Coupon Purchase Prediction

Dmitry Efimov and Lucas Silva

October 10, 2015

Outline

Provided data

Data preprocessing

Likelihoods

Final model and results

Competition

Completed • $50,000 • 1,076 teams

Coupon Purchase Prediction

Thu 16 Jul 2015 – Wed 30 Sep 2015 (8 days ago)

Dashboard

Home
- Data
- Make a submission

Information
- Description
- Evaluation
- Rules
- Prizes
- Timeline

Forum

Scripts
- New Script

Leaderboard
- Public
- Private

My Team

Competition Details » Get the Data » Make a submission

Predict which coupons a customer will buy

Recruit Ponpare is Japan's leading joint coupon site, offering huge discounts on everything from hot yoga, to gourmet sushi, to a summer concert bonanza. Ponpare's coupons open doors for customers they've only dreamed of stepping through. They can learn difficult to acquire skills, go on unheard of adventures, and dine like (and with) the stars.

Investing in a new experience is not cheap. We fear wasting our time and money on a product or service that we may not enjoy or fully understand. Ponpare takes the high price out of this equation, making it easier for you to take the leap towards your first sky-dive or diamond engagement ring.

Using past purchase and browsing behavior, this competition asks you to predict which coupons a customer will buy in a given period of time. The resulting models will be used to improve Ponpare's recommendation system, so they can make sure their

Provided data

- Train: **user-coupon purchases** for 52 weeks
- Train: **user-coupon visits** for 52 weeks
- **User list**: gender, age, locations
  - $\approx$ 20 000 users
- **Coupon list**: price, discount, genre name, locations
  - train: $\approx$ 18 000 coupons
  - test: $\approx$ 400 coupons

Predict: user-coupon purchases for the 53rd week

Evaluation

Mean Average Precision @ 10 (MAP@10):

$$MAP@10 = \frac{1}{|U|} \sum_{u=1}^{|U|} \frac{1}{\min(m, 10)} \sum_{k=1}^{\min(n,10)} P(k)$$

where

- $|U|$ is a number of users
- m is the number of purchased coupons for the given user
- n is the number of predicted coupons
- $P(k)$ is the precision at cutoff k:

$$P(k) = \frac{N \text{ correct coupons from the first } k \text{ predicted}}{k},$$

if k-th coupon is predicted correctly, 0 - otherwise.

Cross validation

Problem: possible pairs user-coupon in train: $\approx 360\,000\,000$

- To decrease the size of train, for each week:
  - take coupons with at least one purchase
  - take users with at least one purchase
  - it gives $\approx 600\,000$ pairs for each week, or $\approx 30\,000\,000$ pairs for the whole train
  - use last few weeks to predict test week (we used last 5 weeks)
- Validation set: pairs for the last week
- To match CV and LB score: use multiplier (MAP@10 = 0 for users without purchases)

Feature engineering

- **Dummies:** one-hot encoding
- **Counts:** number of samples for different feature values
- **Counts unique:** number of different values of one feature for fixed value of another feature
- **Likelihoods**
- **Similarities**

Likelihood features

- **Type 1: using sliding window by weeks**

  Example:
  - for each week calculate the rate of purchases by each GENRE NAME based on the previous 10 weeks
- **Type 2: using multi class algorithms**

  Example:
  - one purchase — sample with target GENRE NAME
  - for the test week: predict what will be next GENRE NAME
  - $\Rightarrow$ XGBOOST with 13 classes

Similarities

- the idea: for each user find the similarity between test coupons and coupons purchased before
- use cosine distance with weights
- coordinates for each coupon — coupon features

Final model and resutls

- XGBOOST with **rank:map** objective and **map@10** evaluation metric

Place	**Team**	**Leaderboard score**
1	Herra Huu	0.009973
2	Halla Yang	0.009848
3	threecourse	0.009484
...	...	...
20	**Dmitry and Leustagos**	**0.007642**

Some observations

- Did not help:
  - ensembling of different models
  - models that optimized different loss functions (not MAP@10)
  - reduce dataset
- We did not try but we should:
  - Bayesian approach: work more accurate with likelihoods
  - weeks are too different: choose weeks similar to the test week and train on those weeks only

Thank you! Questions???

Dmitry Efimov
diefimov@gmail.com