# Team #29: Model Explanation

# Yuxuan Zhang, Yutong Wang, Min Li

## Goal

We built an XGBoost model to predict the revenue score of a zip code in different industries and explain the important attributes.

## Data Manipulation

The data we were given is consist of establishments and revenue information by zip code and industry. We only kept the observations with 3-digit NAICS code because the 4-digit and 5 digit NAICS codes are sub-categories of the 3-digit NAICS codes and we are only interested in a larger scale of the industry categories. The response variable, “revenue score”, was created by multiplying the median of the revenue level by the number of establishment. When building the model, we took a log on the response variable.

Our dataset also incorporated GIS information including zip code Tabulation area, population density in 2012 by zip code (population / area), average individual income by zip code in 2012, average state sales tax in 2012, median-age by zipcode in 2014 and unemployment rate by county in 2012.

## Model

XGBoost is a powerful model for both prediction and feature importance analysis. We set the number of trees to be default 100 and tuned the learning rate, column sample rate in each tree, max tree depth and lasso penalty term. We built the model with different combinations of the parameters and evaluated by 10-fold cross validation. We picked the combination which generates the smallest average root mean square error.

The final set of parameters is {learning\_rate = , colsample\_bytree = , max\_depth = , lamda = }. Then we built the model with a training dataset and evaluate the performance on the validation dataset. Finally, we used the model to predict on the whole dataset to predict the revenue score and get a list of zip code with the highest predicted revenue score.

## Result

|  |  |  |
| --- | --- | --- |
| Mean Score Value | Training Error | Testing error |
| 14.6840 | 0.6589 | 0.7791 |

We also found that population density and wage are the most important features for predicting the revenues. Area and median age are also important.