Report: Boston Housing Market Challenge

Team name: The Bandicoots
Team members: Jemima Graham & Travis Gordon

February 21, 2021

1 Introduction

The Boston housing data-set [1] is a well known data-set which has been extensively used to test new machine learning methods. The aim of this project was to investigate the impact of introducing new geo-spatial features into the pre-existing Boston housing market data-set. Specifically, information about the surrounding facilities such as schools, universities, hospitals, train stations, and community health centers has been investigated as this information is often taken into account on modern-day real-estate sites. The data on facilities originates from 2018 [2], 2018 [3], 2018 [4], 2015 [5], and 2019 [6] respectively. This means that a key assumption is that the infrastructure has not changed drastically over the last 41 years; this is reasonable to some extent as universities and hospitals are often longstanding facilities, however, we recommend investigation into the validity of this assumption for future work.

2 Methodology

2.1 The original data-set

The original data-set provided by the ICDSS team was a clean and complete data-set. This data-set contained 506 observations with features such as: town name, longitude, latitude, crime per capita, percentage of lower status population etc. This data, along with the data for hospitals and schools, was stored in a csv file, so this needed to be converted to a pandas dataframe before it could be converted to a geopandas one. The crs also had to be set manually. This did not need to be completed for the university, train station, or community health centre data because these were stored as shapefiles. However, the latitude and longitude of the shapefiles were switched, so these needed to be manually switched back.

2.2 Data cleaning

The following steps were taken to clean the data:

- NaN values: check that no columns contained NaN values. Missing data would have needed to be accounted for - either by removing the row/column or introducing a technique such as label encoding.
- Cell dtypes: check that all values in a column contain the same data-type.
- Longitude and Latitude bounds: check that all of the coordinates are located inside the Boston area.
- Town id: check that each unique town name corresponds to a unique town id.

The original data-set contained two response variables median values of owner-occupied housing (MEDV) and corrected median values of owner-occupied housing (CMEDV). CMEDV was selected as the response variable and the MEDV column was dropped from the data-set to avoid data leakage. The original data-set also included columns for both town name and town id. Therefore, the town name column was also dropped.

The following geo-spatial data was included:

- Hospitals: All hospitals in Massachusetts ([4]).
- Schools: All pre-kindergarten to high-school schools in Massachusetts ([2]).
- Universities: All universities in Massachusetts ([3]).
- Train stations: All train stations in the Boston area ([5]).
- Community health centers: All community health centers in Massachusetts ([6]).

Due to the time constraints of the project, all columns except location were dropped from the external geo-spatial data. It should be noted that the external geo-spatial data added is the most up-to date data available while the Boston housing challenge contains data from 1978. We have assumed that the geo-spatial data is applicable to the year 1978. This is an obvious oversimplification. Given more time on the project, the geo-spatial data added would be trimmed to only contain facilities that were present in the year 1978. The initial justification for using the data is that the types of facilities included in this analysis last a long time and the majority of the facilities were likely present in 1978.

All the longitude and latitude values for both the external data-sets and the original Boston housing data-set needed to be projected from epsg $= 4326$ to epsg $= 900913$ in order to calculate the distance in meters between the different locations. This is because epsg $= 4326$ is measured in degrees while epsg $= 900913$ is measured in meters.

2.3 Feature engineering

Figure 1 illustrates a section of Massachusetts with icons indicating the location of the external facilities used in this study. As can be seen in Figure 1, a considerable amount of data-points have been added. This facilitated the next stage of the project which involved generating new features.

Two types of features were generated: the distance to the nearest facility, and the number of facilities in a $10\,\mathrm{km}$ radius. The list of features generated are as follows:

- Distance to nearest school.
- Distance to nearest university.
- Distance to nearest hospital.
- Distance to nearest train station.
- Distance to nearest community health center.
- The number of schools in a $10\,\mathrm{km}$ radius.
- The number of universities in a $10\,\mathrm{km}$ radius.
- The number of hospitals in a $10\,\mathrm{km}$ radius.
- The number of train stations in a $10\,\mathrm{km}$ radius.

