CKME136 Capstone Presentation

TalkingData AdTracking Fraud Detection Challenge

https://github.com/PHLHY/Capstone

Phi Huynh
500777278
Monday, April 8, 2019

Kaggle Competition

Kaggle competition started on March 8, 2019 and ended on May 5, 2019

Prize money:

1st Place - $12,500

2nd Place - $7,500

3rd Place - $5,000

Competition found on this link: https://www.kaggle.com/c/talkingdata-adtracking-fraud-detection/overview

Research Question

TalkingData provided a dataset of around 200 million clicks spanning over 4 days for a Kaggle challenge with the aim of building an algorithm for determination of users that will download a mobile application after clicking on a relevant advertisement

As per the competition, evaluation is based on the area under the ROC curve with a submission file using the attributes, “click_id” and “is_attributed” from the testing set

Techniques for classification analysis will be employed on the R platform (RStudio) to tackle this challenge

TalkingData Datasets

Training and testing datasets are provided by TalkingData. The training set includes 184903890 data points with 8 attributes while the testing set includes 18790469 data points with 7 attributes

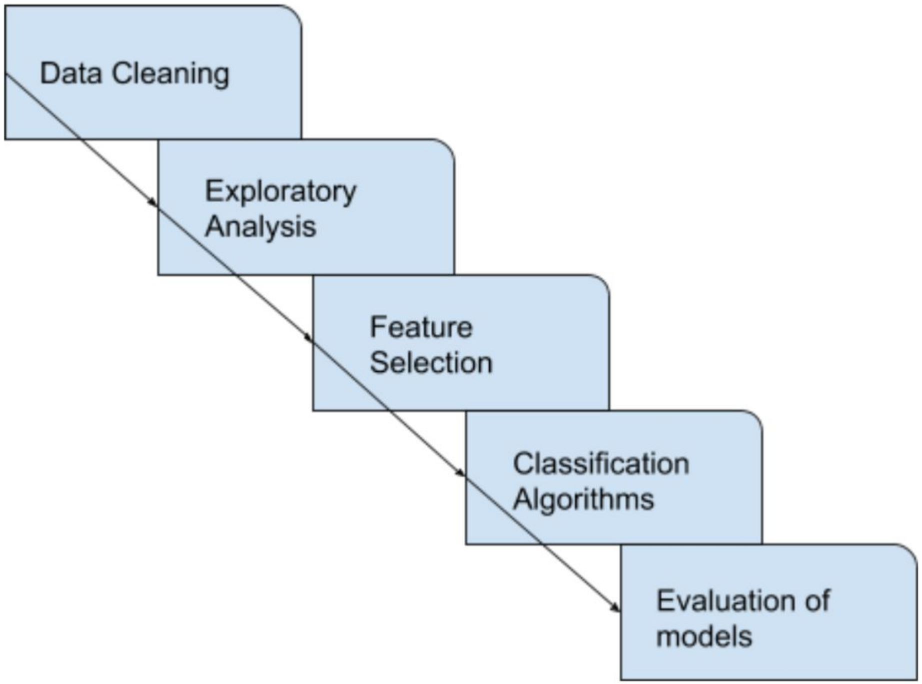
A direct copy of the description of the attributes was provided by Talking data and can be found from this link:
https://www.kaggle.com/c/talkingdata-adtracking-fraud-detection/data

- 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, huawei mate 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

Approach



Data Cleaning
Exploratory Analysis
Feature Selection
Classification Algorithms
Evaluation of models

Data Cleaning

No missing values

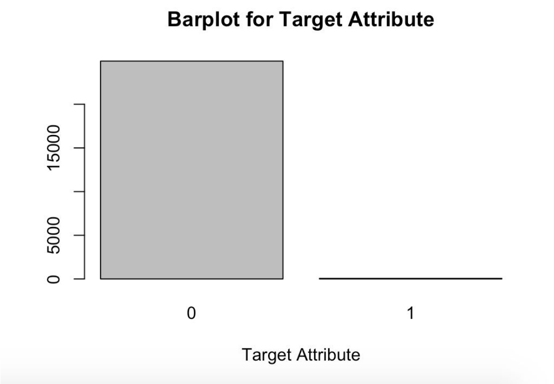
Data was reduced into 25000 observations

