Final Report: Homework 3

Dibyajyoti Pati 111482836
Mohaddeseh Bastan 111743223

In this homework we are supposed to do some datascience processing on the Kaggle Zillow data and predict the logerror of zestimate and real value of the properties. This task contains 7 different parts and this report contains what we have done for each part and the result of our work. All the codes and the comments in the notes including important notes are in the notebook HW3.ipynb. The other files in the git account are the output of preprocessing, making model, merging with other dataset.

Question1:Build a scoring function to rank houses by "desirability", presumably a notion related to cost or value. Identify what the ten most desirable and least desirable houses in the Kaggle data set are, and write a one page description of which variables your function used and how well you think it worked

In this question we are supposed to do ranking for the properties based on their popularity. We explain what we did for this part in brief here and in the last part of including external dataset we explain it more.

For ranking the properties we assumed that clustering would help us. It means that if we can cluster the dataset in 2 or 3 clusters and find the center of each cluster, the items which are closer to the center of clusters should be more popular. For example, if we just cluster by latitude and longitude, we see that there are many properties in specific areas and some areas are so sparse. So the properties which are in the sparse areas should be less popular, so their ranking would be lower that the properties which are in more populated areas. Therefore, we decided to cluster the dataset properly and then calculate the distance of each property with the center of the cluster it belongs to and rank the properties based on their distance from the center. The closer a property is the better its ranking is.

For clustering first we just took into account the latitude and longitude of each property. But this is not enough. Because other factors also have effect on being popular that just the position of the property. So we took into account some other features which were not available in the dataset. We used some external dataset to use the school locations and crime coordination for clustering. Finally by having some information about the position

of the property, the age of the property, how big it is, its closeness to the schools and how far it is from the crime area we clustered the dataset. How we preprocessed the external dataset is explained in using external dataset part, but after adding these features we saw that the clustering is more meaningful. We plot the result of the clustering in the notebook. And the result of the ranking is in *rankingResult.csv* file. More explanation of this part is included in using external dataset part.

Question 2 Define a house "pairwise distance function", which measures the similarity of two properties. Like a distance metric, similar pairs of very similar properties should be distance near zero, with distance increasing as the properties grow more dissimilar. Experiment with your distance function, and write a one page discussion evaluating how well you think it worked. Your function should include geographic as well as property-specific variables.

For question number 2 we are supposed to define a distance function which works properly for this dataset. For calculating the distance between properties, we have different approaches. As there are some categorical and some numerical features, we need to process the categorical features to be able to use in calculating the distance function. What we do in this problem is to change categorical features to binary.

A simple way is to add each type of each categorical feature as a new feature and if a property value for that categorical feature is of that type, the corresponding binary column would be 1 otherwise that would be zero. But this makes the dimensional of the dataset so big at the same time it may be so sparse. So the better work is to decide whether it is necessary to add that categorical type as a new feature or not. To do so, first for each categorical feature we count how many of each type is in the dataset. Then by looking at the numbers we decide whether to add that type as a new binary feature or not. For the 'airconditioningtypeid' feature there are 744291 properties which have value of 1, 58487 have value of 13, 1905 have value of 1905, 8795 have value of 5, 1818 have value of 11, 59 have value of 12. 7 have value of 3 and 2169855 have value of NaN. So we decided to make 4 new binary features for these categorical types. First is for airconditioningtypeid' equal to 1, second for airconditioningtypeid' equal to 13, airconditioningtypeid' equal to whatever else except NaN, airconditioningtypeid' equal to NaN. Instead of 'heatingorsystemtypeid' we create 6 binary features, one for row which has value of 2.0 for 'heatingorsystemtypeid' one for 7.0, one for 6.0, one for 24.0, one for other non-nan values and one for nan values. Instead of 'propertylandusetypeid' we create 8 different new feature, one for row has value of 269.0, one for value of 261.0, one for value of 266.0, one for value of 246.0, one for value of 247.0, one for value of 248.0, one

for value of 263.0 and one for other values. because nan value of other features are too many, we disregard other features. We did the same for other categorical features. If one feature is so sparse we disregard it. After that we defined a pairwise distance function as follow:

1. For numerical feature calculate Euclidean distance
2. For binary features calculate inner multiplication. For example if two features have 10 binary feature the same, their distance is 1/10

The result is using both categorical and numerical features and also everything is simple (KISS!). By normalizing the numerical distance and adding the binary distance to it we defined the whole distance function. Which is not only simple but also considers all features and is very precise.

