GA Customer Revenue Prediction

Objective

Predict revenue

Data

train.csv
test.csv
sample_submission:

What to Predict?

Natural log of the sum of all transactions per user.

Evaluation

RMSE

Work Process

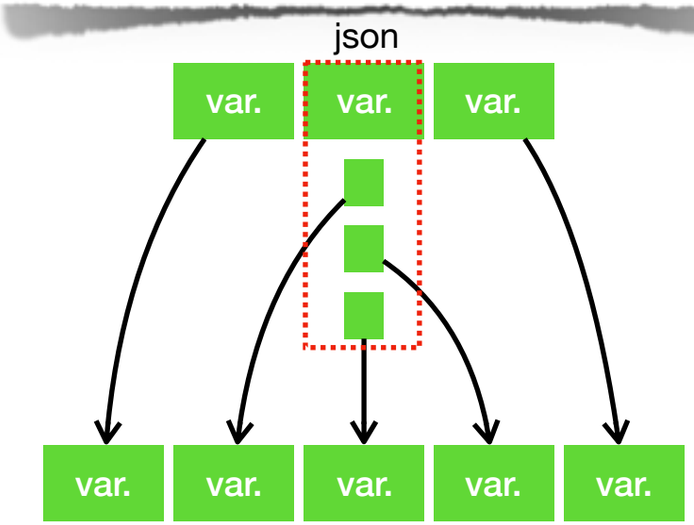
Data management
Exploratory data analysis
Algorithm
Featuring
Modeling
Predict & submission
Summary
Recommendation
Reporting

Software Program

R programing

Data Management

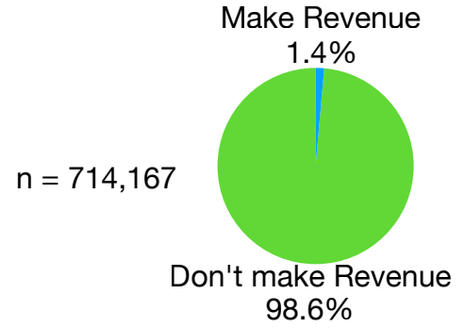
1. Read data file : 12 variables (4 variables are json)
2. Manage json column : device, geoNetwork, totals and trafficeSource (un-nest). got 55 variables.
3. Clean data : “(not set”), “(not provided)”, “not available in demo dataset” etc.
4. Summaries each user, group by “fullVisitorId”
   - label variable - count
   - numeric variable - sum
5. Now, each row represented for unique customer



Number of row before and after group by user

	Train	Test	Submission
Before group by user	903,653	804,684	617,242
After group by user	714,167	617,242	617,242

Exploratory Data Analysis

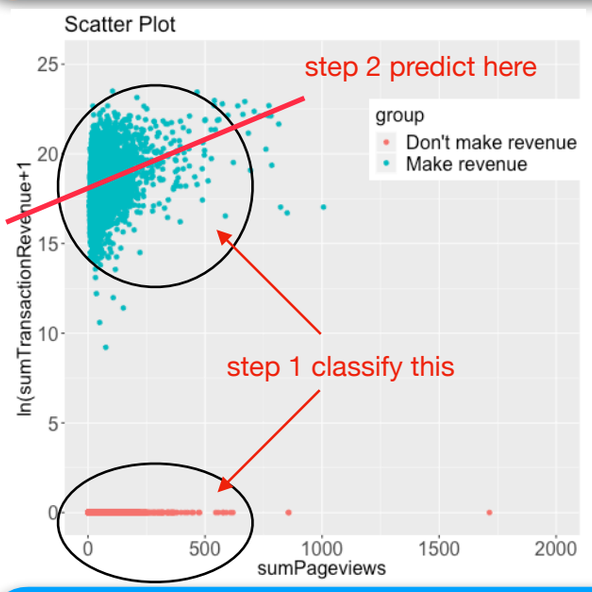


Where the Revenue come from
(Base on customer connection)

Device	Make Revenue	Don’t Make Revenue	n
DeskTop	1.7	98.3	523,690
Mobile	0.5	99.5	190,477
Channel			
Affiliates	0.1	99.9	13,400
Direct	2.0	98.0	109,830
Display	6.7	93.3	4,103
Organic	1.2	98.8	311,607
Paid	3.6	96.4	18,702
Referral	7.4	92.6	65,611
Social	0.1	99.9	212,374
Continent			
Africa	0.1	99.9	13,,488
America	3.0	97.0	323,208
Asia	0.1	99.9	196,416
Euope	0.1	99.9	167,966
Oceania	0.1	99.0	12,901