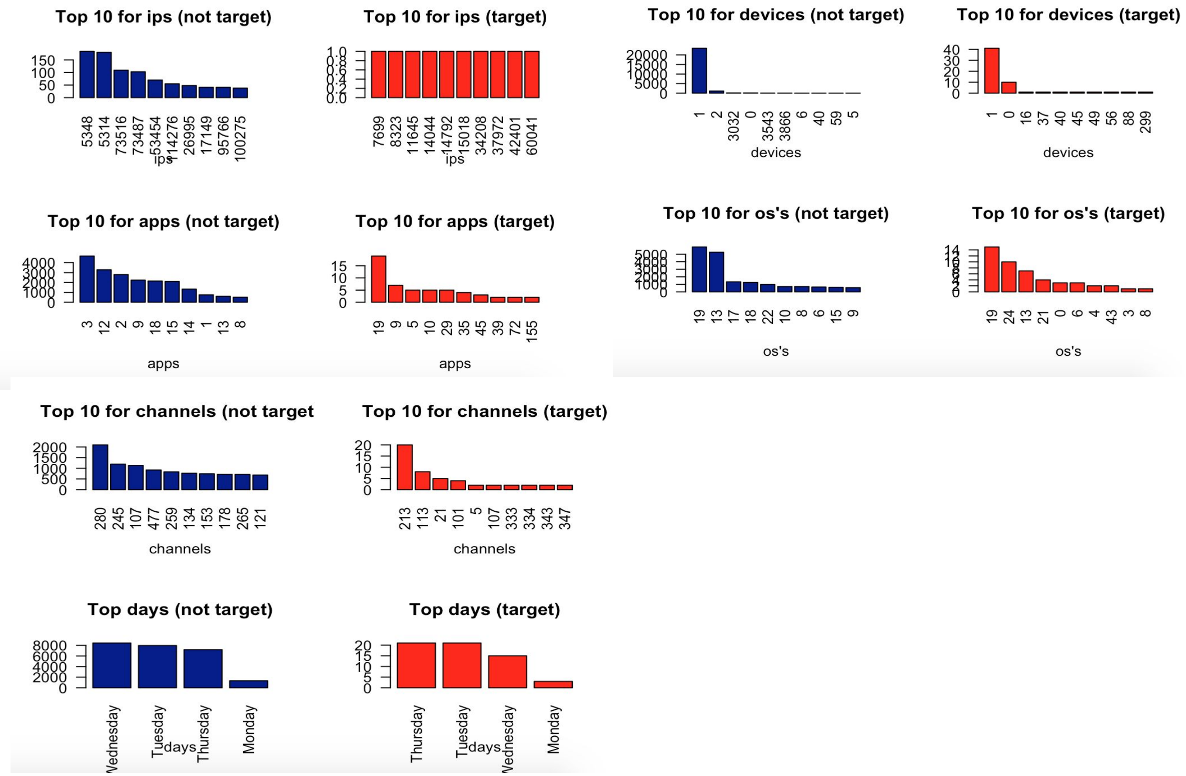
Target attribute, “is_attributed” converted to categorical data type