Question 3 Using your distance function and an appropriate clustering algorithm, cluster the houses using your distance function into 10 to 100 classes, as you see best. Present a dot-plot/map (with tiny dots colored to reflect the clustering) illustrating the clusters your method produced. Write a one page discussion/analysis of what your clusters seem to be capturing, and how well they work

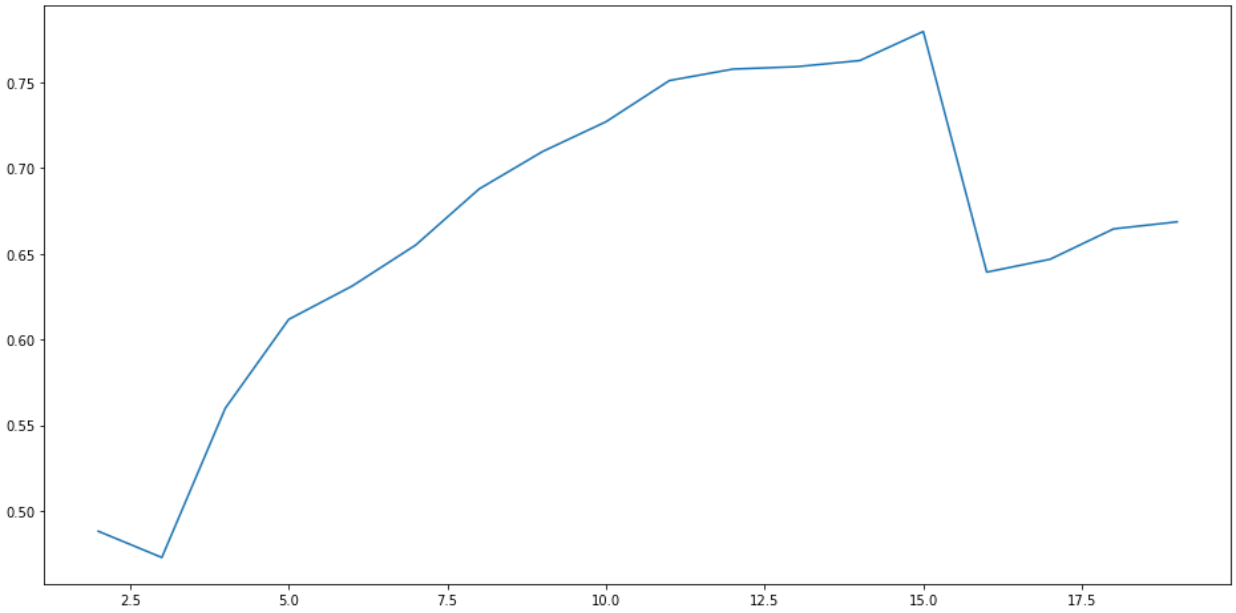
By defining the distance function now we can cluster our dataset. What we did is that we used fcluster to use our predefined distance function and cluster the dataset hierarchically. This function helps us to use our distance function (other clustering function does not let us define our distance function) and also decide what is the best number for number of clusters. By using this unfortunately the process took a long time and our computer ran out of memory. The main reason is that the number of properties is big and the hierarchical clustering is not suitable for big dataset when we have limited memory. So we decided to use another distance function and clustering method for solving this question.

We explored k-means clustering on the following features, 1.Derived:minimum distance from school,crime and maximum crime at the location 2.Few of the property type formed to binary variable where the property was Single Family Residential,Condominium,Inferred Single Family Residential,Duplex (2 Units, Any Combination),Planned Unit Development: the purpose of this being they have maximum number of non-zero values, i.e. most of the properties belong to any of these categories 2.Some of the numerical features finished square feet, room count,tax value assessed and the tax amount collected for the property.

