EECS 510: Preliminary Report

Influencers in Social Networks

Tian Zhang, Can Wang

Goal

We hope to predict which people are influential in Twitter via this project. Our task is to predict human judgement on who is more influential between the given pairs of Twitter users.

Data

Data Format

- Training Data: Here, we have 5500 samples. In each row, we have a pair of data belonging to two users (A and B), including numeric continuous attributes like count of followers, number of mentions received, etc. And in the first row, ”choice” (binary number) is the result that which user has more influence.
- Test Data: it has the same structure with Training Data but it lacks the result row: ”choice”. 5952 samples in total.
- Sample_predictions: It’s the data we submitted in the end. It has two columns: the first column is ID number and the second column is result (It’s represented by probability that “choice” equals 1 which means A has more influence)

Data Preprocessing

- To normalize the dataset, we make log transformation on data+1.
- Then, we get the result after subtracting attribute of B from that of A.
- In this condition, we get 11 columns storing subtracting result from last steps, and take this new dataset as training data.

Models

Baseline

We use ZeroR as our baseline model. This model is extremely naive, which will assign the majority of the choices in training data to all testing data as predictions.

Naive Guess

This guess model is based on the assumption that the user with higher Twitter features is more influential. There is no need to train a model, we just directly analyze the testing data. We normalize and compute the difference for each pair of corresponding features of A and B,

and get the sum of all differences as $sum = \sum \frac{feature_A - feature_B}{feature_A + feature_B}$. If the sum is greater than 0, then we consider A as the more influential one, and assign 1 to Choice; otherwise, we assign 0.

Logistic Regression (sklearn.linear_model.LogisticRegression)

- When the Y-dependent Variable has only two choices (0/1, yes/no, etc…), Logistic Regression is the appropriate model
- Logistic Regression returns the probability of the two choices
- Logistic Regression has a different format:

$$\log\left(\frac{p_i}{1-p_i}\right) = \alpha + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_p X_{ip}$$

- Can't solve with normal regression equations- but gives similar outputs

Regression Tree (sklearn.tree.DecisionTreeRegressor)

A Regression tree may be considered as a variant of decision trees, designed to approximate real-valued functions instead of being used for classification methods.

A Regression tree is built through a process known as binary recursive partitioning. This is an iterative process that splits the data into partitions or "branches", and then continues splitting each partition into smaller groups as the method moves up each branch.

- First, all records in the training set (the pre-classified records that are used to determine the structure of the tree) are grouped into the same partition.
- Second, the algorithm then begins allocating the data into the first two partitions or "branches", using every possible binary split on every field.
- The algorithm selects the split that minimizes the sum of the squared deviations from the mean in the two separate partitions. This splitting "rule" is then applied to each of the new branches.
- This process continues until each node reaches a user-specified minimum node size and becomes a terminal node. (If the sum of squared deviations from the mean in a node is zero, then that node is considered a terminal node even if it has not reached the minimum size.)

Evaluation

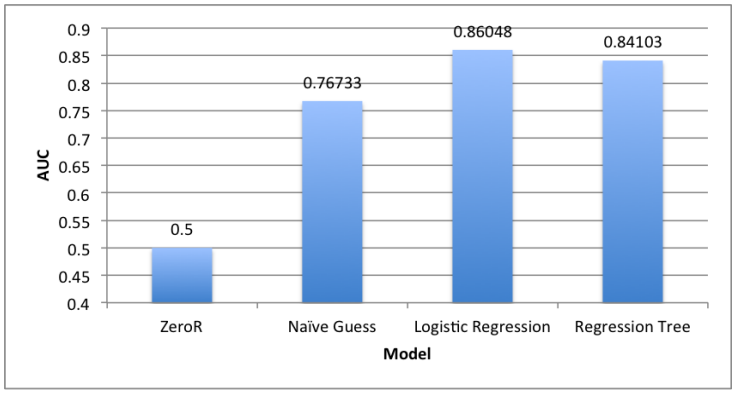
We evaluate the accuracy of our predictions with an AUC score computed by ROC.

In statistics, a receiver operating characteristic (ROC), or ROC curve, is a graphical plot, illustrating performance of a binary classifier system as its discrimination threshold is varied.

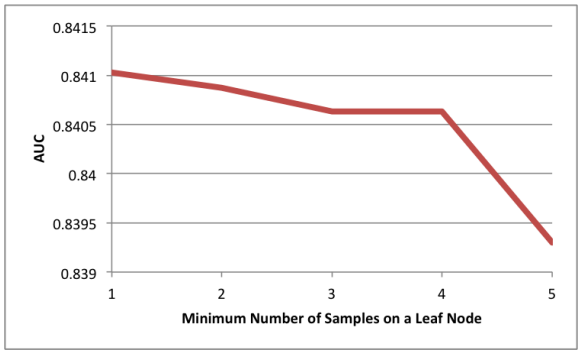
The curve is created by plotting the true positive rate against the false positive rate at various threshold settings. The ROC curve is sensitivity as a function of fall-out. ROC analysis is related in a direct and natural way to cost/benefit analysis of diagnosis decision making.

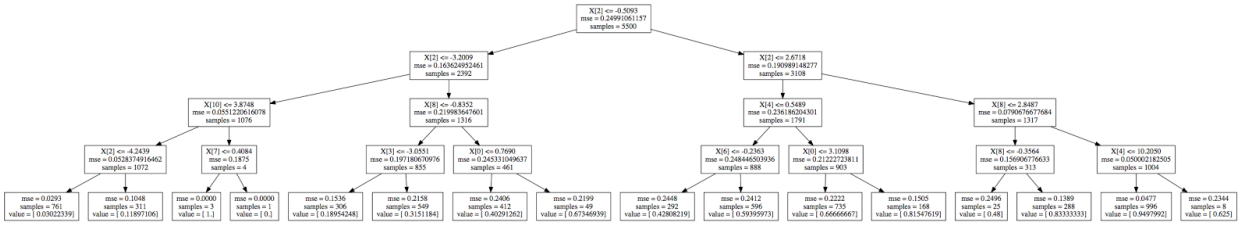
Results

The AUC score of each model is shown below. The Logistic Regression model has the highest AUC score, and then the Regression Tree model.



For the Regression Tree model, we tried different factors of tree depth and minimum number of samples on a leaf node, as shown below. When the tree depth increases, the model is probably overfitting, so the AUC decreases. And in this case, the less the minimum number of samples on a leaf node, the high the prediction accuracy. We also demonstrate a regression tree with max_depth = 4 and min_samples_leaf = 1.





Remaining Work

On the forum of Kaggle, we also find some other popular models used on this case, including Gradient Descent Boosting, SVM and etc. We will run two or more models and modify their factors, and then compare with the models we have now. Besides, we tend to implement cross-validation to improve the accuracy of our models.

** All materials of our project including source code are available on Github.*