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

Executive Summary

Mai Linh corporation was founded in 1993. Before 2015, the time Uber entered the Vietnamese market, Mai Linh was the largest taxi company in Vietnam, accounting for over 50% of the market share and was the only company operating in 63/63 provinces. Since 2017, due to the competition of tech-based taxi services (Uber, Grab), Mai Linh's revenue has decreased significantly. The revenue for the year 2018 is about 70 million USD, equivalent to 43% of 2016 revenue. From around the middle of 2017, Mai Linh started to apply the taxi dispatch management system to optimize resources so that the taxi fare has also decreased to be equivalent to Uber/Grab. Currently, Mai Linh still faces fierce competition from money-burning promotional campaigns of tech-based taxi service, thus maintaining market share is a vital task.

Currently, Mai Linh serves about 1.5 to 2 million successful trips per month, of which the number of trips booked from the App accounts for only about 7% -10%. Although the marketing department was trying to attract more App booking customers, about 10% of regular customers leave the service every month. Therefore, the business goal is to implement precision marketing to target customers to retain customers with a limited marketing budget. To accomplish this business goal, we perform data mining to predict which customers will leave the service next month based on the lastest three months of transaction data.

After exploring the data and based on the domain expert, we selected seven columns in the booking request table, which would be derived into 18 predictors. To select the prediction method, we randomly selected 60% of the regular customer as training data, and the remaining 40% of customers were holdout samples for evaluation. After comparing and evaluating the predictive performance of different prediction methods based on the ROC curve and Lift curve, we chose the logistic regression method with seven predictors. Finally, to evaluate the business performance of the forecasting model, we use 3-month data (8-10/2019) to predict the customers will leave the service in November 2019. The results show that if the company uses the prediction model to select the top 10% of customers with the highest probability of leaving the service, the company will reach customers who actually leave the service **1.8 times** (the lift) better than the random selection method.

In terms of implementation, Predictive models can be set to run automatically after the end of the month. Predicting results can be automatically sent to the Marketing department to support promotion planning for the next month. Also, the accepted time of the driver and the driver- customer familiarity are two predictors that can be improved. We also make recommendations so the company can improve the driver accepted time and the familiarity between drivers and customers, thereby contributing to regular customer retention.

Although we have made our best effort, this project has some limitations. First, we have not used location data to classify booking requests according to population density. Next, we have not yet categorized wait times according to different time frames (e.g., rush hours). In the future, the prediction model can be improved by exploiting location data, time data by hour and combined with data from the customer care center.

1. Problem Description

❖ Business problem

Business Goal: Use the limited promotional budget effectively to carry out the promotion campaign to the target customers, those who intend to leave the service, to retain the customer.

Stakeholder: the main stakeholder is the Marketing department who runs a promotion campaign to the target customer. Because only a group of regular customers receive promotions, complaints may likely arise, so the Customer service department is also the stakeholder.

Opportunities: it is very difficult and expensive to acquire a new customer than it is to retain a current paying customer. If we can predict which customers will leave the service and implement the appropriate promotion, it will bring a great benefit to the business.

Benefit: Retain the customer so that the company can maintain the market share and contribute to revenue and profit. Besides, from the prediction model, we may the factors that influence the customer's decision to leave and thus improve the service.

Humanity considerations: (1) Data Privacy: customer phone and trip information are exposed that may harm customers themselves and others. **(2) Fairness:** Inequity in distributing promotions to customers: only a small portion of customers receive promotions.

❖ Data mining Problem

The Data mining goal is predicting which regular customers will leave the taxi service in the next month. Here, based on discussions with the company, we define **regular customers** as customers who book at least 1 time per month and 1 time per week in average (equivalent to a number of bookings greater than 12 times in 3 months) in 3 consecutive months. The main outcome variable is a binary variable (Leave/not Leave) in which "Leave" means customers do not book any trip within next month. The outcome (Leave/not Leave) is built based on transaction data of one month; then, this label is used for a supervised learning model.