For finding the best number for clusters we used the Silhoutte score. Silhouette refers to a method of interpretation and validation of consistency within clusters of data. The

silhouette value is a measure of how similar an object is to its own cluster (cohesion) compared to other clusters (separation). The silhouette ranges from −1 to +1, where a high value indicates that the object is well matched to its own cluster and poorly matched to neighboring clusters. If most objects have a high value, then the clustering configuration is appropriate. If many points have a low or negative value, then the clustering configuration may have too many or too few clusters.

Silhoutte Score Analysis: Silhoutte score analysis the clusters on the basis of how similar the picked clusters are wrt to other clusters i.e. A silhoutte score of -1 indicates that the picked sample belongs to another cluster and a value of plus 1 indicates that the sample s are not stongly related to cluster, whereas a value close to 0 indicates that the samples in the cluster are strongly related to each other and have a very low likehood of shifting to another cluster on starting at another init. This is the image of using different number of clusters Vs Silhouette score :



Number of clusters were analysed till n=20, the two driving factors for this analysis were: 1.understand the silhoutte score 2.maximize the distribution ant prevent skewing of clusters.

Silhoutte score has a minimum value at cluster=3

There is a distinct separation of three counties on the basis of the cluster even though the clustering involved no parameters related to position directly. The result of clustering based on features is in the notebook.

Question 4 Identify at least one external data set which you can integrate into your price prediction analysis to make it better. Perhaps it can be financial, such as the historical effects of interest rates, consumer confidence, etc. on housing prices. Perhaps it can be geographic, like the crime rate, educational scores, income levels, etc. Write a one page discussion/analysis on whether this data helps with the prediction tasks

For this question we decided to use 2 different external datasets and see the result of merging them with our dataset. First we explaing about these datasets and then we explaing about the result of adding these datasets to the model:

Integrating school location data:

In this section, we tried to integrate data pertaining to the locations of schools with respect to the property. There was no segregation of property on the property type, rather the distance was determined for all parcel ids. The candidates for data integration were the distance from top Private schools and top public schools. The ranking(top) of a school was determined on the basis of a number of enrolled students as the average SAT score for the selected school wasn't available for ranking purpose. The data used for this purpose was assembled from California dept of education and Eric Hurst's gist (git link) which in turn was collected and composed from US Census data. In this attempt, we have integrated the database of private schools only(with the assumption that private schools are more sorted after than a public school and thereby its location would have more impact on the desirability factor). The data from the private school database was sorted and filtered with county code filter and the top 15 schools from each county were selected. The selected data was then mapped to the zip code to Lat/Lon mapping file by performing an inner join on the zip code(which is available for both). This code is submitted as part 'SchoolLocations' notebook uploaded in git. This resultant of this operation was a database with schools in each county and their respective coordinates(the assumption being, as these are densely populated regions the average error between the actual coordinate of the school and the average coordinate of the zip code would be relatively small)

The database obtained in the above step was used to determine the distance between the property and each of the fifteen top schools in each county. The haversine method or flat earth method was used for this purpose. The distance computation was done at county level i.e. properties belonging to LA were matched with schools in LA to avoid humongous computations. The resultant of this was a distance parameter for each property representing the nearest private school distance. The resultant dataframe was

saved with parcel id, minimum school distance, and coordinates for later consumption by performing join operation over the parcel id

Integration of Crime data

A dataset from Open justice was used to arrange and integrate the crime statistics. The Crime statistics record came with two caveats, it had recorded crime for all the years from 1985-2016 overall crime codes/NCIC Codes, they had to be filtered by year and a summation was applied to all crimes at a given location for a given year. This dataset was not directly mapped to a zip code or coordinate, but a rather to a region. Which in turn had to be mapped to a zip code and zip code to coordinates. The purpose of this dataset was to collect two types of statistics

1. Whats the crime rate near a given property
2. How far is the nearest crime reported from the property

For simplification If a region had multiple zip codes, then average crime of all zip codes for that region was assigned to that region instead of maintaining multiple zip codes for a single region. Following a similar approach of joining and mapping, we finally reached a database with each property's parcel id, the recorded number of crimes in the nearest crim location. This data provide us with a very sophisticated method for analyzing a property on the basis of its location which in fact is considered to be the most valuable attribute of a property.

