Milestone 1

Ethan Li, Saheli Saha

Indiana University Bloomington

Group name: Mesa

I. House Prices

The goal of this project is to predict the price of the houses in Ames, Iowa.We are almost done with the prediction and have some preliminary results. At first we did data visualization to have a better understanding about the data. Following that we cleaned the data by imputing missing values and normalized the data. We use sample mean to impute numerical data and sample mode to impute categorical data. In order to be able to do regression, we transformed the categorical data into dummy vectors. As a result we got more than 300 features. Apparently, we felt the necessity to reduce the dimension of the datas to save computing resources and running time. So we applied PCA as a feature selection algorithm to the cleaned data. Finally we applied prediction models, such as linear regression and regression forest. We got fair results and we will continue to work on it to get better accuracy.

The main challenge of this project is to make better prediction for the house price. We are still working on this to get better result by tuning different parameters and applying different algorithm.

II. Outbrain Click Prediction

The goal of this project is to rank the recommendations in each display group by decreasing predicted likelihood of being clicked. We are in the preliminary stage of this project. As the project is bit different than the conventional data mining projects we are using different perspective to deal with the large amount of data. Currently we are forming our ideas to deal with the project and experimenting with the same. The ideas are as follows:

1) Transform the problem to a regression problem. From the datasets, we can form a data table, with ad_id as the primary id and all information related to the ad_ids, such as document topics, document categories, and campaign_id as the features. The number of clicked times related to the ad_ids is the result we want to predict. Then we can sort ad_id in display_id in test data according to the predicted number of clicked times as the final results.

2) use collaborative filtering (CF) algorithm. Use user id and ad id to form a data matrix indicating the relation between users and ads. The CF algorithm predicts user's preferences by considering the similar interests between some other users. By applying the CF algorithm, we can attain the preferences of all ads to each of these users. It then can be used very naturally to the test data.

The main challenges are to deal with large datasets, and get a good prediction. To deal with large datasets, we can use some existing packages, and if necessary, implement distributed algorithms on clusters. For getting good prediction, we will need to do experiments and tune our algorithms.