CSCI59000 Big Data Management Project Proposal

Project Title: Zillow's Home Value Prediction [1]

Team Members:

Jingyi Hui hui@iu.edu

Tianze Jiang

Jasvir Kaur

Tejaswi Parasa

Dataset:

Name	Size	Shape
Properties_2016.csv	648.8MB	2.99m * 58
Properties_2017.csv	650.1MB	2.99m * 58
train__2016_v2.csv	2.5MB	90.3k * 3
train__2017.csv	2.8MB	77.6k * 3

Objectives:

A home is often the largest and most expensive purchase a person makes in his or her lifetime. Our aim is to analyze hundreds or even thousands of data points on each property and estimate home values based in statistical and machine learning models.

Research Approach:

- Perform data cleansing and preprocessing
- Use tools like Parallel Coordinates, Radial Boxplots and Box Plots to visualize and get a brief understanding of the dataset
- Find out key features of the dataset and apply different Machine Learning algorithms with Hadoop Spark
- Evaluate the prediction with Mean Absolute Error. For each property, we will predict a log error for each time point.

Deliverables:

- Source code
- Project report (models applied, evaluation and visualization graphs)
- Power Point presentation slides

References: [1] https://www.kaggle.com/c/zillow-prize-1