DWDM Project Report

Real Estate Search based on Data Mining

Submitted by

T. Subramaniam
14CO245
Amit Shrikrishna Wadatkar
14CO253

VII Sem B.Tech (CSE)

in partial fulfillment for the award of

COMPUTER ENGINEERING

Department of Computer Science & Engineering Technology
National Institute of Technology Karnataka, Surathkal.

Summary of the Project

The problem statement is to predict the price of real estate based on related parameters which are taken as input from the user. The main objective of this project is to develop a price prediction model which can guide the buyer to make a right choice, based on his required house preferences.

The idea is to build the price prediction model using two layers of cascaded classifiers. The first layer contains four different regression models which are individually trained on 70% of the dataset. The remaining test data is used against each of the regression models to produce the predicted price. The second layer contains a Gradient Boosting Regression model whose input is the output of the predicted price obtained from the first layer, appended to few or all of the other features in the original dataset. The output of the second layer is compared against the expected output value of the test data and thus the second layer is trained. A comparative study is done across the possible combinations of feature inputs to the second layer. Principal Component Analysis (PCA) is used to select the most relevant features in the case when only few features are passed on as input to the second layer. The buyer does not have the need to know about the features that are influential in price prediction as the system takes care of it. The buyer can also optionally leave the some features blank and the system handles such missing values as well. Hence, the buyer can get a good estimate of the price in accordance with his preferences regarding the real estate.

Approach used

- ## Data mining technique used

Pre-processing of the dataset:

- The NA values were replaced by taking the mean of the existing values of the same feature across the entire dataset.
- Duplicate training examples were removed and the dataset examples were made unique.
- Noisy data such as data type mismatch and inappropriate values were handled through assertion statements.
- Outlier analysis was done using Z-score statistics and the outliers were removed.
- Feature scaling was done to ensure uniformity. The variance was scaled to one and the data was centered on the mean.

Building the model:

The model consists of two layers of classifiers. The first layer consists of the following four regression models:

- Linear Regression
- Gradient Boosting Regression
- Random Forest Regression
- Decision Tree Regression

The second layer consists of a single Gradient Boosting Regression model which predicts the output based on the output from the first layer and the other features in the dataset.

Training and testing the model:

The training dataset consists of 70% of the entire dataset and the remaining goes to the test dataset. The first layer is trained using the training data and four prices (one from each of the four regression models) are predicted for each training example of the corpus.

A new training dataset is created for the second layer. The new corpus consists of the price predicted as one of the features. Either none, some or all of the features in the original corpus are appended to the new corpus.

- **Stacking**: None of the features in the original corpus are included in the new corpus.
- **Blending**: All of the features in the original corpus are included in the new corpus.
- **PCA**: Influential features in the original corpus are used to derive new features which are included in the new corpus using Principal Component Analysis. These features influence the price of the real estate the most.

The decisional attribute of the training examples in the new corpus is same as the decisional attribute of the corresponding examples in the original corpus

The new corpus is split into training and testing data with a ratio of 80:20 respectively. The Gradient Boosting Regression model is trained with the training data and the accuracy of the entire model is tested against the test data to compute the accuracy.

- **Data set used**

  The corpus consists of 17 attributes to relevant to real estate and price of the real estate as the numerical label. The real estates are based out of America.

  The following features are selected in the corpus:

  - Price – Selling price of the house.
  - Number of bedrooms
  - Number of bathrooms – 0.5 indicates the need for a toilet but not a shower.
  - Number of floors
  - Area of the interior living space (in square feet)
  - Area of land space (in square feet)
  - Requirement of location abutting a waterfront (on a scale of 0 to 1.)
  - View – Rating of how good the view of the property was (on a scale of 0 to 4.)
  - Condition – Rating of the condition of the property (on a scale of 1 to 4.)
  - Grade – Rating based on quality level of construction and design (on a scale of 1 to 13.)
  - Area of the interior housing space above ground level (in square feet)
  - Year built
  - Year of latest renovation
  - Zip code
  - Latitude
  - Longitude
  - Area of the interior living space of the nearest 15 neighbors (in square feet)
  - Area of land space of the nearest 15 neighbors (in square feet)

  The selected features cover a variety of preferences for a variety of buyers. Location of the house is covered along with the preferred density of neighbors. Measures of the dimensions of the house and the number of various rooms are covered. Aesthetics of the house is also accounted for in the corpus.

Results and Discussions

The cascaded prediction model estimates the price of the real estate based on the input features. It produces a single floating point value as the estimated price. The real estate search is based out of King County Area, America.

