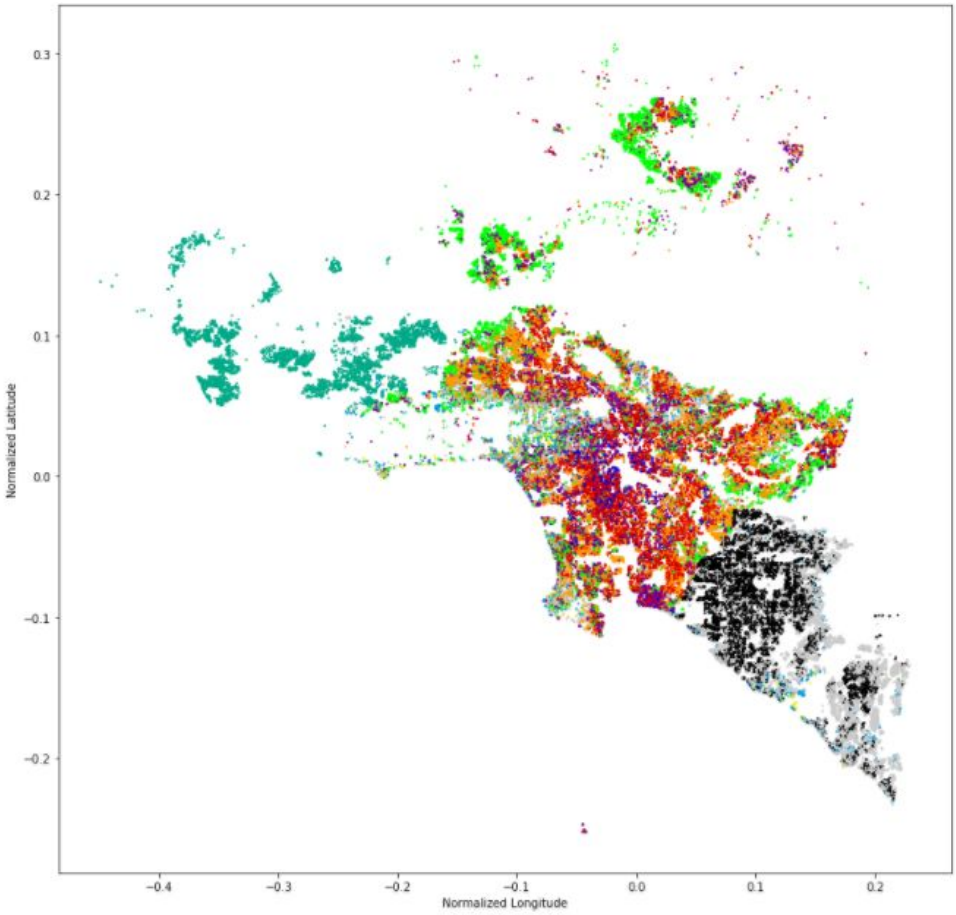
```
actual_area              6.600000e+02
poolcnt                  1.000000e+00
latitude                 3.404580e+07
longitude               -1.182610e+08
unitcnt                  1.000000e+00
lotsizesquarefeet        7.481800e+04
bedroomcnt               0.000000e+00
calculatedbathnbr        1.000000e+00
hashottuborspa           0.000000e+00
fireplacecnt             0.000000e+00
taxvaluedollarcnt        4.420000e+05
buildingqualitytypeid    7.000000e+00
garagecarcnt             2.000000e+00
age                      1.100000e+01
taxamount                5.407740e+03
airconditioningtypeid    1.000000e+00
regionidcounty           3.101000e+03
Name: 13926, dtype: float64
```

In the above image, we try to print the values of the two houses which are different but have the smallest value of distance in the whole dataset. As it can be observed that most of the values in both the properties are the same, the only difference is that of a small amount in the tax values of the properties.
This is a good indication that the distance metric is performing well.

Now we find out the two properties that are the most different from each other. We can observe that there are stark differences in most of the features of both the properties. For example, the first property is almost thrice as large as the second property (based on actual area), both the properties are physically far away from each other and also belong to different counties.

This leads us to believe that our distance function is doing a good job of finding out the similarity between two properties. This function can then be used when we implement clustering

3. Implementing Clustering



For clustering we have used kMeans clustering algorithm. In this algorithm, we take a random set of k points on the plot and assume that they are the centers of the clusters.
In this first pass the distance between the temporary cluster centers and all the points is taken.
Then, the cluster centers of the newly formed clusters are moved.
After which, in the next pass, we find the distance between all the points in and the cluster centers and move the cluster centers accordingly.
We keep on repeating the above steps till the time the cluster centers stop moving.

We prepared a dot-plot map for latitude vs longitude.This plot represents the actual physical distance between the parcels. We have normalized the values of Latitude and Longitude and have divided the points into 15 clusters. As expected the plot also clearly shows the shoreline towards the left side of the plot.
There are some cluster distributions which are a little skewed, that is because there are some clusters which has some empty spaces in between.