The attribute, “click_time” was split into data and time format

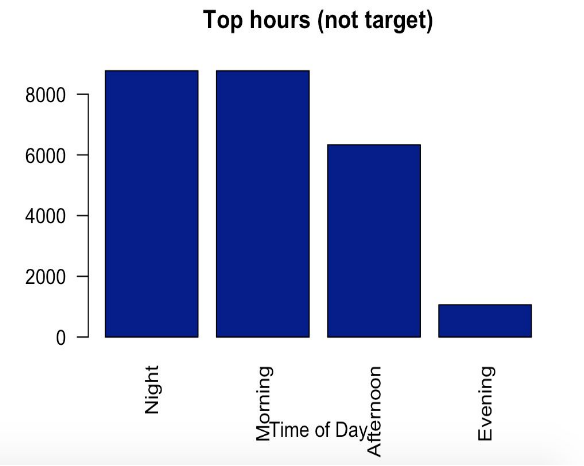
Redundancy in “attributed_time” and “is_attributed”

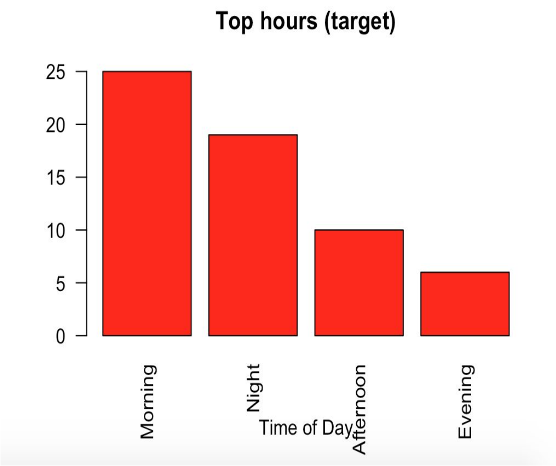
Exploratory Analysis

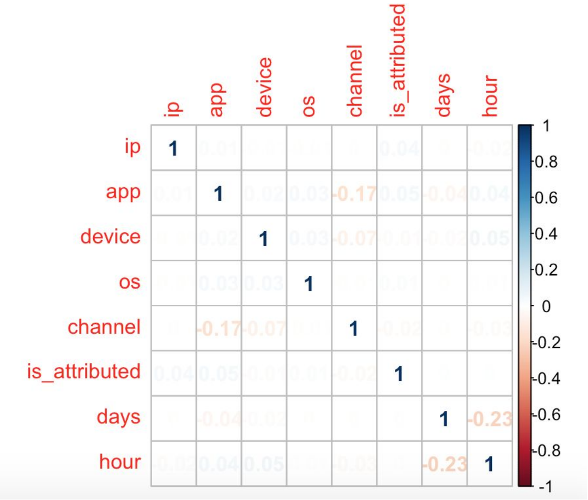




Exploratory Analysis (cont.)







Feature Selection

StepAIC - Forward direction

The follow attributes will be carried forward for the modelling process: “ip”, “app”, channel”, “os”, and with our target attribute, “is_attributed”

```
> summary(forward)

Call:
lm(formula = is_attributed ~ ip + app + channel + os, data = reduced.set)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.12450 -0.00415 -0.00184  0.00014  0.99995 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept) -5.851e-04  8.526e-04  -0.686 0.492544    
ip           3.910e-08  4.454e-09   8.779  < 2e-16 ***
app          1.702e-04  1.984e-05   8.582  < 2e-16 ***
channel     -8.531e-06  2.377e-06  -3.588 0.000334 ***
os          -1.533e-05  5.339e-06  -2.870 0.004105 ** 
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.04877 on 24995 degrees of freedom
Multiple R-squared:  0.006577,	Adjusted R-squared:  0.006418 
F-statistic: 41.37 on 4 and 24995 DF,  p-value: < 2.2e-16
```

Algorithms and Balancing Methods

Reduced set (25000 observations) split into testing set (70%) and validation set (30%)

Repeated cross validation done (5 folds and 3 repeats)

Constant seed set for all models in regard to the folds and resampling process for comparison of models

6 models produced based on combination of classification algorithms and balancing methods

Algorithms/Balancing	None	SMOTE	Upsampling
RandomForest	-	-	-
XGBoost	-	-	-

Evaluations of the Models

Confusion matrix statistics - Accuracy, sensitivity, specificity, precision, recall, and F1 score