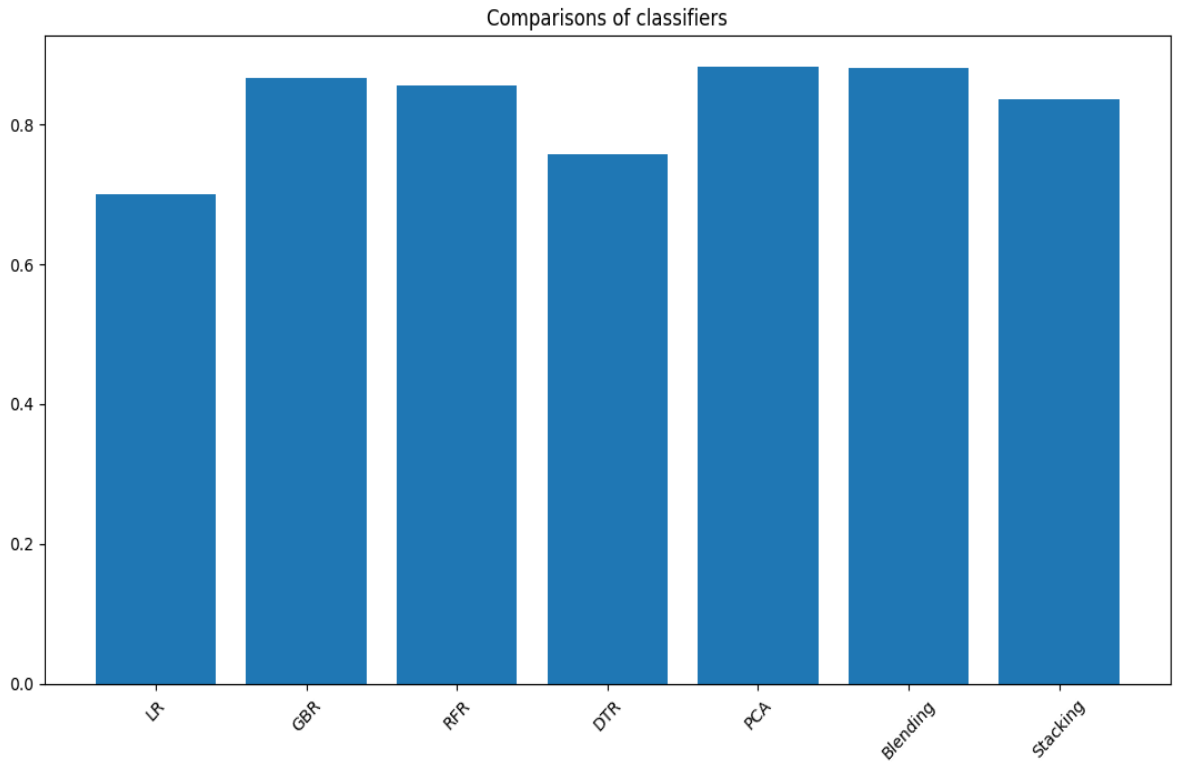
Here is the tabular form of the accuracies produced against individual classifiers present in the first layer (no cascading):-

Regression Model	**Accuracy**
Linear Regression	70.1714%
Gradient Boosting Regression	86.6261%
Random Forest Regression	85.1936%
Decision Tree Regression	74.5019%

Here is the tabular form of accuracies produced against different combinations of features selected in the new corpus as explained above:-

Type of Feature Combination	**Accuracy**
Stacking	86.8925%
Blending	88.3265%
Principal Component Analysis (PCA)	90.0926%

Graphical representation of the comparative study:-



Legend:-

- LR - Linear Regression
- GBR - Gradient Boosting Regression
- RFR - Random Forest Regression
- DTR - Decision Tree Regression

Major findings include the fact that cascaded classifiers perform better when there are a lot of features in the corpus while not all of them influence the result strongly. PCA is a helpful technique to filter only the most influential features, use it to train the cascaded classifier and then to predict the price according to the preferences. The most influential features as computed by PCA are:-

- Number of bedrooms
- Requirement of location abutting a waterfront
- View
- Condition
- Area of the interior housing space above ground level
- Year of latest renovation
- Latitude
- Longitude
- Area of land space of the nearest 15 neighbors

Conclusion

The buyer is given an interface to input his preferences regarding the real estate. The buyer need not enter all the preferences. He can choose to enter selectively. The un-entered preferences are filled by the system using statistical measures. The buyer need not know about the most influential features as the system takes care of it.

The system consists of two cascaded layers of regression models. The first layer consists of four different regression models whose prediction is used to train the model in the second layer. PCA is used to select the most influential features for training the model in the second layer. A comparative study is done against different combinations of feature selections used for generating the corpus for the second layer. A comparative study is also done against individual classifiers and cascaded classifiers. It turns out that cascaded classifiers perform better than the individual classifiers and feature selection using PCA outperforms other cascaded classifiers with an accuracy of 90.0926%.

References

[1] Omid Poursaeed, Tomas Matera and Serge Belongie ,*Vision-based Real Estate Price Estimation*,School of Electrical and Computer Engineering, Cornell University, 2017.

[2] Vishal Venkat Raman, Swapnil Vijay, Sharmila Banu K , *Identifying Customer Interest in Real Estate Using Data Mining Techniques* , International Journal of Computer Science and Information Technologies, 2014.