**Data preprocessing**

Data preprocessing is critical in machine learning to (1) reduce dimensions (2) strengthen relationships between features and target attributes. The goal is to remove irrelevant, redundant, noisy and unreliable data. In this project, we focus on (1) removing features containing excessive null values and outliers; (2) removing redundant/interdependent features; (3) categorizing features which by themselves are meaningless.

**Remove irrelevant data**

During the data mining process, we noticed that building types such as apartments, factory, and warehouse would dramatically obscure the relationship between predictive variables and target features. Therefore in this project, we focused on single houses by selecting property type id = 261/279. This reduced data size by ~18%.

A few features were very useful for visualization but less meaningful for machine learning. These includes latitude, longitude, assessmentyear and we removed these features.

**Remove features containing excessive deviating values**

Macintosh HD:Users:swuser:Downloads:CSE6242_Project:data:Zillow:missing_value.pdfBy counting the ratio of entities that contain missing values for each feature, we identified 26 out of 57 features where more than 50% of entities are missing. Removing these features significantly reduced the dimension of our dataset. Three features, buildingqualitytypeid, unitcnt and heatingorsystemtypeid have ~1/3 missing values. Buildingqualitytypeid and heatingorsystemtypeid has 5 and 4 major classes each so that we can categorize this feature and including missing value as another class. Unitcnt has only one dominant class = “1” so we removed this feature too. The other features typically have much less than 1% missing values. We simply removed the entities which contained these missing values. Alternatively, we could perform clustering and fill in the median/mean values of corresponding clusters.

Propertylandusetypeid only has one value 261, so we removed this feature.

**Remove redundant/interdependent features**

We found pairs of features that were highly redundant or indeterpdendent. We only maintained one feature shown by red.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Correlation | Feature 1 | Feature 2 | Feature 3 | Feature 4 |
| 1 | bathroomcnt | calculatedbathnbr | fullbathcnt |  |
| 1 | calculatedfinishedsquarefeet | finishedsquarefeet12 |  |  |
|  | regionidzip | regionidcounty | regionidcounty | fips |
| ~0.95 | taxvaluedollarcnt | structuretaxvaluedollarcnt | landtaxvaluedollarcnt | taxamount |

**Other features**

Propertycountylandusecode

Propertyzoningdesc has too many values without dominant classes.

Rawcensustractandblock and censustractandblock

Regionidneighborhood

It is unclear how these features above relate with house value. We will use training house value to study the relationship score between these features and house value and decide how to process them.

**Categorize features**

During data mining, we found that some features might strongly affect house values but using the original format is meaningless or misleading, suggesting that certain transformation is required. These includes buildingqualitytypeid, heatingorsystemtypeid and regionidzip. For the first two features, we chose the dominant classes and used one-hot encoding strategy to categorize them. Regionidzip has too many values with similar representation. We clustered it according to the corresponding house value and binned it to a few categories (to be done).