4. Merging External Dataset

Datasets that can be merged:

Schools available:

This data can be useful for those people who have children who are school going age. If the house is near a school, then the residents have peace of mind knowing that the children are close by and can be accessed as and when needed. If the children are old enough to walk to school by themselves, then that saves a lot a transportation cost.
Closer proximity to schools increases the desirability of the school and hence the resale value. Since, generally, there is more police presence near schools, this leads to better safety around them.
Presence of a school nearby means that there would be a playground as well nearby. This factor attracts those people who might not have children but are interested in staying fit.
The above-mentioned factors increase the desirability of the property and hence the price also increases. However, there are some disadvantages of living near schools as well. For example, increased traffic during school start and end times, noise from school events, parents parking their cars in the street causing traffic snarls which can be very annoying.

Final Verdict: Living in an area near a school is like a double-edged sword. There are some factors which can jack up the price significantly but on the other side, there are some factors which could negatively affect the price. Final verdict varies from house to house, but we feel that the general trend is that having a school near your property is more advantageous from the standpoint of desirability.

Recreational Facilities available:

Availability of recreational facilities nearby is desired by people of every age. It can be helpful in maintaining both mental and physical fitness. It has been found that living in a non-green area can cause issues like loneliness and depression. In such areas, rates of aggression and violence has also been found higher.
On the flip side, living near a recreational facility can have some problems like traffic snarls because of people trying to reach the facility especially on holidays and weekends, green areas attract wild animals like snakes which can be dangerous sometimes.

Final Verdict: We think that availability of a recreational area near a property helps increase the property value because generally, people find such properties more desirable.

Effects of such data-set on the Zestimate: The presence of these facilities near a property increases its value. In our case, if Zillow had not factoring them in, then this could cause an increase in the value of the logerror. Also, if a new school or a recreational facility opens up after the assessment by Zillow, then that could also lead to an increase in the logerror of that property.

Integration of data:

We plan to integrate this data into our model by measuring the distance between the coordinates of the properties and the recreational facility or school zone. The school zone is calculated by taking its coordinate up to 1 decimal place, this creates a circular zone of around 11 miles around the school, which then can be compared with the property’s co-ordinates.

5. Building models

Implementing Neural Network

Description of MLP

- This model is based on the way a Human brain works.
- It is basically an arrangement of small units called Neurons.
- These neurons are connected with other neurons in the next layer.
- The neurons are activated on the basis of the input to the activation function.

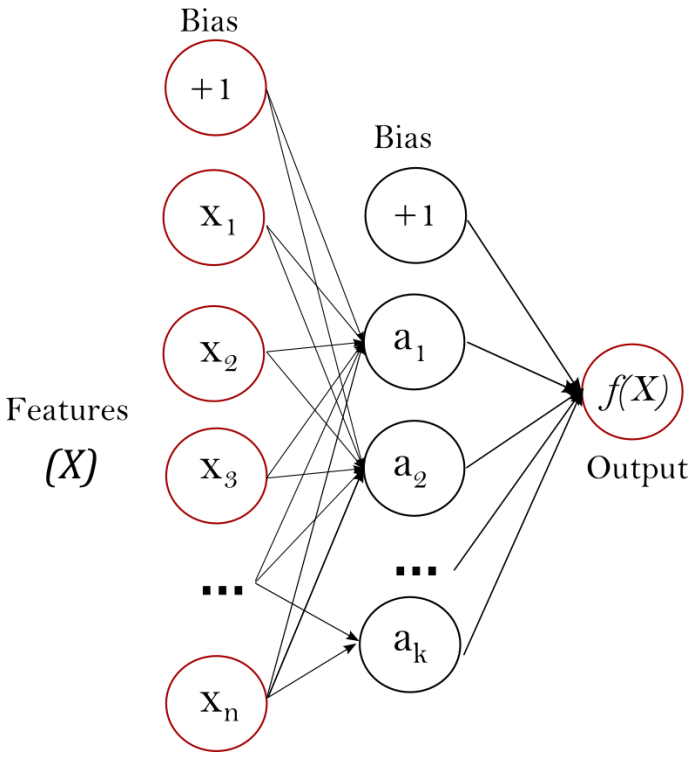


Image source - "http://scikit-learn.org/stable/modules/neural_*networks*_supervised.html"

Each Layer has assigned weights to each incoming connection to the neurons in the layer and also each layer has a bias term. The Image represents a basic neural network.

The leftmost layer, known as the input layer, consists of a set of neurons $\{x_i \mid x_1, x_2, .., x_m\}$ representing the input features. Each neuron in the hidden layer transforms the values from the previous layer with a weighted linear summation $\{ w_1*x_1 + w_2*x_2 + ... + w_m*x_m \}$ followed by a nonlinear activation function like the hyperbolic tan function. The output layer receives the values from the last hidden layer and transforms them into output values.

