数领先机·智赢未来

第九届中国数据库技术大会

DATABASE TECHNOLOGY CONFERENCE CHINA 2018

A Machine Learning-based Discount Marketing

杜睿桓

京东-广告数据部

The Problem

We want to find the customers who

- have some probability to buy a product, and
- will increase the probability to buy if they receive a coupon

So we are going to

- predict the probability of purchase for a group of customers and products; and
- evaluate for each customer the probability of purchase that he increases when he receives a coupon

The Problem

Problem definition

- Select a group of users and products;
- for each pair of user and product:
  - **predict** the probability that the user orders for the product in a future time window; and
  - **evaluate** the difference of ordering probability of whether or not the user receives a coupon of the product

Outline

1. Feature Engineering
2. Algorithm Selection and Combination
3. Potential Customers Mining
4. Spark-based Algorithm Platform Tuning

Outline

1. Feature Engineering
2. Algorithm Selection and Combination
3. Potential Customers Mining
4. Spark-based Algorithm Platform Tuning

1. Feature Engineering

Try to make it independent with the model and convenient for the model.

1.1 Feature Selection for Big Data

1.2 Feature Preprocessing

1.3 Decreasing the Data Size

1.1 Feature Selection for Big Data

(user, product) All user data is anonymized and encrypted

Static features

- User: static user preferences and profiles, etc.
- Product: color, category, brand, gender, etc.

Dynamic features

- User: average score, pv, adding cart and transactions over a period of time, etc.
- Product: average score, pv, uv, adding cart and transactions over a period of time, etc.
- (User, Product): actions of the user on the same or similar products of the product

1.2 Feature Preprocessing

One-hot Encoding with spark

id	feature
1	a
2	b
3	a

StringIndexer →

id	feature	Indexed
1	a	1.0
2	b	2.0
3	a	1.0

OneHotEncoder ↙

id	feature	Indexed	onehot	
			a (1.0)	b(2.0)
1	a	1.0	1	0
2	b	2.0	0	1
3	a	1.0	1	0

1.2 Feature Preprocessing

Standardization with spark: train a StandardScaler and then apply it on the training data and test data.

id	feature
1	-1
2	1
3	1

$$x' = \frac{x - \bar{x}}{\sigma}$$

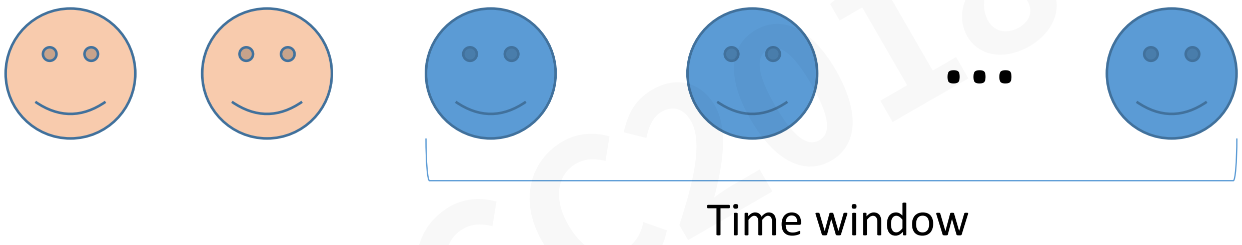
StandardScaler

id	feature	standardized
1	-1	-0.654653670707977**2**
2	1	0.654653670707977**2**
3	2	1.3093073414159544

