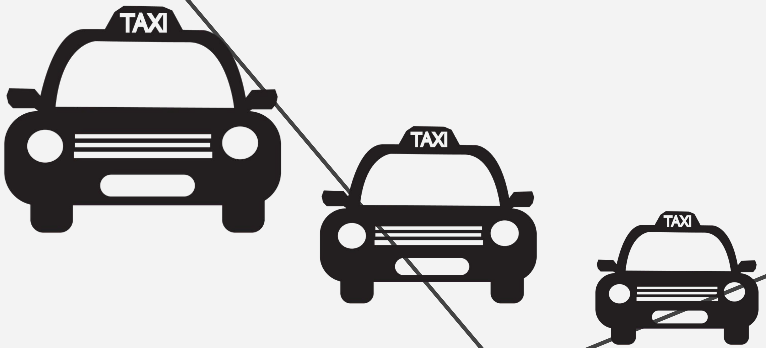
Predicting customer churn for targeting promotions for a traditional taxi company in Vietnam



Team 8
Meng-Hen Huang, Sammi Yien Lu,
Viet-Cuong Trieu (Daniel), Xin Wang

Company Introduction

Founded in 1993

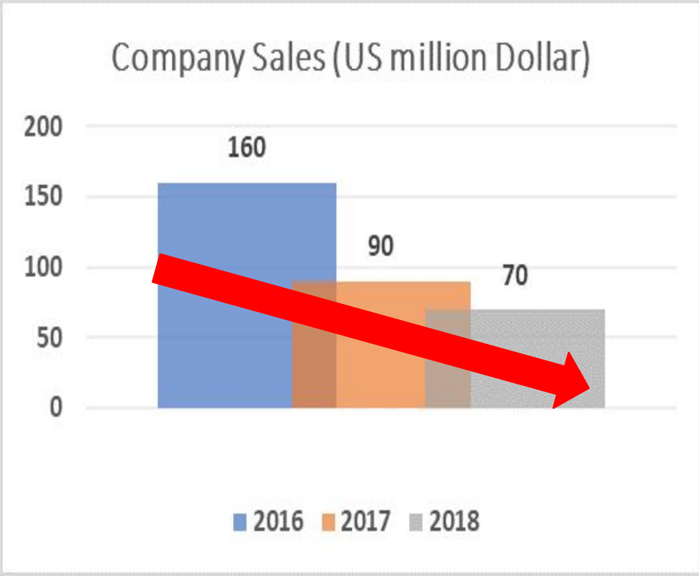
Occupy 63/63 province in Vietnam.

The biggest traditional taxi company in Vietnam

- 15.000 taxi cabs
- 19.000 full time drivers
- 1.5 - 2 million successful trips/month

App booking percentage (7%-10%)

Entered VN market (2015)



Business Problem

Business Goal : To seize/maintain our customer with a very limited budget.

Challenges : Under competition from money-burning promotional campaigns of tech-based taxi services (Uber / Grab)

Opportunity : Precision marketing. If we can predict which customers will leave the service and implement appropriate portion.

Humanity considerations : (1) Data Privacy; (2) Fairness

Stakeholder : (1) Marketing department

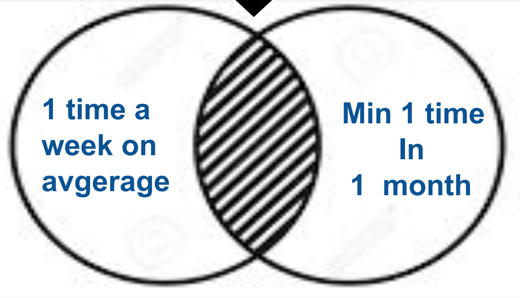
(2) Customer service department

(3) Planning department and board of directors

(4) Users

Data Mining Problem

Regular Customers Definition



Goal — To classify whether regular customers will leave the services or not

Challenge — Lack of customer features

Outcome variables — Regular (leaving) as 1 "Target User : Zero booking in one month"
Regular (loyal) as 0

Task — Supervised & Predictive

Training

Validation

	1st Month	2nd Month	3rd Month	4th Month	5th Month	
Customer A	15	1	10	12		regular (loyal) 0
Customer B √	8	4	1	0		regular (leaving) 1
Customer C	4	0	8	2		irregular

Data Descriptions

Time Period: 7/2019 to 11/2019

Row: Booking transaction data.