There were a few surprises when using this model.
One of them was that the model while training had random jumps in the Loss,
which was counter-intuitive, because as the training is done, the loss should decrease. Later, on reading about it. I realised that it because this SKlearn implementation uses Stochastic Gradient Descent by default, which make random jumps in gradients to converge faster and to avoid a local minima.

Hyperparameters to tune the Model

1. Number of hidden layer and the size of each layer - a single hidden layer is enough to model a linear data.
   However, as the data becomes more complex, it can be modelled by using more number of hidden layer.
2. Solver and optimizer - There are multiple solvers used for weight optimization of the MLP, these include
   - ‘lbfgs’ is an optimizer in the family of quasi-Newton methods.
   - ‘sgd’ refers to stochastic gradient descent.
   - ‘adam’ refers to a stochastic gradient-based optimizer proposed by Kingma, Diederik, and Jimmy Ba
3. Learning rate - A high learning rate means that the model is sensitive to outliers and variance. And a small learning rate means that the model is biased towards the data it observed previously. It is important to strike a balance here.
4. Activation Function - {‘identity’, ‘logistic’, ‘tanh’, ‘relu’}
   Activation function for the hidden layer.
   - ‘identity’, no-op activation, useful to implement linear bottleneck, returns f(x) = x
   - ‘logistic’, the logistic sigmoid function, returns f(x) = 1 / (1 + exp(-x)).
   - ‘tanh’, the hyperbolic tan function, returns f(x) = tanh(x).
   - ‘relu’, the rectified linear unit function, returns f(x) = max(0, x)
5. alpha - L2 penalty (regularization term) parameter.

These are the major parameters to tune while training a MLP neural network.

Features Selection

We will be using the following features:
- "transactiondate_year"
- "transactiondate_month"
- "transactiondate_quarter"
- "actual_area",
- "poolcnt",
- "latitude",
- "longitude",
- "unitcnt",
- "lotsizesquarefeet",
- "bedroomcnt",

- "calculatedbathnbr",
- "hashottuborspa",
- "fireplacecnt",
- "taxvaluedollarcnt",
- "buildingqualitytypeid",
- "garagecarcnt",
- "age",
- "taxamount",
- "airconditioningtypeid",
- "regionidcounty"

Architecture

Our model uses 3 Densely connected Layers :
1st with 350 neurons
2nd with 150 neurons
and
3rd with 50 neurons

and the output layer is a dense layer with 1 neuron.

Each layer is followed by a Parametrized ReLU, for activation of neurons.
Which in turn is followed by Dropout layer with a dropout rate of 50%.
Doing so prevents the network from overfitting.

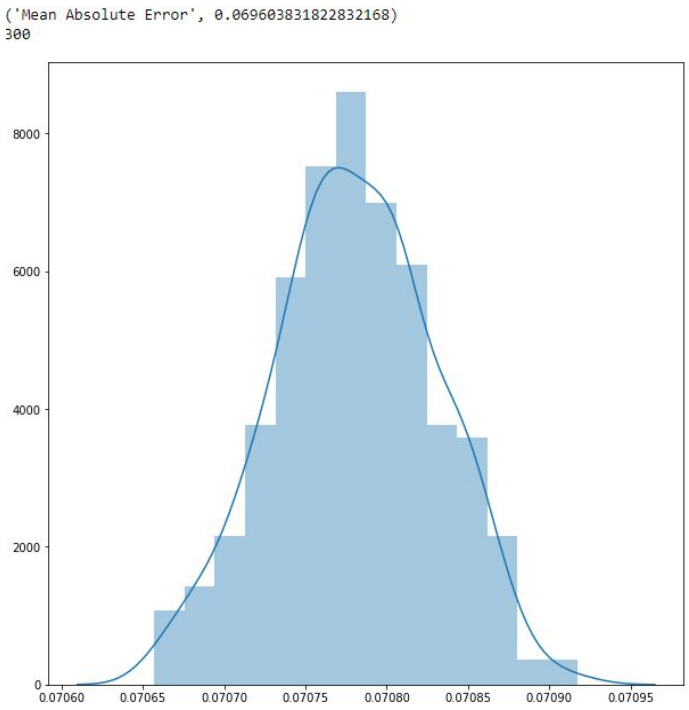
Evaluation of the Model

Here we are evaluating the model by splitting the data into training and validation sets. We then fit the models on the training sets and then predict the values of the validation set. These prediction are checked for finding the MAE.

NN on Validation Data:
MAE - 0.0690

NN on Testing Data:
MAE - 0.0648571

6. P-value



The P-value is 0.0033 for 300 substitutions

7. Submit the result on Kaggle

The above was our best submission. This submission used the above described neural network in the prediction. The result was relatively better than that of Linear Regression.

The above was our second best submission. This submission used Linear Regression that we developed in the first assignment.