Task

The Online-To-Offline (O2O) lifestyle services have become a big phenomenon in China. In accordance with INNOVATION IS EVERYWHERE, O2O could be described as the link between 'online discovery' and actual commerce in the physical world. More specifically, as declared in the blog of Martin Pasquier, the business of O2O begins with the customer that go online: with the help of online component, consumers could discover a product and pay for it. Afterwards they go to the physical store to complete the actual purchase. Concerning the types of services, according to a research of Tencent Penguin Intelligence, food delivery, transportation, travel and groceries etc. are the most popular kinds of O2O services. Thanks to the technological innovations and the hunt for convenience, O2O industry rise dramatically and become a must-have for many Chinese residents.

As the main marketing tool of O2O services, coupon is widely used as purchase incentives[1] in customer acquisition and customer retention[2]. However, randomly placed coupons are annoying to most users. Especially when it comes to merchants, useless coupons may take effect on brand image and make it difficult to estimate marketing costs. In this sense, based on the rich data in the O2O scene provided by this competition, participants are required to predict if a coupon would be redeemed by the user in the specific time window.

Original Data

Table 1: Overview of Original Data

File	Name	Description	Time Window
ccf_offline_stage1_train.csv	Train Table Offline	offline purchasing and coupon redeem	01.01.2016 - 30.06.2016
ccf_online_stage1_train.csv	Train Table Online	online click/receive/purchase and coupon redeem	01.01.2016 - 30.06.2016
ccf_offline_stage1_test_revised.csv	Test Table Offline	offline purchasing and coupon redeem prediction	01.07.2016 - 31.07.2016

In accordance with the official document, the data[3] describes online and offline user behaviors between 01.01.2016 to 30.06.2016, with this information participants are to predict the coupon redemption in the following 15 days for coupons that are received in July 2016. As can be seen from the table above, the competition provides two data sets for training and one for prediction.

[1] http://www.data-mining-cup.de/en/review/goto/article/dmc-2015.html
[2] https://tianchi.aliyun.com/competition/introduction.htm?spm=5176.100068.5678.1.9Igo9O&raceId=231587
[3] https://pan.baidu.com/s/1nvFG2ff

Concerning the first data set, as shown below, ‘Train Table Offline’ provides information on the user's offline purchasing behaviors as well as attributes of the underlying deals and coupons.

Table 2: Train Table Offline

Field	Description
User_id	id of users
Merchant_id	id of merchants
Coupon_id	id of coupon: 'null' - a purchase without coupon in this sense, Discount_rate and Date_received are nonsense
Discount_rate	a normal discount rate when x ∈ [0,1] while in the format of x:y - user gets x Yuan rebate as purchase amounts at y Yuan for instance, 100:20 means user gets 20 Yuan reduction if the deal costs more than 100 Yuan
Distance	the distance between the place the user haunts and the most accessible store of the merchant the distance is in the format of x*500 meters where x ∈ [0,10], to note, 0 refers to a distance smaller than 500 m while 10 means more than 5 km otherwise 'null' indicates an absent info;
Date_received	date that the user received the coupon
Date	date when the deal occured if Date=null & Coupon_id != null this is a negative instance because the coupon is received but not used; if Date!=null & Coupon_id != null this is the date when the deal occured and coupon redeemed, positive instance In case Date!=null & Coupon_id = null this is a normal purchase without coupon;

Similarly, the second data set, 'Train Table Online' (see table below) describes the user's online actions of receiving the coupons and underlying offline transaction details.

Table 3: Train Table Online

Field	Description
User_id	id of users
Merchant_id	id of merchants
Action	actions on coupon: 0 click, 1 purchase, 2 receive
Coupon_id	id of coupon: 'null' - a purchase without coupon in this sense, Discount_rate and Date_received are nonsense
Discount_rate	a normal discount rate when x ∈ [0,1] while in the format of x:y - user gets x Yuan rebate as purchase amounts at y Yuan for instance, 100:20 means user gets 20 Yuan reduction if the deal costs more than 100 Yuan
Distance	the distance between the place the user haunts and the most accessible store of the merchant the distance is in the format of x*500 meters where x ∈ [0,10], to note, 0 refers to a distance smaller than 500 m while 10 means more than 5 km otherwise 'null' indicates an absent info;
Date_received	date that the user received the coupon
Date	date when the deal occured if Date=null & Coupon_id != null this is a negative instance because the coupon is received but not used; if Date!=null & Coupon_id != null this is the date when the deal occured and coupon redeemed, positive instance In case Date!=null & Coupon_id = null this is a normal purchase without coupon;

Additionly, the third table is similar to the ‘Train Table Offline’ but lacks of ‘date’ information that tells the condition if the coupon is redeemed.