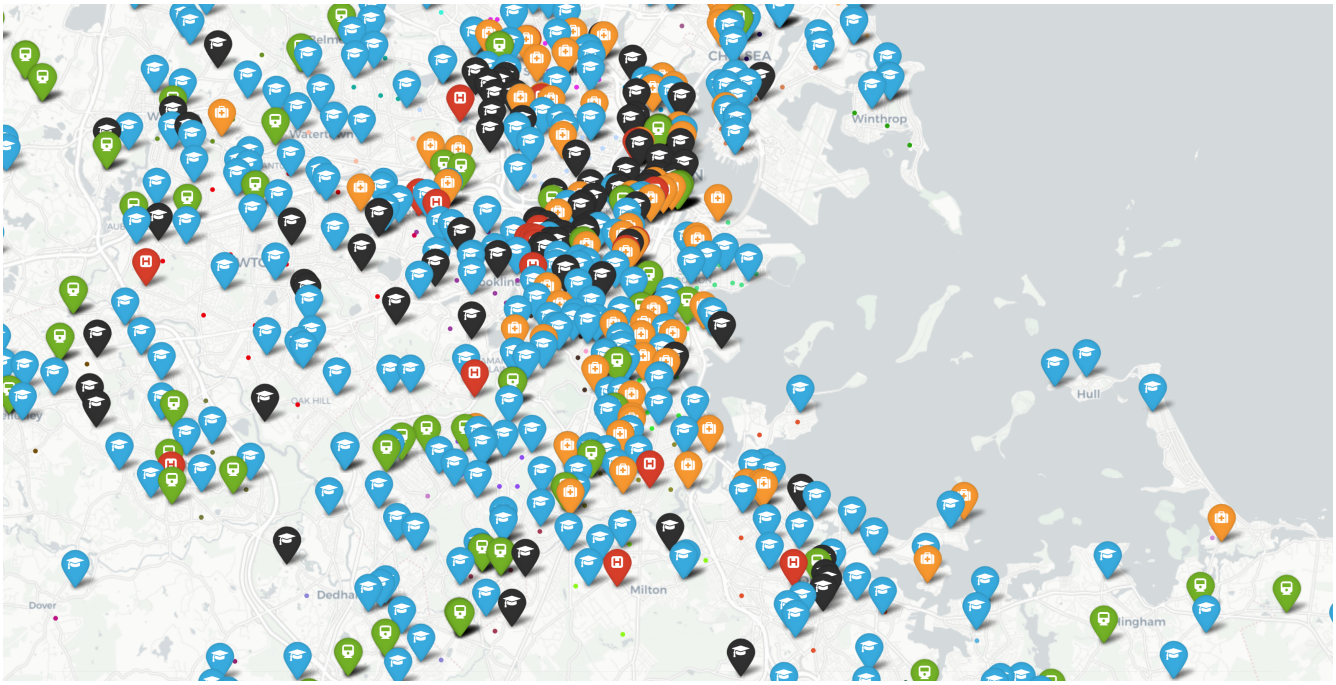

Figure 1: Screenshot of part of map of Massachusetts where circles represent houses coloured according to town and icons represent utilities which symbols that represent their function.

- The number of community health centers in a 10 km radius.

The range of data generated for the distance to nearest facility, and number of facilities within 10 km was investigated by using box-plots. This was to give a simple indication of the spread of the data. Additionally, outliers could easily be identified.

The mutual information score of each of these features (along with features of the original data-set) were compared. This was done to investigate which features most reduce the uncertainty of the target variable. With this knowledge, a smaller subset of the features could be selected for the final model. This allows redundant features to be dropped and improves the explainability of the final model.

A correlation matrix of the features was also generated to help gain an understanding of the diversity of the features. This concept was taken further by calculating the PCA components of each feature. Due to time constraints, the results of the PCA and correlation analysis were not incorporated into the feature selection process. However, with more time, features would only be considered for the final model if they scored an acceptable mutual information score and diverse enough from other high scoring features.

2.4 Model training

Before data is fed into any models, it is normalized using sklearn's StandardScaler function in order to ensure that the models perform well; it was also normalized before using principal component analysis (PCA) which is mentioned in Appendix B but this cannot be discussed here due to time constraints.

Three models (a linear regressor, an elastic net regressor, and a decision tree regressor) were trained on the combination of the Boston Housing dataset and the newly created features, while an additional decision tree regressor was trained using the top ten most insightful features according to mutual information score (as shown in Fig.2 where higher mutual information scores are better).

Five-fold cross validation used to evaluate the efficacy of each model, with a 80:20 training set:test set split. A hyper-parameter search was carried out for all elastic net and decision tree models in order to find the optimum penalty term strength and complexity parameter respectively.

