CS 6350 Project Status Report

TalkingData AdTracking Fraud Detection

Students: hxz133530 yxh165530 zxl165030 txz160930

Kaggle Link: https://www.kaggle.com/c/talkingdata-adtracking-fraud-detection/

Goal:

To build an algorithm that predicts whether a user will download an app after clicking a mobile app ad.

Data fields:

Each row of the training data contains a click record, with the following features.

- ip: ip address of click.
- app: app id for marketing.
- device: device type id of user mobile phone (e.g., iphone 6 plus, iphone 7, etc.)
- os: os version id of user mobile phone
- channel: channel id of mobile ad publisher
- click_time: timestamp of click (UTC)
- attributed_time: if user download the app for after clicking an ad, this is the time of the app download
- is_attributed: the target that is to be predicted, indicating the app was downloaded.

The test data is similar, with the following differences:

- click_id: reference for making predictions
- is_attributed: not included

Data pre-processing:

- Standardization: Transform the data so that it has a mean of zero and a standard deviation.
- Scaling: A standard scaling transformation is to map the data from the original scale to a scale between zero and one. This is typically called data normalization.
- Remove Skew: Normally distributed data and can perform better if the skew is removed. Try replacing the attribute with the log, square root or inverse of the values.
- Feature engineering: Extract useful features for the training models.

Experimental methodology:

Build pipelines that contain data pre-processing and building machine learning models.

Use cross-validation to select best models which will be evaluated on area under the ROC curve between the predicted probability and the observed target.

Coding language: Scala

CS 6350 Project Status Report

Technique to be used:

We will cover most classification methods from Spark MLib library: Random Forest, GBT, Multilayer perceptron, SVM and Naïve Bayes.

We will then apply XGBoost to achieve a better ranking for the Kaggle competition. We will also explore Deep Learning methods if we still have time.

Schedule:

1. Prepare data
2. Implement Spark MLib Classification methods
3. Implement a XGBoost model
4. Implement a deep learning model
5. Training on the complete training dataset(5G) on EMR
6. Tuning
7. Submit result and report