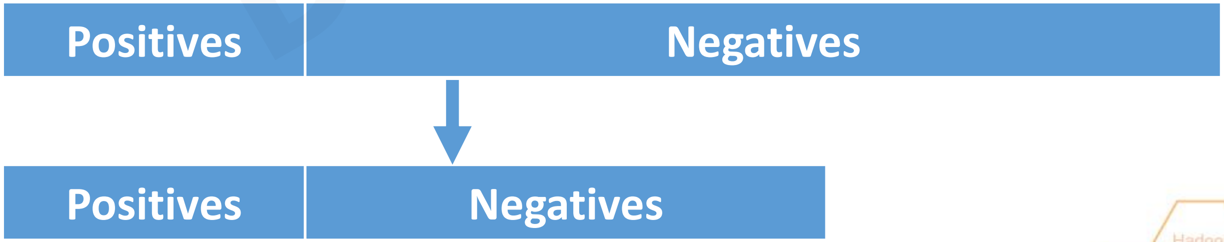
1.3 Decreasing the Data Size

Two methods are adopted to decrease the data size.

- User selection.



- Balance the samples.



Outline

1. Feature Engineering
2. Algorithm Selection and Combination
3. Potential Customers Mining
4. Spark-based Algorithm Platform Tuning

2. Algorithm Selection and Combination

Several algorithms have been experimented on the data. Three of them have pretty good performance

2.1 Logistic Regression (LR)

2.2 Gradient-boosted Tree Regression (GBDT)

2.3 The combination of LR and GBDT

Evaluation of the influence of coupon will also be covered in this chapter.

2.1 Logistic Regression (LR)

$$P(y\,|x) = sigmod(\beta x) = \frac{1}{1+e^{-\beta x}}$$

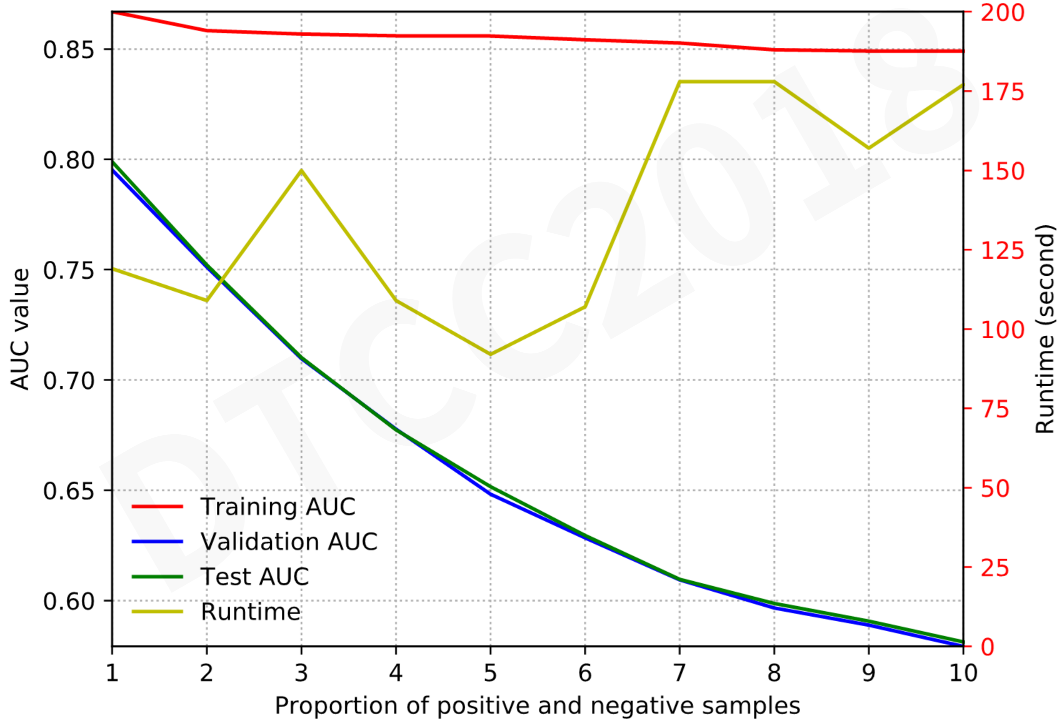


Fig 1. Varying the proportion of the positive and negative samples

2.2 Gradient-boosted Tree Regression (GBDT)

The main idea of GBDT:

- use the forward stagewise method to fit a list of trees iteratively as the weak classifiers and adding them to a strong classifier;
- for each stage, the tree is fitted to pseudo-residuals and the multiplier is computed by the updated loss function.