Exploring the result of adding these datasets:

In this question, we try to evaluate the performance improvement upon addition of the new external variables, The steps followed are: 1.Read training and testing notebook 2.read clustering csv, this comprised of cluster label and the three extra parameters fetched from external sources namely, minimum school distance,minimum distances from crime location and the maximum occurrence of the crime in the nearest location 3.The variable pool count was systematically treated for removal of NA: In this the pool count was imputed by this method-> if its Bungalow,Condominium,Rural Residence,Townhouse,Cluster Homeset the pool count was set to 1 else 0 4. After reading the clustering csv and the properties csv (only the required columns were read from each one of them), a join was performed on 'parcelid' to obtain the record of properties along with their external variables. 5.Following this another join was performed with the resultant dataframe obtained in the above step with the training set, resulting in a record of 92715 rows 6.Mean Imputation, data type conversion and scaling of data was performed before modeling to achieve faster convergence Through this exercise we tried to evaluate if there are in gaps in the zillows property evaluation by imputing external data into our system and evaluating if there are any improvement in

prediction of log error. It is evident from the mean absolute error on the validation set and zillow score that the external data we choose to impute into our system had no or very minimal impact to our overall score.i.e. Our Hypothesis that zillow would have ignored facts like the distance from nearest school or nearest distance at which a crime has occurred is proven incorrect: Upon analysis, there were multiple reasons for the failure of this hypothesis 1.Zillow would have already accounted for this 2. there is no real impact of these parameters on the cost of a property(which we are very skeptical about) 3.Gross error in computation of these parameters: There were certain assumptions made while computing these parameters, Ex. We considered distance from private school as more determining than distance from public schools, there were top 15 schools picked from each county, whereas LA has more population,so more number of schools should have been a better idea, the criteria for selecting best schools was based on the total number of enrollments ,which might be the best approximation in the absence of other params like average SAT score. Modeling: In this exercise we tried to model on a subset of the data set picked by k-fold cross validation. A total number of 200 folds were made and run in iteration, saving the best model so far to a variable . The swapping was performed on the basis of the mean absolute error on the validation set. This process was repeated for records in 2016 ad 2017 to populate the rows for the submission file, the reason being that the tax values in the property set have seen a updation across all counties.

Question 5 Finally, build the best prediction model you can to solve the Zillow task. Use any data, ideas, and approach that you like. Predict the logerror for instances at file "sample_submission.csv". Report the score/rank you get. You are allowed to merge your prediction teams to the extent that Zillow allows it. Write a 2-3 page report about how it works, an evaluation, and any interesting experiences along the way

We used different method for this task. The first one which was linear regression we used in previous homework. It is simple and it works. But the result is not very good. So for this homework we decided to make use of our new features and use a random forest to see whether the model gets better or not. Also the dataset is different from previous task. The main difference between 2017 and 2016 datasets is that the first one contains the information about the properties which has been sold in 2017. Therefore the log error and transaction date is very different. But some other small features are also different. After analysing the datasets we saw that some binary features has been changed too. So for making categorical features as binary, the classes are different for two datasets and we handle it by running different preprocessing methods for both datasets. We used the

random forest for the new dataset. The value of R^2 (the score of our model) is completely different from the value for 2016 dataset. The 2016 dataset has range of 0.009 and this dataset has range around 0.14 and first we thought our model is better but after uploading at kaggle we saw that our score in test sets decreased. So we decided to use other methods for solving this problem and predicting the log error.

The other attempt was to use our new features instead of the usual one. We changed the categorical features into binary and used them for calculating distance function. We decided to used these new features and random forest but the result didn't get better.

The other interesting thing we found is that when tuning the parameters of random forest model, we found that increasing the depth of the tree is not improving the model, however it increases the complexity of the program. It may be because of the types of the data we have that the depth has effect to some degree.

Also we used another method for solving this part, we used a neural network. We created a simple MLP and tried to tune the hyperparameters (number of layers, number of node in each layer, activation function, solver method and alpha, beta parameters in adadelta) but none of them make the model better. We think the most reason for this result is that the simple MLP may not work fine for this dataset and may be other methods work better.