- **Overall, Gstore can make revenue only 1.4% of user.**
- **GStore have a chance to make revenue from desktop user more than mobile user.**
- **GStore have a chance to make revenue from American user more than other continent.**

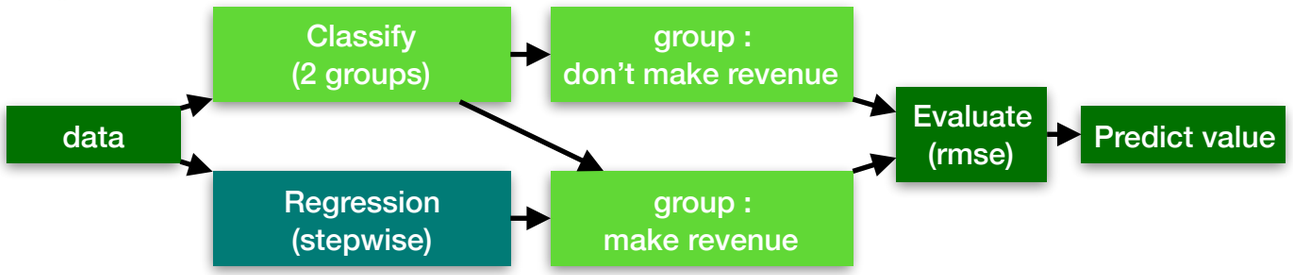
Algorithm



Algorithm 1 : Regression

data → select data (r2) → Regression (stepwise) → Evaluate (rmse) → Predict value

Algorithm 2 : Classify and Regression



Modeling

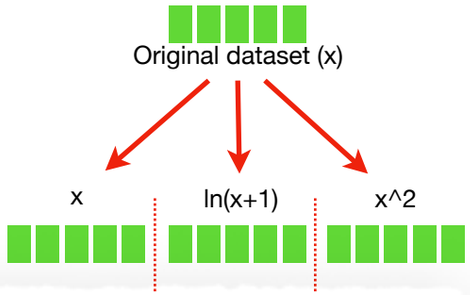
3 Models

1. Linear regression model (LM) use stepwise to select variable.
2. Logistic linear regression and LM (GLM_LM): use default parameter
3. Generalized boosted regression with LM (GBM_LM): defined parameter.

Use RMSE to evaluate and compare performance of model.

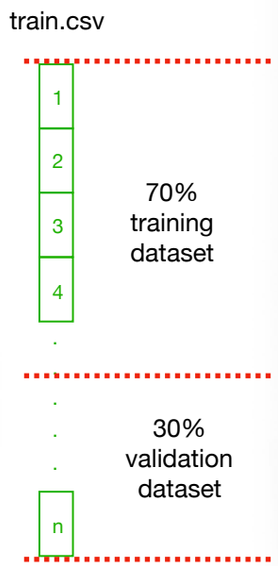
Data Featuring

Data featuring : make more feature. Firstly, try log(x+1), x^2



- We can make a lot of feature such as ratio, sd, (x-min)/ (max-min) etc.
- We use EDA help to select the features, such as correlation, mean difference between group, scatterplot etc.

Data Preparering



1. Divide data to 70:30, 70%, is training dataset, use to train the model. 30% is validation dataset, use to validate the model.
2. Set and train model and predict on the validation dataset.
3. Use RMSE to evaluate the model., tune parameter and try to find low RMSE.
4. Use the model that lowest RMSE predict on test dataset.

Predict & Submission

Result : Model Performance

method	Classified	Predict	Rmse Validation	Submission Score
(1)	-	LM	1.7735	**1.5822**
(2)	GLM	LM	2.1901	**2.0071**
(3)	GBM	LM	1.9451	**1.6975**
(4)	Combine	(1)+(2)	-	**1.6650**
(5)	Combine	(1)+(3)	-	**1.5757**
(6)	Combine	(2)+(3)	-	**1.7073**

Summary

- After a lot of work to do with limitation of time and hardware, found LM model is better than GLM with LM model and GBM with LM model.
- However, when try to use average of prediction value from model, some is better than original model.

Recommendation

- Dataset is imbalance (a lot of zero value) : bias parameter estimation when use LM model.
- Algorithm that classified first then predict revenue from user who trend to make revenue is make sense and explainable.
- Classification method : GLM may not suitable due to imbalance dataset.
- Try to tune parameter in GBM model.
- Try more feature data.
- Use others classified model.
- Beware more error : from classified and prediction.