Capstone Proposal

Kaggle Competition : Zillow Prize: Zillow's Home Value Prediction (Zestimate)
Can you improve the algorithm that changed the world of real estate?

Domain Background

A home is often the largest and most expensive purchase a person makes in his or her lifetime. Ensuring homeowners have a trusted way to monitor this asset is incredibly important. Zillow's Zestimate home valuation was created to give consumers as much information as possible about homes and the housing market, marking the first time consumers had access to this type of home value information at no cost. "Zestimates" are estimated home values based on 7.5 million statistical and machine learning models that analyze hundreds of data points on each property. And, by continually improving the median margin of error (from 14% at the onset to 5% today), Zillow has since become established as one of the largest, most trusted marketplaces for real estate information in the U.S. and a leading example of impactful machine learning.

I am personally interested in this Kaggle challenge as it's a prestigious competition which is being participated by people across the world and has a huge prize money of $1,200,000. Zillow Prize, a competition with a one million dollar grand prize, is challenging the data science community to help push the accuracy of the Zestimate even further. Winning algorithms stand to impact the home values of 110M homes across the U.S.

Kaggle Competition Details:
https://www.kaggle.com/c/zillow-prize-1

Problem Statement

The problem statement is to build a model to improve the Zestimate residual error by engineering new features that give model an edge over the competition.

Datasets and Inputs

All the real estate transactions in the U.S. are publicly available. We are provided with a full list of real estate properties in three counties (Los Angeles, Orange and Ventura, California) data in 2016. The train data has all the transactions before October 15, 2016, plus some of the transactions after October 15, 2016. The dataset is available in the Kaggle website under the competition “ Zillow Prize: Zillow’s Home Value Prediction (Zestimate)”. The shape of the training data set is 90275,3 and that of the properties data set is 2985217,58. The shape of the merged dataset is 90275, 60. Target variable for this competition is variable "logerror". There are few outliers but the logerror has a normal distribution of data. Out of the 60 variables; 53 are float, 5 objects, 2 ints and 1 datatime. The target variable is no unbalanced, also we will have to convert objects to integers by transforming the data before feeding the same to the model.

Link to Dataset :
https://www.kaggle.com/c/zillow-prize-1/data

Solution Statement

For each property (unique parcelid), we must predict a log error for each time point. As part of the competition, The test data logerror values are private, so we cannot calculate the score locally. Hence, we will estimate score locally by using cross validation, where the data is sliced into N slices and the model’s performance is averaged across these slices.

Before feeding the data to the model, we will preprocess the data by handling missing data, removing outliers and transforming data where required after doing a univariate and multivariate analysis of the data. Based on analysis of the data, we will remove data which has more than 85% of missing data, impute data mean values and h add some features based on domain knowledge. We will be using gradient boosting models such as XGBoost or Light GBM.

Benchmark Model

The Benchmark score for my model using RandomForestRegressor is a mean score of -0.08274 and a standard average score of 0.00539.

Evaluation Metrics

Submissions are evaluated on 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.

Project Design

My project design will be as follows :

a. Exploratory Analysis of the data. This will include Missing value analysis, Correlation analysis, Univariate and Bivariate analysis of the data. Distribution of the Numerical data will be done using histograms, Categorical data through Bar plots and Correlations will be studied through Heatmaps.

b. Data preprocessing and cleaning will handling missing data by either dropping observations (null values above a certain percentage) , imputing missing data with mean values or -999, removing outliers etc

c. Feature Engineering will including adding new fields using domain knowledge or creating dummy variables, remove redundant feature from the dataset etc.

d. Model Selection : We will be using gradient boosting models such as XGBoost or Light GBM.

e. Model Training will include gridsearch on hyperparameters and cross-validation of the data using k-fold cross validation methods with performance metric as Mean Absolute Error (MAE).