Team member's details:

Anusha Ramamurthy anuramam@umail.iu.edu

Vandana Kolli kolliv@iu.edu

Project 1: House Prices: Advanced Regression Techniques

You will need more in introduction. What would be a motivating real life problem?

is project is to predict the final price of each home With 79 explanatory
g (almost) every aspect of residential homes in Ames, Iowa.

Data Analysis:

The housing data set has 1460 rows and 81 features with the target feature Sale Price. As part of data analysis we made the following observation on categorical and non-categorical data.

Missing values in the data:

You will want to put this into a more narrative form.

e categorical variables with the largest number of missing values are: Alley, FirePlaceQu, PoolQC, Fence, and MiscFeature.

- The numeric variables do not have as many missing values but there are still some present. There are 259 values for the LotFrontage, 8 missing values for MasVnrArea and 81 missing values for GarageYrBlt.
- We applied a technique where missing values are replaced by its mean.

Future Implementation Ideas:

We are trying to identify a better method to replace missing values. Replacing by mean doesn't make sense, since the values like LotFrontage may depend on the street where the house is located.

Apply KNN to identify how to replace the missing values.

Add categorical data as a separate column to be 0 for absent, 1 for presence and check how the linear regression values change.

Feature Extraction: Identify features that strongly influence the regression line, and eliminate features that don't influence the Sale Price. Thus reducing the feature set.

Apply other algorithms like SVM or Random forests instead of Linear Regression to check if the sale price improves.

Histograms:

We tried to plot different histograms to get an idea of how the sale price is impacted by various attributes.

Algorithm:
We used some part of train data to train the system and rest of the train data (last 20 records) to test the actual sale price to predicted sale price.
We applied linear regression on the clean data to predict sale price.

Questions:

Try few different ones and see which ones perform bests.

1. We are confused about which algorithm to use for predicting the house price – we have been working on linear regression, but thinking of KNN and random forest also.
2. How do we use Categorical Data in linear regression? We are confused how they will influence the regression line.
3. Is clustering a better method for this dataset?

Come to office hours if you need help with that still

We have the labels, why would you want to use clustering? Clustering can be used for missing values though.

Project [illegible] Prediction

In Outbrai[illegible] are expected to predict which adds the users are more likely to cl[illegible]mmendation algorithm is needed.

Data Ana
Dataset: Because the dataset is huge, we are trying to figure out a way on how and where to use the dataset. Also we are trying to understand different attributes in different files.

database seems like a good choice, perhaps working with a sample of the original dataset would be easier as well.

Questions:

1. For huge datasets like this one, what [illegible]ze it – we are considering postgresql, SQL Server are there any other better options?
2. How to interpret confidence values?
3. There are duplicate values, which values should we retain?
4. Can we consider this to be a clustering or Estimation Maximization problem?