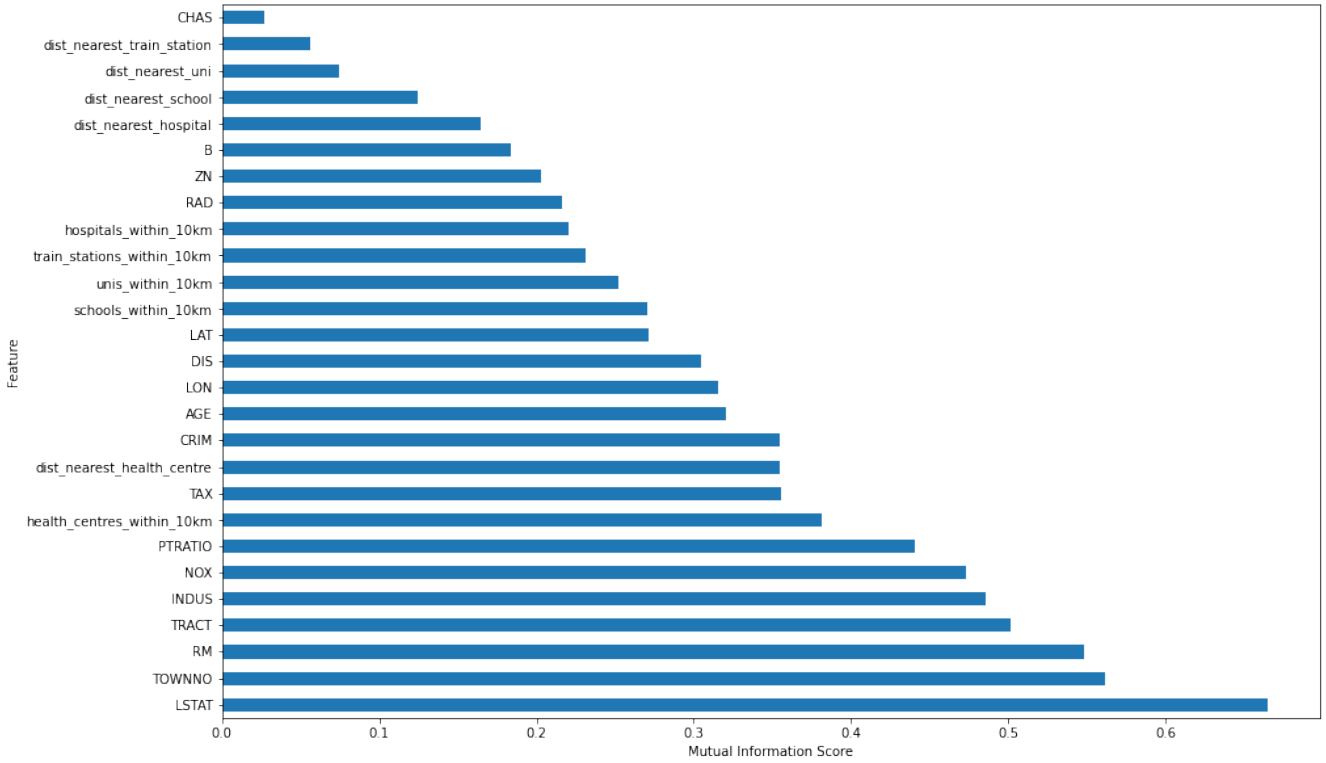

Figure 2: Mutual information score between each feature and the target variable (corrected house price in $1000s).

3 Results & Discussion

The root mean squared error (RMSE) and mean absolute percentage error (MAPE) are given in Tab.3 for each model discussed in Section 2.4. It can be seen that the elastic net performs slightly worse than the benchmark (the linear regression model with no regularization) despite the use of a hyper-parameter search to find the optimum penalty term strength. The decision tree regression without feature selection clearly outperforms these models suggesting that it is the most suitable model considered so far. Feature selection based on mutual information score improves on this further as shown by the reduction in RMSE and MAPE shown in the last row of Tab.3.