Next, we have explored ensemble models in attempt to improve the performance. After trying out multiple methods, we picked out extra tree regressor for this problem. The reason being the best score in validation set.

Extra tree regressor stands for extreme randomized trees and in way similar to random forest but uses optimal cut points instead of boot strapping for introducing the variance.

The performance of this model was further improved by trying out multiple values of tuning parameters like ,number of estimators: which is the number of trees to used to make the prediction(the output being average of these values) and the maximum depth of the tree.

We tried our hands on k-fold cross validation method, and it certainly improved the model performance. K-fold cross validation divides the train and test set to k- different subsets and uses k-1 sets in each iteration for training and just 1 set for validation purpose.

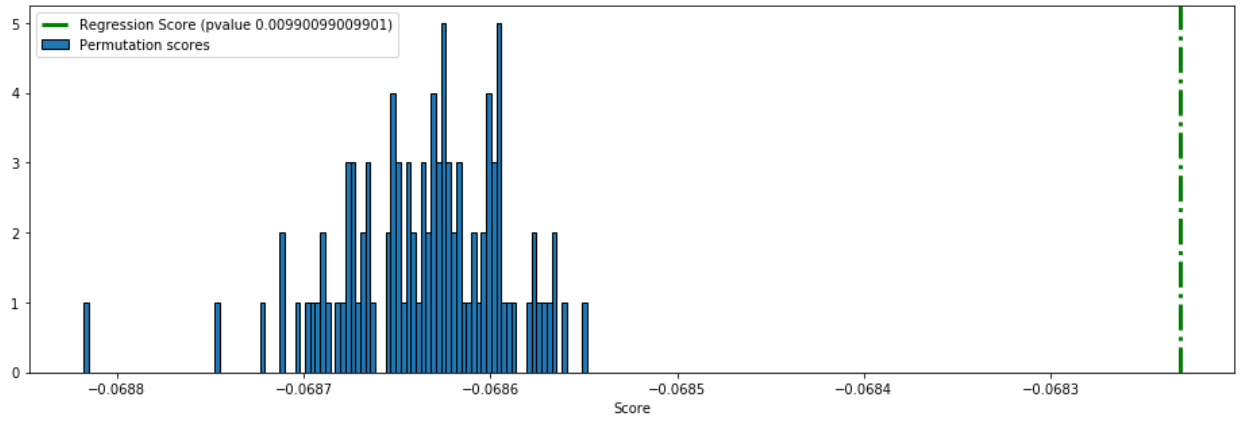
This randomizes the train and test set in total of k- iterations.

We observed that increasing the number of folds improved the mae score to an extent, after which the increase in k- has no impact.

The values of k were tried from 20-500, and the optimal value was found to be at 200

Question 6 Do a permutation test to determine a p-value of how good your predictions of logerror are. You can use whatever metric you wish to score your model (like mean absolute error). For a large enough sample of the evaluation data, compare how your model ranks by this metric on the real data compared to 100 (or more) random permutations of the logerror assigned to the real data records. What fraction of permutations produce at least error at least as good at the real data set? If necessary, sample your data so these 100+ runs do not take too much time.

Evaluation of model on p-test: we used sklearn's permutation package to evaluate the effectiveness of the model. This used 100 permutations tests to determine the significance of the model , the model used in this case was the best model derived by extra tree regressor in the steps above it, inorder to break the tasks the model was pickled to file and read later.
the outcome of this result was plotted as a histogram
the magic number 0.00990 for the significance value comes from the fact that the input model was better on actual target than all the permutation inputs than provided to it



As seen the model prediction is seen to have outperformed all other random permutation tests performed, its affinity to the target variable is evident from the fact that the average mean absolute score reduces when applied on random shuffled target

Question 7 Submit your results on the real test data to Kaggle before deadline. Write the result into a csv file and submit it to the website. Actually, submit two for your two best models, to the extent that Zillow allows it

With new model (bootstrap) we could make our model better and the final score was 0.0648558 and the final ranking was 2377.