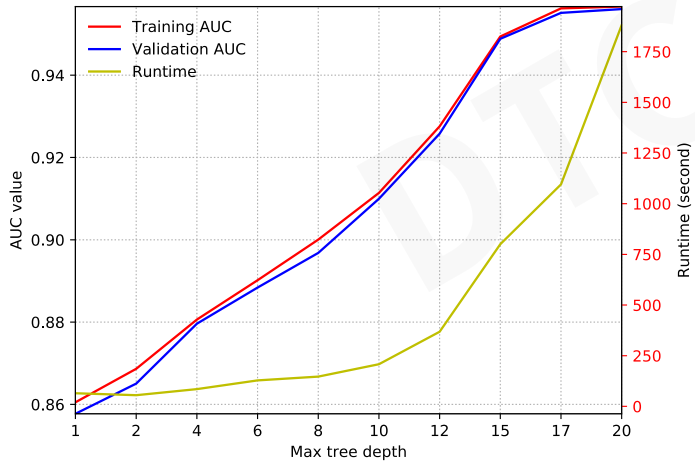


Fig 2. Varying max tree depth

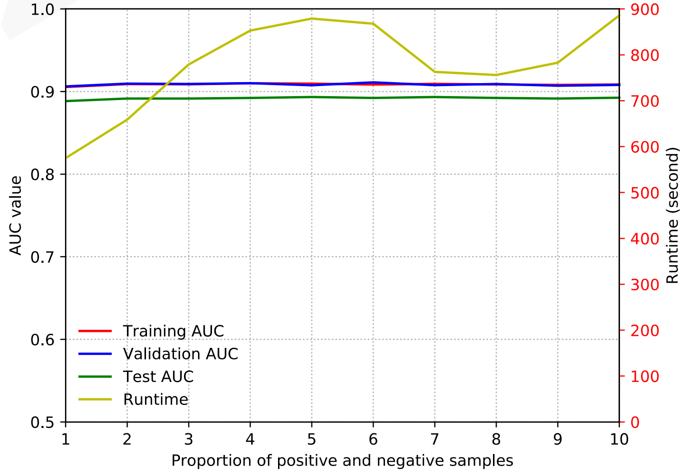


Fig 3. Varying the proportion of the positves and negatives

2.3 The combination of LR and GBDT

A hybrid model is built according to He et al. [1].

- Train a GBDT model;
- Treat each individual tree as a category feature;
- Take as value the index of leaf an instance ends up falling in;
- Use the transformed feature as the feature to train a LR model.