Table 1: The final results of the train and test sets. The quantities displayed are the root-mean-squared error (RMSE) and mean absolute percentage error (MAPE).

	trian_RMSE$(\times 10^3\$)$	**test_RMSE**$(\times 10^3\$)$	**train_MAPE**(%)	**test_MAPE**(%)
Linear	4.258474	4.692458	16.266513	18.297933
Elastic Net	5.428836	5.466787	18.330701	18.784390
Decision Tree	2.475156	4.852431	10.355333	17.039425
Feature Selection Decision Tree	1.750009	4.035640	7.429926	14.764868

However, it could be argued that the regularization in the elastic net model clearly serves its purpose to prevent overfitting of the training data as the MAPE for the training data (train_MAPE) is only marginally smaller than the MAPE for the test data (test_MAPE) while all of the other models show a much larger difference; this implies that the benchmark model and both decision tree models are overfitting the data.

In addition, Fig.3 and Fig.4 are box-plots of the features created to consider the distance between the houses and desirable facilities. Both box-plots are highly skewed, while Fig.3 also indicates that there are a lot of outliers. Further checks must be completed to ensure that the projection from degrees to meters and the distance calculation are completed correctly. If no bugs are found during these steps, development of less skewed features must be considered. One interesting

observation is that the features represented by health centres are still highly skewed and have many outliers even though they have a good mutual information score when comparing the feature to the target variable. Investigation into whether the mutual score improves once there are fewer outliers is recommended.

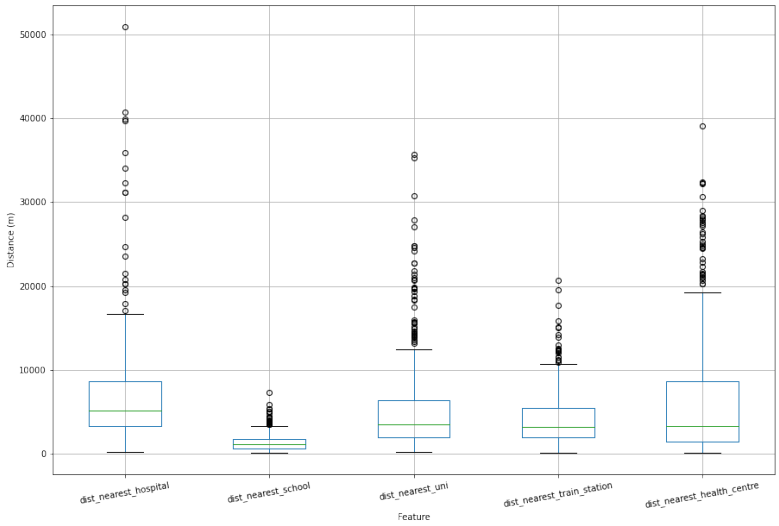

Figure 3: Box-plot of distance between each house and the nearest hospital, school, university, train station, and community health centre.

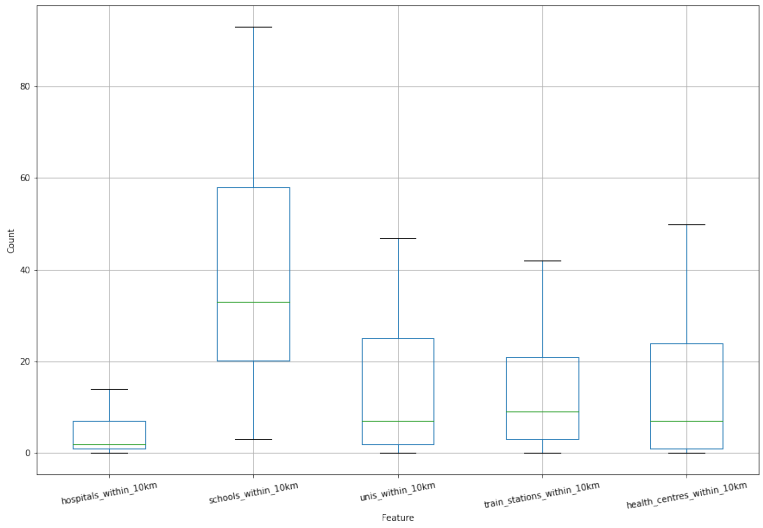

Figure 4: Box-plot of number of hospitals, schools, universities, train stations, and community health centres within 10km of each house.

Altogether, given more time it would be beneficial to only use data corresponding to facilities present in 1978. In addition, more tailored feature generation may be required. For example, house prices may be more highly correlated to the number of schools within a smaller radius than the 10 km chosen in this study.