Table 4: Test Table Offline

Field	Description
User_id	id of users
Merchant_id	id of merchants
Coupon_id	id of coupon: 'null' - a purchase without coupon in this sense, Discount_rate and Date_received are nonsense
Discount_rate	a normal discount rate when x ∈ [0,1] while in the format of x:y - user gets x Yuan rebate as purchase amounts at y Yuan for instance, 100:20 means user gets 20 Yuan reduction if the deal costs more than 100 Yuan
Distance	the distance between the place the user haunts and the most accessible store of the merchant the distance is in the format of x*500 meters where x ∈ [0,10], to note, 0 refers to a distance smaller than 500 m while 10 means more than 5 km otherwise 'null' indicates an absent info;
Date_received	date that the user received the coupon

In addition to the three data, the format of data to submit is shown below, the first four columns are the ones in ‘Test Table Offline’ and the ‘Probability’ is the prediction outputs.

Table 5: Submit Table

Field	Description
User_id	id of users
Merchant_id	id of merchants
Coupon_id	id of coupon: 'null' - a purchase without coupon in this sense, Discount_rate and Date_received are nonsense
Discount_rate	a normal discount rate when x ∈ [0,1] while in the format of x:y - user gets x Yuan rebate as purchase amounts at y Yuan for instance, 100:20 means user gets 20 Yuan reduction if the deal costs more than 100 Yuan
Probability	probability of the coupon is redeemed in the following 15 days

Evaluation

In this task, the prediction accuracy is evaluated by average AUC. More specifically, AUC is calculated for each coupon_id and the final AUC is the mean of all coupon_id that are in need to predict. The reasoning for AUC is that, ROC is insensitive to the change of the positive to negative ratio of the unseen instances. This is in line with the real business world where the data is highly likely to be imbalanced. In this sense AUC is an appropriate evaluation metric.

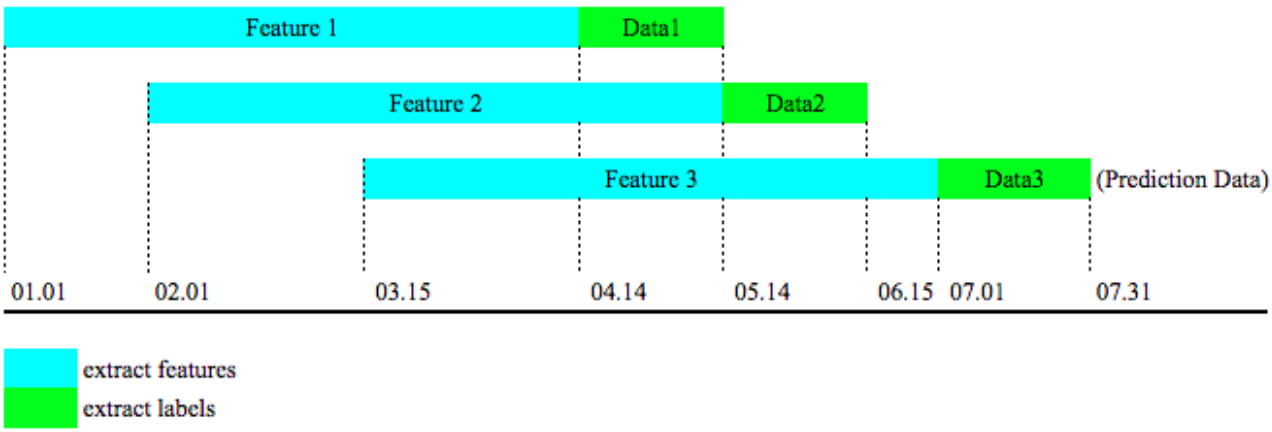
Solution Analysis

The following sections elaborate the winner tactics in details.

Data Splitting

The winner solution implements sliding window method to partition the original data so as to obtain more training data. The video provided by the winner provides more details on this strategy[4].

Figure 1: Data Splitting by Sliding Window



As can be seen in Figure 1, the original Train Table Offline is splitted into five tables. Three data sets (Feature1, Feature2 and Feature3) which are visually presented by the bars in light blue, are used for extracting features. In addition, two sub parts of Train Table Offline are used for extracting labels (Dataset1 and Dataset2), which are presented in green. Important to note, Dataset3 is exactly the Test Table Offline that provided by the original data.

[4] https://tianchi.aliyun.com/video.htm?spm=5176.100258.100258.3.1O7LLR

More specifically, as can be seen from the table below, the three data sets used for extracting features have the same criterias that either the deal is completed (date is not null) or the deal is not completed (date is null) but coupon is received in the underlying time window. The time window differs among the three data sets that Feature1 starts from the beginning of January and ends at the middle of April, Feature2 is one month later than Feature1 and lasts for 3.5 months as well. Similarly, Feature3 contains data from the middle of March until the last day of June.

