**Boosting House Price Predictions using Geo-Spatial Network**

**Embedding**

We propose a novel geo-spatial network embedding (GSNE) framework that can accurately capture the geo-spatial neighborhood context in terms of different types of POI and their features, and the relationships among these POIs in a weighted, attributed multipartite graph.

– We adopt and extend the Gaussian embedding methods to realize our GSNE framework, which is highly efficient and can work with heterogeneous types of nodes and features.

– Our comprehensive evaluation on a large real-estate dataset shows that for the house prediction task, combining geo-spatial embedded vectors learned by GSNE with the housing features results in consistently better prediction performance than raw feature only, regardless of the downstream regression model.

**Dataset:**

We conducted our experiments on the house transaction records obtained from a real-estate website2 for Melbourne, which is the second largest city in Australia by population.We extracted a total of the 52,851 house transaction records of years from 2013 to 2015. Our dataset also includes the three types of POIs: regions, schools, and train stations and their corresponding features. Houses are situated in regions which capture the geographical contextual information about houses. Intuitively, information about nearby schools and train stations may influence house prices. Our dataset contains information of the 13,340 regions, 709 schools, and 218 train stations.

**Housing Features**: Our dataset contains information about a wide range of housing features. In

total, we consider 43 housing features for each house for in depth exploration of the effect of GSNE. To the best of our knowledge, none of the prior works considered such a wide range of feature sets in a large dataset like ours for house price prediction task. Although the dataset in Kaggle competition has 86 features, it has only 3000 samples in total and lots of columns are highly sparse rendering only a few of those columns truly useful. Besides, no information regarding neighbourhood amenities is available in that dataset. In our dataset, each house record contains information ranging from basic housing features like number of bedrooms, number of bathrooms, number of parking spaces, location, type of property, etc. to detailed facility features like air-conditioning, balcony, city-view,river-view, swimming, tennis-court, etc.

**Region Features**: Our dataset contains Melbourne region information at SA1 level3. SA1 is the

most granular unit for the release of census data of Australia. The SA1 data typically has a

population of 200 to 800 people with an average of 400 people per region. For each region, our

dataset contains comprehensive information about the number of residents, average age, median personal income, percentage of Australian citizens, educational qualification, median house rent,location as the centroid of the region, etc. Since these aspects can be useful for determining house prices, we consider all of them as the features for regions.

**School Features**: The schools in our dataset are attributed with the type of school (primary

or secondary), school category by gender(single gender or co-ed), ranking, location, number of

students, zone restrictions, number of students enrolled in Victorian Certificate of Education(VCE), percentage of students securing 40% marks, etc

**Train Stations**: The train stations in the dataset contain information about their location and

average time to reach to other stations.

You can get the experiment dataset [here](https://drive.google.com/drive/folders/1vTkD3zGCwvNomUDmsdsbgLINpYHrQMiR?usp=sharing).

**Project Structure:**

Repository

|

|-- GSNE\_Boosting\_House\_Price

|

|-- 1. Dataset

|

|-- Check\_Performance\_Dataset

\*\*\* contains check performance required datasets \*\*\*

|-- Preprocessed\_Dataset

\*\*\* Main raw data \*\*\*

|-- Processed\_Dataset

\*\*\* Processed npz files \*\*\*

|-- 2. Data\_Preprocessing

\*\*\* data-preprocessing codes \*\*\*

|-- 3. Graph\_embedding

|

|-- Code

|-- utils.py

|-- model.py

|-- train.py

|-- job.sh

|-- requirements.txt

|-- 4. Checking\_performance

|-- embedding\_1

\*\*\* contains embedded pickle file \*\*\*

|-- embedding\_2

\*\*\* contains embedded pickle file \*\*\*

|-- check\_performance.py

|--ReadMe.md

To run the the project,you have the followings to do-

1. Check the requirements in the requirements.txt in the Graph\_embedding folder 3 and install them, Then just do *python train.py ‘cora-ml’ ‘glace’*

or simply run the command ***./job.sh*** in linux system.

1. Now for the price prediction from the embedded pickle files in folder 4.Chechking\_performance, run ***python chechk\_performance.py***