4 Conclusion

In conclusion, this project highlighted the influence of health centre location on house price as both the feature related to distance to the closest health centre and the feature related to number of health centres in a 10 km radius ranked in the top ten according to mutual information score. Additionally, this project emphasizes the need for feature selection as the model that only includes the top 10 features according to mutual information score far outperforms the other three models discussed here. Contrastingly, despite the low mutual information scores seen for the other distance features created as part of this project, distance to hospitals, schools, universities, and train stations cannot be ruled out as an important factor at this stage due to the large gap in time between the Boston Housing Challenge dataset and the location data for the facilities discussed here. Investigation into the impact of this is recommended for future work. Another recommendation for future work is use of the correlation matrix and principal component analysis (PCA) shown in Appendix A and Appendix B respectively as this contains a lot of valuable information that could not be used due to time constraints. A final recommendation is to improve the accuracy and reduce the amount of overfitting seen in the best performing model (Feature Selection Decision Tree) as this can hinder performance when applied to new data.

A Correlation matrix

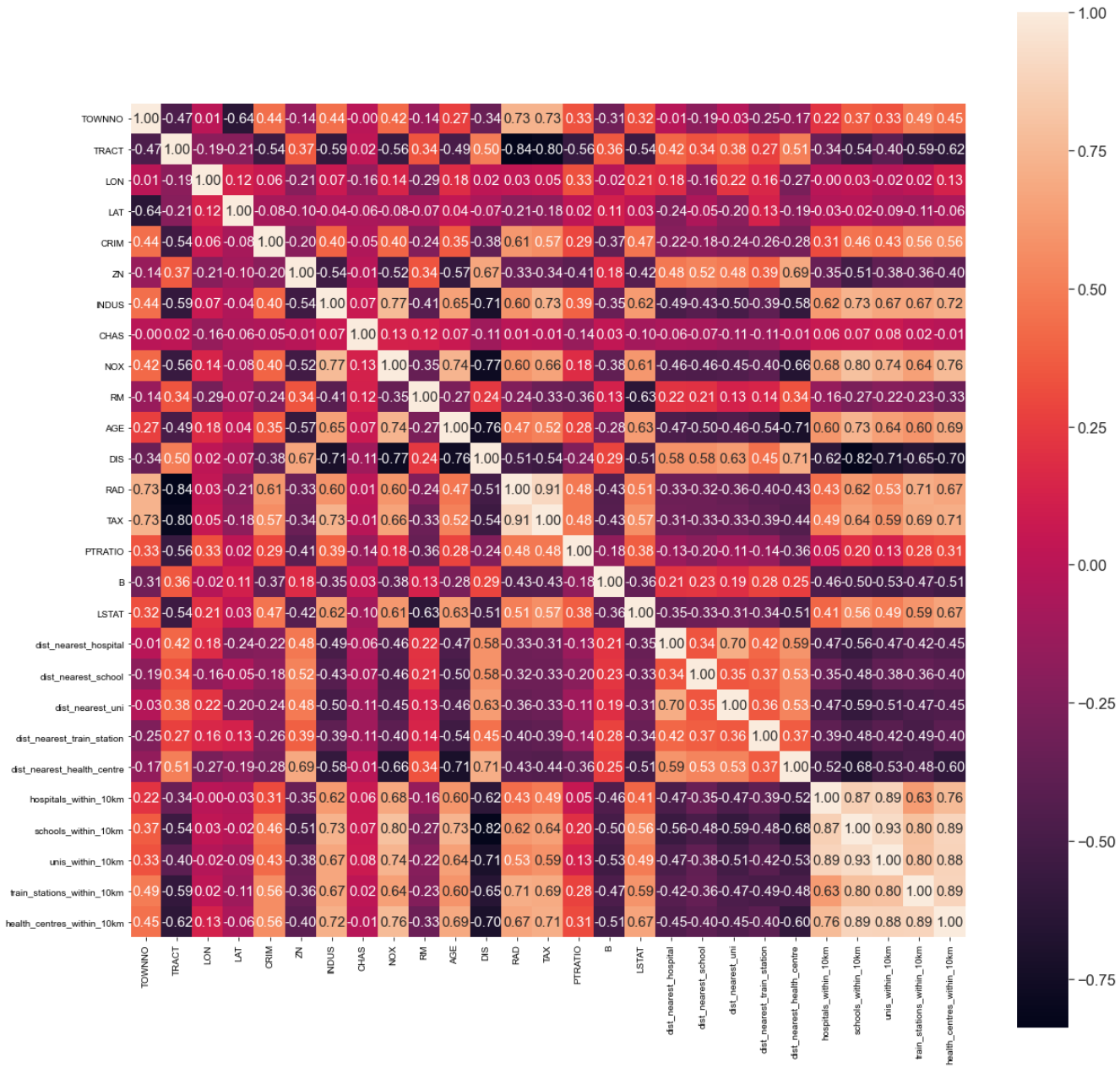

Figure 5: Heatmap of the correlations between each feature

B Principal Component Analysis (PCA)

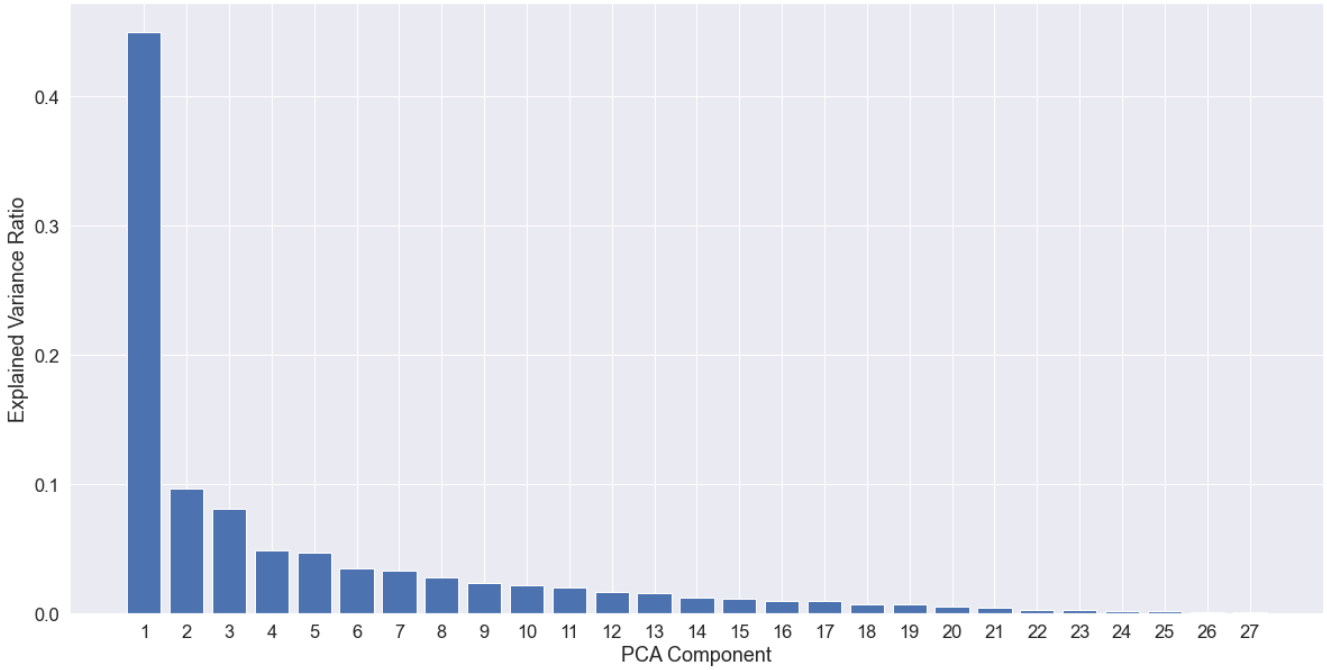

Figure 6: Explained variance ratio of each PCA component. The higher the value, the more important the component is for determining differences between features.