Number of used column: 7 / 66

(1) customer id, (2) status

(3-5) time: (request, accept, pickup)

(6) province_id, (7) driver_id

Predictors (initial 18)

- Waiting time (driver accept, pickup)
- Number of trip (success/fail/cancel)
- Number of driver over number of trip
- Day of the week
- Booking in Big city (categorical)

client_id	status	time_client_request	time_driver_accept	time_up_taxi	province_id	driver_id
801550	3	8/19/2019 12:04:17 AM	8/19/2019 12:04:21 AM	8/19/2019 12:06:47 AM	14	16068
216389	3	8/19/2019 12:04:29 AM	8/19/2019 12:04:32 AM	8/19/2019 12:14:29 AM	6	32526
10101	5	8/19/2019 12:07:47 AM			18	
661369	5	8/19/2019 12:08:26 AM			34	
611801	5	8/19/2019 12:10:01 AM			2	
519254	3	8/19/2019 12:10:43 AM	8/19/2019 12:12:00 AM	8/19/2019 12:15:12 AM	2	31098
611801	6	8/19/2019 12:12:24 AM	8/19/2019 12:13:17 AM		2	70459
1016262	3	8/19/2019 12:15:47 AM	8/19/2019 12:15:56 AM	8/19/2019 12:19:40 AM	2	58200

Data Preprocessing

Extract the latest 3 month data → Aggregate and build the possible predictors → Choose regular customers

No.	Transaction data	Predictors construct	Measures in 3 month
1-3	time request- time accept	Driver accept waiting time	max, min, average
4-6	time accept - time pickup	Pick up waiting time	max, min, average
7-9	customer_id, status	Number of trip	successful, fail, user cancel
10-16	time request, number of trip	Day of week booking ratio (7 days)	trip percentage of day of week.
17	driver_id, trip_id	Driver - customer relationship	number of driver over number of trip
18	province id	Big city booking	Customer live in Big city

