MLEND YANQING CAPSTONE PROPOSAL

June 1, 2018

1 Machine Learning Engineer ND Capstone Proposal

Yanqing Zhu May 31st 2018

1.1 Domain Background

Bay area is a expensive place to live, especially for housing, purchasing or selling a home is no small deal, the correct buy/sell action at the right time, might be more profitable than working for 20 years. I'm hoping to predict the actual home sale price with all housing related feature plus tech sector stock price for bayarea's house price.

1.2 Problem Statement

1. You are provided with a full list of real estate properties in three counties (Los Angeles, Orange and Ventura, California) data in 2016.
2. The train data has all the transactions before October 15, 2016, plus some of the transactions after October 15, 2016.
3. The test data in the public leaderboard has the rest of the transactions between October 15 and December 31, 2016. The rest of the test data, which is used for calculating the private leaderboard, is all the properties in October 15, 2017, to December 15, 2017. This period is called the "sales tracking period", during which we will not be taking any submissions.
4. You are asked to predict 6 time points for all properties: October 2016 (201610), November 2016 (201611), December 2016 (201612), October 2017 (201710), November 2017 (201711), and December 2017 (201712).
5. Not all the properties are sold in each time period. If a property was not sold in a certain time period, that particular row will be ignored when calculating your score.
6. If a property is sold multiple times within 31 days, we take the first reasonable value as the ground truth. By "reasonable", we mean if the data seems wrong, we will take the transaction that has a value that makes more sense.

1.3 Datasets and Inputs

All data will come from [https://www.kaggle.com/c/zillow-prize-1/data]

1. properties_2016.csv - all the properties with their home features for 2016. Note: Some 2017 new properties don't have any data yet except for their parcelid's. Those data points should be populated when properties_2017.csv is available.

2. properties_2017.csv - all the properties with their home features for 2017 (released on 10/2/2017)
3. train_2016.csv - the training set with transactions from 1/1/2016 to 12/31/2016
4. train_2017.csv - the training set with transactions from 1/1/2017 to 9/15/2017 (released on 10/2/2017)
5. All features

- 'airconditioningtypeid'
- 'architecturalstyletypeid'
- 'basementsqft'
- 'bathroomcnt'
- 'bedroomcnt'
- 'buildingqualitytypeid'
- 'buildingclasstypeid'
- 'calculatedbathnbr'
- decktypeid'
- threequarterbathnbr'
- finishedfloor1squarefeet'
- calculatedfinishedsquarefeet'
- finishedsquarefeet6'
- finishedsquarefeet12'
- finishedsquarefeet13'
- finishedsquarefeet15'
- finishedsquarefeet50'
- fips'
- fireplacecnt'
- fireplaceflag'
- fullbathcnt'
- garagecarcnt'
- garagetotalsqft'
- hashottuborspa'
- heatingorsystemtypeid'
- latitude'
- longitude'
- lotsizesquarefeet'
- numberofstories'
- parcelid'
- poolcnt'
- poolsizesum'
- pooltypeid10'
- pooltypeid2'
- pooltypeid7'
- propertycountylandusecode'
- propertylandusetypeid'
- propertyzoningdesc'
- rawcensustractandblock'
- censustractandblock'
- regionidcounty'

- regionidcity'
- regionidzip'
- regionidneighborhood'
- roomcnt'
- storytypeid'
- typeconstructiontypeid'
- unitcnt'
- yardbuildingsqft17'
- yardbuildingsqft26'
- 'yearbuilt'
- 'taxvaluedollarcnt'
- 'structuretaxvaluedollarcnt'
- 'landtaxvaluedollarcnt'
- 'taxamount'
- 'assessmentyear'
- 'taxdelinquencyflag'
- 'taxdelinquencyyear'

1.4 Solution Statement

A linear regression model with PCA will be implemented using keras/scikit, additional feature might be added in, the bayared model will be trained using the model trained from all states with additional stock information.

1.5 Benchmark Model

leader board model (global score)

1.6 Evaluation Metrics

Mean Absolute Error between the predicted log error and the actual log error. The log error is defined as logerror=log(Zestimate)log(SalePrice) and it is recorded in the transactions training data. If a transaction didn't happen for a property during that period of time, that row is ignored and not counted in the calculation of MAE.

1.7 Project Design

From the info above, we can inferred that a linear model would be best to fit this question, with additional tech sector stock market as an additional feature, we will start with some standardlization and feature extraction PCA, after that will be training using logistic regression with regularization, if model not performing well, probably will start with half trainied gold model from kaggle then move forward to retrain specifically for Bay are