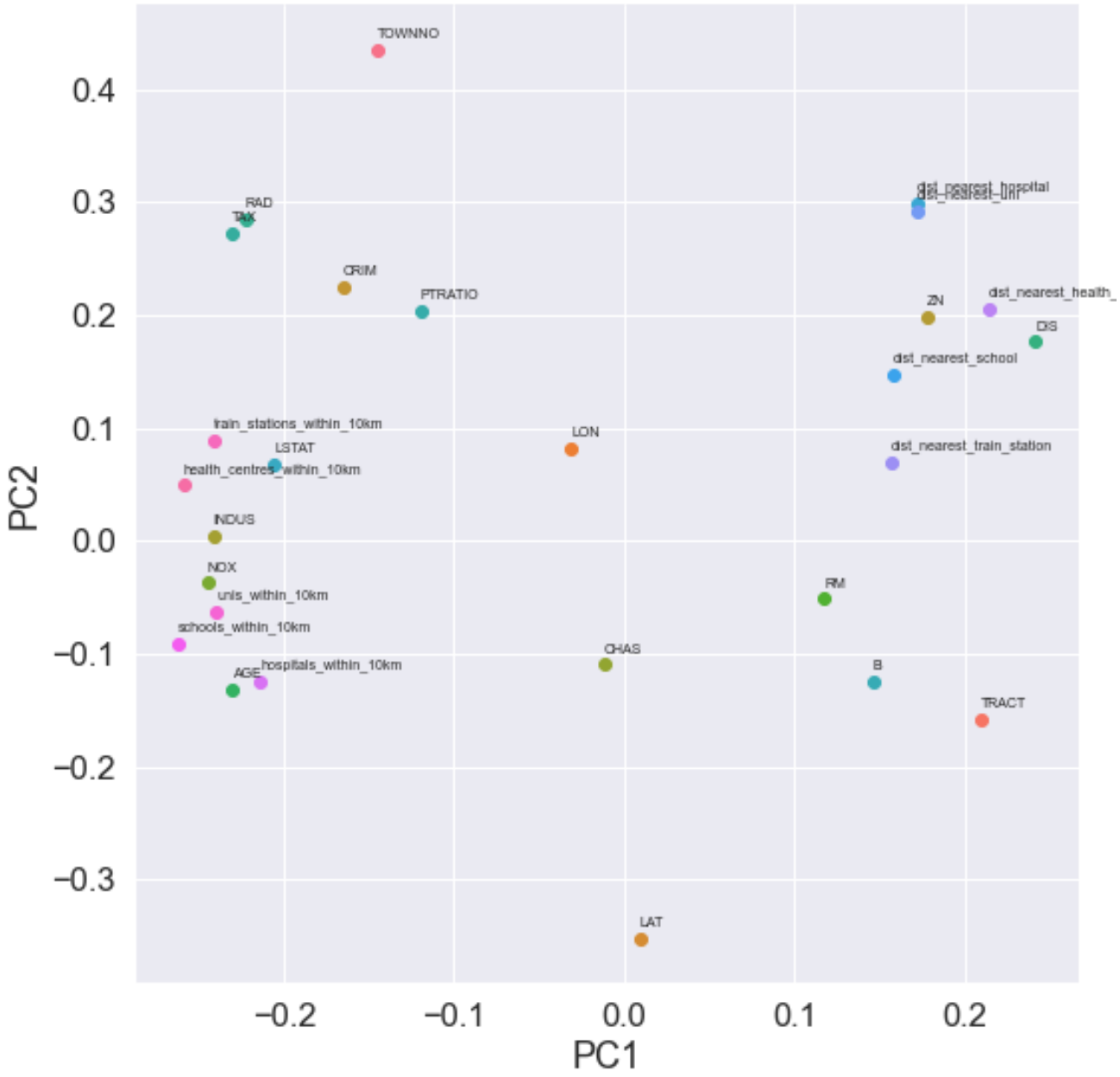

Figure 7: PCA component 2 against PCA component 1 to show which features are the most similar; separation along the x-axis carries more meaning than separation along the y-axis because PCA 1 has a higher explained variance ratio.

References

[1] D. Harrison and D. L. Rubenfield. Hedonic housing prices and the demand for clean air. Volume 5:81–102.

[2] Koordinates. Massachusetts schools (pre-k through high school). https://koordinates.com/layer/98813-massachusetts-schools-pre-k-through-high-school/.

[3] Massachusetts Document Repository. Massgis data: Colleges and universities. https://docs.digital.mass.gov/dataset/massgis-data-colleges-and-universities.

[4] Koordinates. Massachusetts hospitals. https://koordinates.com/layer/97766-massachusetts-hospitals/.

[5] Massachusetts Document Repository. Massgis data: Trains. https://docs.digital.mass.gov/dataset/massgis-data-trains.

[6] Massachusetts Document Repository. Massgis data: Community health centers. https://docs.digital.mass.gov/dataset/massgis-data-community-health-centers.