Area Under the Curve (AUC) based on the Receiver Operating Characteristic (ROC) curve

Statistical significance - constant folds and resampling processes, paired T-tests, bonferroni correction

Results - Confusion Matrix Statistics

Confusion Matrix Statistics for No Balancing

	Random Forests	XGBoost
Accuracy	0.9983	0.9985
Sensitivity	0	0
Specificity	0.9997329	1
Precision	0	NA
Recall	0	0
F1	NA	NA

```
          Reference
Prediction   No Target
    No     7487     11
    Target    2      0

          Reference
Prediction   No Target
    No     7489     11
    Target    0      0
```

Confusion Matrix Statistics for SMOTE Balancing

	Random Forests	XGBoost
Accuracy	0.9447	0.9131
Sensitivity	0.909091	0.818182
Specificity	0.944719	0.913206
Precision	0.023585	0.013657
Recall	0.909091	0.818182
F1	0.045977	0.026866

```
          Reference
Prediction   No Target
    No     7074      1
    Target  415     10

          Reference
Prediction   No Target
    No     6839      2
    Target  650      9
```

Confusion Matrix Statistics for Oversampling

	Random Forests	XGBoost
Accuracy	0.9973	0.9953
Sensitivity	0.1818182	0.4545455
Specificity	0.9985312	0.9961277
Precision	0.1538462	0.1470588
Recall	0.1818182	0.4545455
F1	0.1666667	0.2222222

```
          Reference
Prediction   No Target
    No     7478      9
    Target   11      2

          Reference
Prediction   No Target
    No     7460      6
    Target   29      5
```

Results - AUC ROC

AUC ROC Results			
	None	**SMOTE**	**Oversampling**
Random Forests	0.9429	0.97	0.9314
XGBoost	0.9687	0.9351	0.9726

Results - Statistical Significance on the ROC

Main comparison model - Upsampling XGBoost algorithm

Noted to be statistically different to two other models (Unbalanced RandomForest and Upsampling RandomForest)

```
p-value adjustment: bonferroni
Upper diagonal: estimates of the difference
Lower diagonal: p-value for H0: difference = 0

ROC
          RF_NB     XGB_NB    RF_SMOTE  XBG_SMOTE RF_UP     XGB_UP
RF_NB               -0.094942 -0.036986 -0.070935 -0.045130 -0.080962
XGB_NB    0.0010032            0.057956  0.024007  0.049812  0.013980
RF_SMOTE  0.1038082 0.0003148           -0.033949 -0.008144 -0.043976
XBG_SMOTE 0.0190770 0.0348973 0.0351131            0.025805 -0.010027
RF_UP     0.5750569 0.0040562 1.0000000 1.0000000           -0.035832
XGB_UP    0.0212108 1.0000000 0.0665815 1.0000000 0.0421181
```

Statistical Differences Between Models (P-Values) - ROC Metric	
Main Model for Comparison = Oversampling XGBoost	
	Oversampling
NB - RF	2.12E-02
NB - XBG	1
SMOTE - RF	6.66E-02
SMOTE - XBG	1
OVER - RF	0.0421181

Conclusion

Main method of evaluation for this competition was the area under the ROC curve

Thus, upsampling XGBoost model will be the choice selection for this Kaggle challenge

- Based on the highest AUC - ROC score and is significantly different compared to two other models
- In addition, with consideration of the statistics from the confusion matrix, this model has the second highest precision score and F1 score

Limitations and Future Considerations

Computational limitations

- Dataset was reduced significantly to only 25000 points for training and testing
- Some of the attributes were not converted to categorical data type

Further exploratory analysis and feature engineering

- “Attributed_time” for feature engineered
- Potentially all attributes could have been included in the models for benchmarking

Modelling

- Preprocessing was not done
- Could have considered different types of balancing (ie. ROSE, undersampling etc.)