Midterm: Working With Large Data Sets and Machine Learning

Team 4: Emily Strong and Raksha Kaverappa

Code Repository: <https://github.com/erstrong/INFO-7390-ADS-Fall-17-TeamNo.4/tree/master/Mid-Term>

**Part 1: Data Ingestion, EDA, Wrangling**

AWS Bucket:

EDA Notebook: https://github.com/erstrong/INFO-7390-ADS-Fall-17-TeamNo.4/blob/master/Mid-Term/Part%201%20Exploratory%20Data%20Analysis.ipynb

**Data Cleansing Design and Implementation:**

**Data Cleaning Analysis:**

Our first step in missing data analysis was to determine the percent missing data for each feature. The following features have more than 80% missing data and cannot be calculated so we are dropping them:

* 'architecturalstyletypeid'
* 'basementsqft'
* 'buildingclasstypeid'
* 'decktypeid'
* 'threequarterbathnbr'
* 'finishedfloor1squarefeet'
* 'finishedsquarefeet6'
* 'finishedsquarefeet13'
* 'finishedsquarefeet15'
* 'finishedsquarefeet50'
* 'poolsizesum'
* 'storytypeid'
* 'typeconstructiontypeid'
* 'yardbuildingsqft17'
* 'yardbuildingsqft26'

Since in part 4 we need to use Latitude and Longitude to calculate distances, we chose to drop any rows with NaN in those columns. After doing this, the following features did not have any missing data and were not eliminated for other reasons: bedroomcnt, fips, parcelid (key), propertylandusetypeid, assessmentyear, taxamount, taxvaluedollarcnt, transactiondate.

The following features represent physical attributes of the properties and have skewed data because of the real-world frequencies and differences between condos, small houses, large houses, etc. For these we have substituted missing values with the median:

* 'calculatedbathnbr'
* 'calculatedfinishedsquarefeet'
* 'numberofstories'
* 'yearbuilt'
* 'taxvaluedollarcnt'
* 'structuretaxvaluedollarcnt'
* 'landtaxvaluedollarcnt'
* 'taxamount'

|  |  |
| --- | --- |
| **Feature** | **Decision** |
| 'airconditioningtypeid' | Per the 2005 study cited in this article most homes in the Los Angeles area do not have air conditioning so we are replacing missing values with the value for "None": <http://articles.latimes.com/2006/jul/28/business/fi-air28> |
| 'bathroomcnt' | Per the data dictionary this is redundant with calculatedbathnbr and the two columns are highly correlated. This column has no missing values but it does have 0s which we assume is impossible, and the 0s match to NaNs in the other column. We are dropping this one. |
| 'buildingqualitytypeid' | We are replacing missing values with the mean. |
| 'finishedsquarefeet12' | This has the same definition as calculatedfinishedsquarefeet, the two are highly correlated, and this column has more missing data so we are dropping it. |
| 'fireplacecnt' | There is more than 80% missing data, but there are no 0s. We assume a NaN is a 0. |
| 'fireplaceflag' | There is more than 80% missing data and count does not match that for 'fireplacecnt'. Recalculating based on 'fireplacecnt'. |
| 'fullbathcnt' | There is more than 80% missing data and the column is correlated with calculatedbathnbr so we are dropping it. |
| 'garagecarcnt' | There are no 0s so we are replacing missing values with 0. |
| 'garagetotalsqft' | For properties with a garagarcarcnt > 0 we are replacing missing values and 0s with the median. For properties with a garagecarcnt=0 we are setting this value to 0. |
| 'hashottuborspa' | Though there is more than 80% missing data, there are no False values, and the count for True matches pool type 2 + pool type 10. We are filling missing values with False. |
| 'heatingorsystemtypeid' | This is categorical so we are replacing the missing values with the mode. |
| 'lotsizesquarefeet' | Data is skewed so we are replacing missing values with the median. There are two outlier values (3589145 and 6971010) that repeat suggesting these numbers are automatically generated by some sort of data entry error. Replacing these two values with the median. |
| 'poolcnt' | There are no 0s, and the count is slightly over the sum of counts for pool types 2 and 7. We are filling missing values with 0. |
| 'pooltypeid10' | This is a one-hot of pool type. We are combining the pool types into a single column for our exploratory data analysis, calculated from poolcnt and hashottuborspa. |
| 'pooltypeid2' | This is a one-hot of pool type. We are combining the pool types into a single column for our exploratory data analysis, calculated from poolcnt and hashottuborspa. |
| 'pooltypeid7' | This is a one-hot of pool type. We are combining the pool types into a single column for our exploratory data analysis, calculated from poolcnt and hashottuborspa. |
| 'propertycountylandusecode' |  |
| 'propertyzoningdesc' | Since zoning is done by individual counties, these descriptions would not generalize. There are also too many unique values for this to be useful as a categorical variable. |
| 'rawcensustractandblock' | We are converting this to a "calculated" value to get rid of meaningless trailing digits. There are no missing values after the lat/lon NaNs are dropped. |
| 'censustractandblock' | Some of the values are impossible and this is redundant with 'rawcensustractandblock'. We are dropping this column. |
| 'regionidcounty' | This is redundant with the FIPS codes. We are dropping this column. |
| 'regionidcity' | Per the data dictionary not all properties will have a value in this column. |
| 'regionidzip' | The values in this column are incorrect/impossible. We are recalculating as a new column 'zipcode' and dropping this one. |
| 'regionidneighborhood' | There are too many unique values for this to be useful as a categorical variable and the values are Zillow's internal coding, they do not have real-world meaning that can be matched to other data sets. We are dropping this column. |
| 'roomcnt' | There must be at least as many rooms as there are bathrooms and bedrooms. We are calculating missing values as the sum of those columns. |
| 'unitcnt' | Data is skewed so we are replacing missing values with the median. Outliers appear to be the result of user error – the properties are condos and planned developments based on the propertylandusetypeid, and the unitcnt appears to be for the entire development whereas all other data reflects individual units. We are changing outliers to the median. |
| 'taxdelinquencyflag' | There are only 'Y' values, so though there are more than 80% missing data, we assume those are all 'N' and are filling missing values with that. Confirmed that 'Y' count matches the count for 'taxdelinquencyyear' |
| 'taxdelinquencyyear' | More than 80% have NaN but the count matches that for 'taxdelinquencyflag', so these NaNs are accurate – these properties have never had a value for this column. Due to the small size of the data we are dropping this column. |

**Part 2: Build A Prediction Model**

Model Notebook:

**Multiple Linear Regression Design:**

**Random Forests Design:**

**Neural Networks Design:**

**Analysis:**

**Part 3: Model Deployment**

Azure Webservice:

API Notebook:

**Design and Implementation:**

**Analysis:**

**Part 4: Geospatial Search**

Azure Webservice:

API Notebook:

**Design and Implementation:**

**Analysis:**