Methods

<u>Choose Predictors</u>

- Random forest
- Stepwise

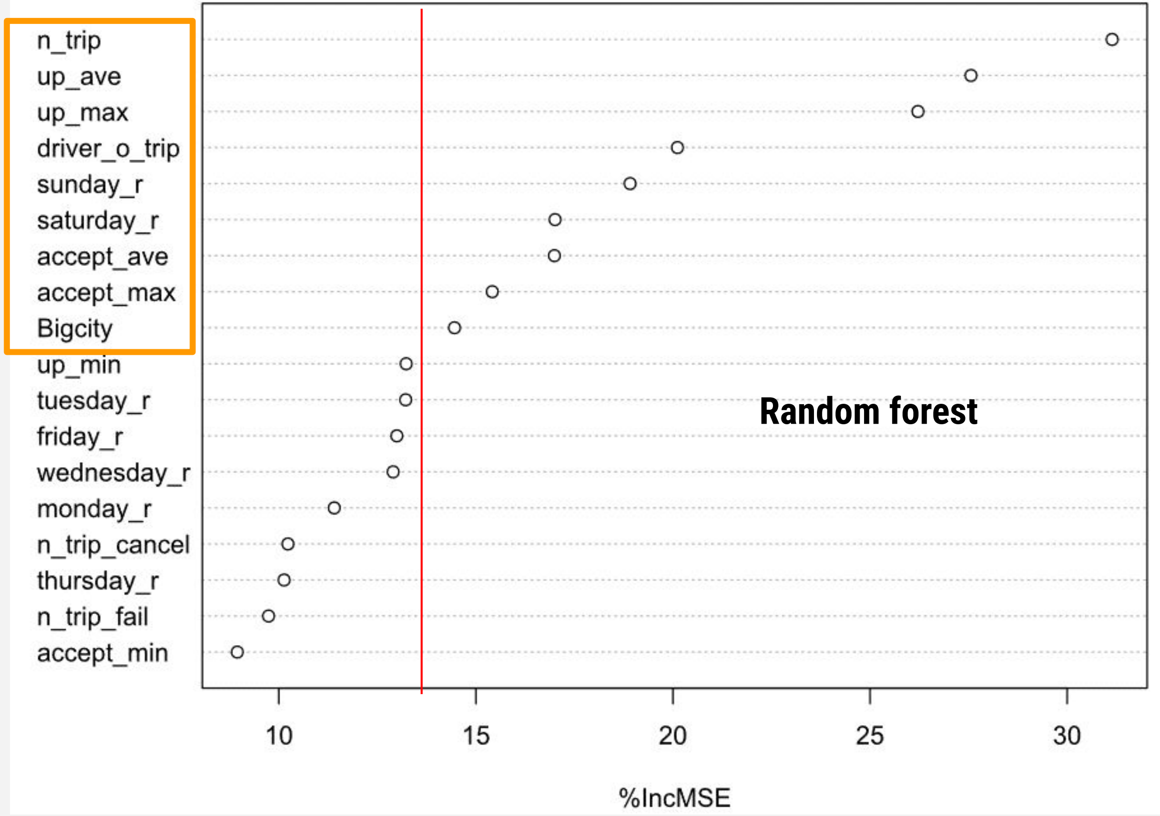
<u>Predict method</u>

Logistic regression

<u>Performance Measure</u>

Decile-wise lift chart

Stepwise: n_trip, driver_o_trip, Saturday_r, accept_max, Bigcity



Run logistic regression

```
Call:
glm(formula = leave ~ n_trip + up_ave + up_max + sunday_r + driver_o_trip +
    saturday_r + accept_ave + accept_max + Bigcity, family = "binomial",
    data = data_taxi)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.7557  -0.5358  -0.4598  -0.3307   3.2512

Coefficients:
               Estimate Std. Error z value Pr(>|z|)
(Intercept)   -1.228782   0.330614  -3.717 0.000202 ***
n_trip        -0.027229   0.004101  -6.639 3.15e-11 ***
up_ave         0.017919   0.032060   0.559 0.576221
up_max        -0.003588   0.003020  -1.188 0.234876
sunday_r      -0.267324   0.407997  -0.655 0.512332
driver_o_trip  0.486232   0.381303   1.275 0.202244
saturday_r    -0.862368   0.434098  -1.987 0.046969 *
accept_ave     0.476231   0.514625   0.925 0.354761
accept_max    -0.299300   0.166497  -1.798 0.072236 .
Bigcity1      -0.384708   0.126602  -3.039 0.002376 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 3356.0  on 4927  degrees of freedom
Residual deviance: 3228.8  on 4918  degrees of freedom
AIC: 3248.8
```

Waiting time
(+)