Table 6: Training Data by Sliding Window

Dataset	Source	Time Window	Shape	Criteria
Feature1	Train Table Offline	01.01.2016 - 13.04.2016	(995k, 7)	transcation is completed (date != 'null'), or no transcation but coupon is received in the given time window
Dataset1	Train Table Offline	14.04.2016 - 14.05.2016	(137k, 7)	Date_received!= 'null' and in given period
Feature2	Train Table Offline	01.02.2016 - 14.05.2016	(813k, 7)	transcation is completed (date!='null'), or no transcation but coupon is received in the given time window
Dataset2	Train Table Offline	15.05.2016 - 15.06.2016	(258k, 7)	Date_received!= 'null' and in given period
Feature3	Train Table Offline	15.03.2016 - 30.06.2016	(1037k, 7)	transcation is completed (date!='null'), or no transcation but coupon is received in the given time window
Dataset3	Test Offline	01.07.2016 - 31.07.2016	(114k, 6)	exactly the Test Table Offline

In addition to the three data sets for feature extraction, two data sets are created for extracting labels. In this sense, these data sets must have records in 'Date_received', so that user receives coupon in these data sets. Dataset1 is exactly the following month of Feature1, which is shown in the Time Window column of the table above. Dataset2 and Dataset3 follows the same rule.

Feature Construction

There are five types of features being extracted: Other Features (Table 7) make use of the leakage information that all features are extracted from the data to predict, which is actually unavailable in real business. Table 8 demonstrates the coupon-related features, such as information of different types of discount. Similarly Merchant Related Features (Table 9) contains derived features that are concerning the attributes of shops, such as the redeemed rate of coupons for each shop. Additionally, in Table 10 there are features that describe attributes of users, for example if the user is fond of using coupons and the maximum distance the user has traveled to for discount. Lastly, features concerning user-shop interactions are recorded.

Table 7: Other Features

Feature	Description
this_month_user_receive_all_coupon_count	count of all coupons the user received in this month
this_month_user_receive_same_coupon_count	count of different coupons the user received in this month
this_month_user_receive_same_coupon_lastone	binary variable if it is the last time the user receives this coupon
this_month_user_receive_same_coupon_firstone	binary variable if it is the first time the user receives this coupon
this_day_user_receive_all_coupon_count	the sum of all coupons the user received on that date
this_day_user_receive_same_coupon_count	sum of all different coupons the user received on that date
day_gap_before (receive the same coupon)	min day gap btw the user received the same coupon and the last time of receiving the coupon
day_gap_after (receive the same coupon)	min day gap btw the user received the same coupon and the next time of receiving the coupon

Table 8: Coupon Related Features

Feature	Description
day_of_week	which day of the week when the coupon is received
day_of_month	retrive day from date when the coupon is received
days_distance	day gap between the transcation date and begining of the period
discount_aomunts_at	amounts_at value of the 'amounts at - rebate' type of discount
discount_rebate	rebate value of the 'amounts at - rebate' type of discount
is_amounts_at_rebate	binary variable if the discount is in the foramt of 'rebate of x if purchase amounts at y'
discount_rate	all in the range of [0,1]

Table 9: Merchant Related Features

Feature	Description
total_sales	aggregrate number of completed transcations by merchant
sales_use_coupon	number of completed goupon-transcations by merchant
total_coupon	num of total coupons sent out by merchat
coupon_rate	sales_use_coupon/total_sales, how much a shop is dependent on coupon in the given period
transfer_rate	sales_use_coupon/total_coupon, the usage rate of coupons that a merchant sent out
merchant_min_distance	the min user-store distance that has transcations conducted via coupon by merchant
merchant_max_distance	the max user-store distance that has transcations conducted via coupon by merchant
merchant_avg_distance	the average user-store distance that has transcations conducted via coupon by merchant
merchant_median_distance	the median user-store distance that has transcations conducted via coupon by merchant

Table 10: User Related Features

Feature	Description
count_merchant	how many shops the user has been to
user_avg_distance	average distance the user has been traveled for purchasing via coupon
user_median_distance	median distance the user has been traveled for purchasing via coupon
user_min_distance	min distance the user has been traveled for purchasing via coupon
user_max_distance	max distance the user has been traveled for purchasing via coupon
buy_use_coupon	count of transcations via coupon per user
buy_total	count of completed transcations per user
coupon_received	count of coupon received per user
user_coupon_transfer_rate	buy_use_coupon/coupon_received, coupon usage rate per user
buy_use_coupon_rate	buy_use_coupon/buy_total, rate of transcations completed via coupon per user
user_date_datereceived_gap	how many days it takes for the user to use the coupon since the coupon is received
avg_user_date_datereceived_gap	average of user_date_datereceived_gap
min_user_date_datereceived_gap	min of user_date_datereceived_gap
max_user_date_datereceived_gap	max of user_date_datereceived_gap

Table 11: User-Merchant Interaction Related Features

Feature	Description
user_merchant_buy_total	count of times the user went to the specific shop
user_merchant_received	count of coupons the user received from the specific shop
user_merchant_buy_use_coupon	count of coupons the user used in the shop
user_merchant_any	count of all kinds of records regarding the user-shop pair
user_merchant_buy_common	count of transcations without coupons in the shop
user_merchant_coupon_transfer_rate	rate of coupons user received from the store are used
user_merchant_coupon_buy_rate	rate of deals via coupon to overall deals the user in the shop
user_merchant_rate	rate of deals to overall records of the user-merchant pair
user_merchant_common_buy_rate	rate of deals without coupon to overall deals the user in the shop