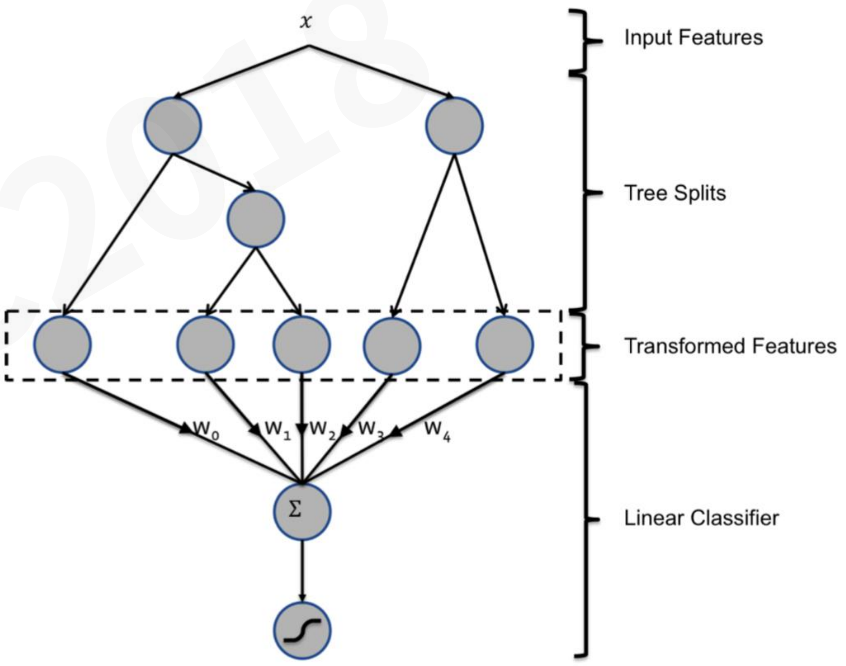


Fig 4. GBDT + LR

2.3 The combination of LR and GBDT

A comparison of LR, GBDT and GBDT + LR.

Table 1. AUC of the three models on test data

	LR	**GBDT**	**GBDT+LR**
Test AUC	0.79891	0.88848	0.90152

2.4 Evaluation of the influence of coupon

There are two methods to evaluate the influence of a coupon to the customer behavior. For a certain user, we denote by

- A: the user orders for some product
- B: the user has a coupon when he is purchasing

The influence will be always positive if we evaluate it by

- $P(A|B) - P(A|\bar{B})$ (both are predicted by the model) or
- $P(B|A)$, where the model is used to predict $P(A)$ (easier to be implemented)

Outline

1. Feature Engineering
2. Algorithm Selection and Combination
3. Potential Customers Mining
4. Spark-based Algorithm Platform Tuning

3. Potential Customers Mining

The customers covered by pv-rule are not enough for us.

Two methods are helpful

3.1 User-similarity based extension

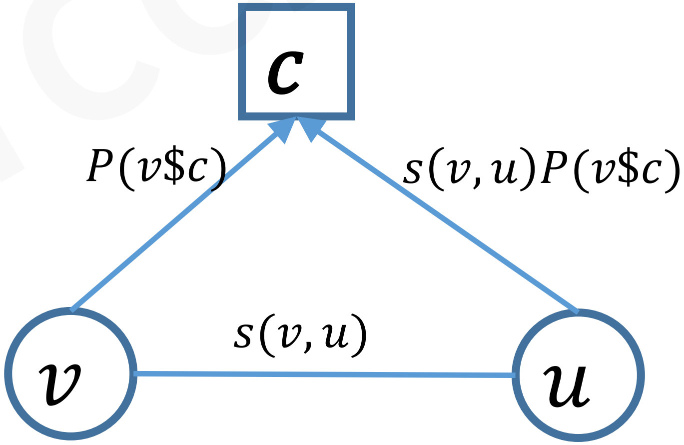
3.2 Product-similarity based extension

3.1 User-similarity based extension

All user data is anonymized and encrypted

Similar users have common interests. To measure the similarity of users, the following aspects are taken into consideration:

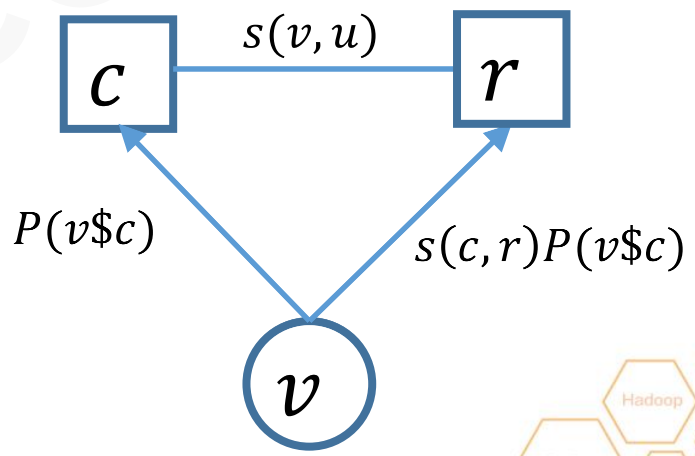
- recently browse and purchase
- static user profiles, etc.