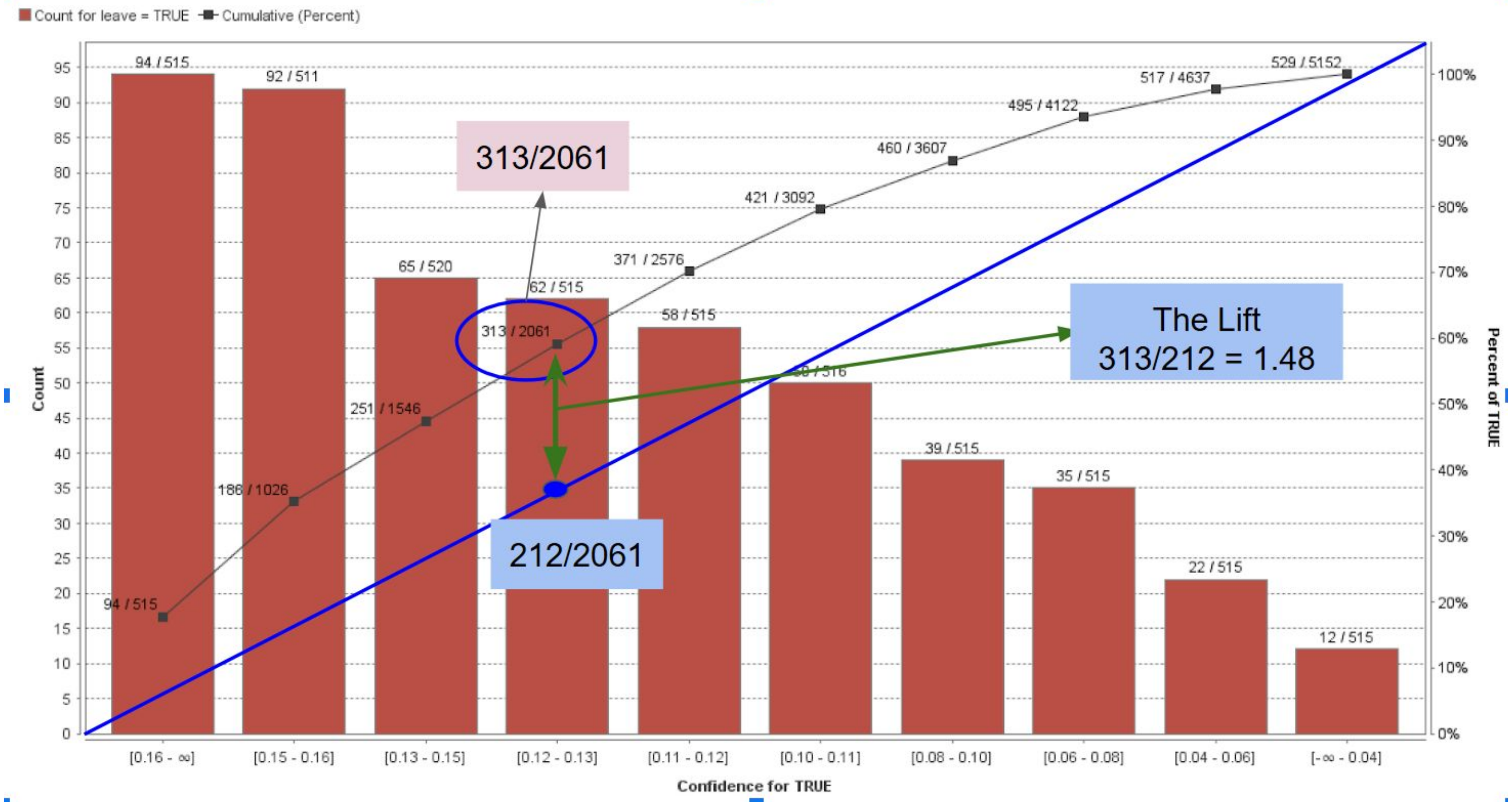
#Driver / #trip
unfamiliar degree
(+)

Weekend
(-)

#Success trip
(-)

Bigcity booking
(-)

Evaluation



Count for leave = TRUE
Cumulative (Percent)
313/2061
The Lift
313/212 = 1.48
212/2061
94 / 515
92 / 511
65 / 520
62 / 515
58 / 515
39 / 515
35 / 515
22 / 515
12 / 515
186 / 1026
251 / 1546
313 / 2061
371 / 2576
421 / 3092
460 / 3607
495 / 4122
517 / 4637
529 / 5152
Count
Percent of TRUE
Confidence for TRUE
[0.16 - ∞]
[0.15 - 0.16]
[0.13 - 0.15]
[0.12 - 0.13]
[0.11 - 0.12]
[0.10 - 0.11]
[0.08 - 0.10]
[0.06 - 0.08]
[0.04 - 0.06]
[-∞ - 0.04]

Limitations & Recommendations

<u>Limitation</u>

1. Not yet exploit location data for prediction
2. Not yet exploit booking data in hourly (e.g., a rush hour)

<u>Recommendations</u>

1. Improve driver - customer relationship
2. Combine with customer support data to get higher prediction performance
3. Predict at the end of month and prepare the promotion