3.2 Product-similarity based extension

Similar products have similar customers. To measure the similarity of products, the following elements are taken into consideration:

- recently browse and sale
- price, weight, gender, etc.
- color, brand, etc.



Comparison

The product similarity based version is better.

Table 2. Increase of recall after extension

	User similarity based	Product similarity based
$\Delta recall$	0.0004	0.0050

Outline

1. Feature Engineering
2. Algorithm Selection and Combination
3. Potential Customers Mining
4. Spark-based Algorithm Platform Tuning

4. Spark-based Algorithm Platform Tuning

4.1 ML vs. Mllib

4.2 Cross-Validation and Parameter Tuning

4.3 Tuning Spark

4.1 ML vs. Mllib

ML

- New
- Pipelines
- Dataframe
- Easier to construct a machine learning pipeline
- More reliable (personally speaking)

MLlib

- Old
- RDD's
- More features

4.2 Cross-Validation and Parameter Tuning

CrossValidator

- K-fold cross validation
- Hyper-parameter tuning: it builds a parameter grid and searches for the best combination of parameters
- Expensive

TrainValidationSplit

- Hyper-parameter tuning

4.3 Tuning Spark

Some simple practical experiences

- Allocate as many cores and memory as possible
- Repartition when reading data from hdfs
- Avoid load tilt after shuffles
- Split a heavy task to steps
- Speculation, rpc.askTimeout, etc.

Summary

1. Feature engineering is the most important part

2. Boosting trees method is better than Logistic method

3. Product similarity based extension is better

4. ML is more advanced than MLLib, but MLLib provides great freedom

References

[1] He, Xinran, et al. "Practical lessons from predicting clicks on ads at facebook." *Proceedings of the Eighth International Workshop on Data Mining for Online Advertising*. ACM, 2014.

[2] Friedman, Jerome, Trevor Hastie, and Robert Tibshirani. *The elements of statistical learning*. Vol. 1. New York: Springer series in statistics, 2001.

[3] 李航. "统计学习方法." *清华大学出版社, 北京* (2012).

Personal Communication and Job Opportunities

Email: duriuhuan@jd.com

THANKS


HANA
Hadoop
SQL
MySQL
DB2
BigDate
Spark
SQL Server
MPP
NoSQL
Oracle
NewSQL
OLAP

讲师申请
联系电话（微信号）：18612470168

关注“ITPUB”更多
技术干货等你来拿~

与百度外卖、京东、魅族等先后合作系列分享活动

微学堂

让学习更简单

微学堂是以ChinaUnix、ITPUB所组建的微信群为载体，定期邀请嘉宾对热点话题、技术难题、新产品发布等进行移动端的在线直播活动。

截至目前，累计举办活动期数60+，参与人次40000+。

ITPUB学院

ITPUB学院是盛拓传媒IT168企业事业部（ITPUB）旗下
企业级在线学习咨询平台
历经18年技术社区平台发展
汇聚5000万技术用户
紧随企业一线IT技术需求
打造全方式技术培训与技术咨询服务
提供包括企业应用方案培训咨询（包括企业内训）
个人实战技能培训（包括认证培训）
在内的全方位IT技术培训咨询服务

ITPUB学院讲师均来自于企业
一些工程师、架构师、技术经理和CTO
大会演讲专家1800+
社区版主和博客专家500+

培训特色

无限次免费播放
随时随地在线观看
碎片化时间集中学习
聚焦知识点详细解读
讲师在线答疑
强大的技术人脉圈

八大课程体系

基础架构设计与建设
大数据平台
应用架构设计与开发
系统运维与数据库
传统企业数字化转型
人工智能
区块链
移动开发与SEO

联系我们

联系人：黄老师
电　话：010-59127187
邮　箱：edu@itpub.net
网　址：edu.itpub.net
培训微信号：18500940168