In the taxi service, customers are not required to declare personal information such as gender, age, etc. Therefore, the challenge is the lack of customer features. The prediction model only uses customer booking data as input data. In terms of deployment, predictive models can run on schedule at the end of the month. The result will be sent to the marketing department. The model should be re-analyzed and re-train after 3-6 months.

2. Data description

Usage data is extracted from the "booking request" table of the taxi dispatching system. The booking request table contains all the information related to customer bookings[1], with about 74 million records from April 2017. Because the predicting model uses the latest 3-month data to forecast, we extract data from July 2019. Besides, we only extract car booking data through the App. The raw data includes 66 columns, and each record is one booking request from the customer. In the next step, based on discussions with the company, we established a data dictionary for each column and then used the domain expert to select candidate columns for the prediction model (see details in appendix 1**).** Finally, we select the appropriate nine columns for the predicting model. However, due to limited geographic data, the final prediction model did not use location data (longitude, latitude). Therefore, the final predicting model only uses data from 7 columns: (1) customer id, (2) status (success, fail), (3-5) time:

[1] Type of transportation: Car or motorbike booking; type of booking: App, phone, Marketing point, Street

(request, accept, pickup), (6) province_id, and (7) driver_id as shown below. Based on these 7 data columns, we built the output variable and 18 input variables (predictors).

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	

3. Data Preparation

Data Preparation consists of 2 steps: (1) Aggregate data and build the possible predictors, and (2) eliminate irregular customers. In the first step, from the raw data, we built 18 predictors including 6 variables related to waiting time, 3 variables related to the number of trips, 7 variables related to the number of trips proportion of the day of weeks, a predictor which measure the level of unfamiliarity between drivers and customers, and a dummy predictor to indicate whether customers live in two big cities of Vietnam (Hochiminh and Hanoi city). In the next step, we only retain regular customers who meet the booking conditions at least 1 time per month and an average of 1 time per week.

No.	Transaction data	Predictors construct	Measures in 3 months
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 trips	Successful, Fail, User cancel
10-16	time request, #trips	Day of week booking ratio	trip percentage of the day of week.
17	driver_id, trip_id	Driver-customer unfamiliarity	#different driver over #trips
18	province id	Big city booking	Customer lives in big city (yes/no)

4. Data mining solution:

❖ Prediction method

Because our goal is to rank the customers according to the highest probability of leaving the service, we choose logistic regression as the main prediction method. We make four different forecasting models then compare and select the best performing model.

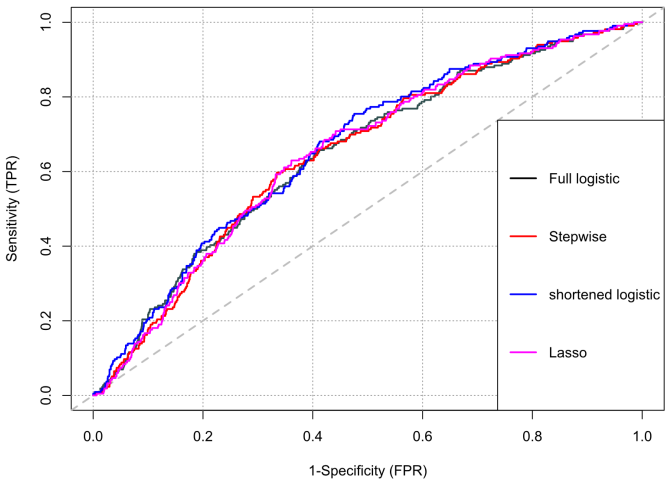
(1) Full logistic regression using all 18 predictors.
(2) Logistic regression with Stepwise using all 18 predictors.
(3) Shortened logistic regression using 7 predictors.
(4) Logistic regression with Lasso using all 18 predictors.

In the shortened logistic regression model, we perform Bootstrapping 10,000 times to determine the variation of the predictor coefficients in the logistic regression equation. Predictors coefficients of which have an 80% confidence interval distribute in both positive and negative sides will be excluded. We also perform collation with the predictor of the stepwise method and the importance of predictors in random forest methods to choose the appropriate predictor. Finally, we only keep seven predictors (table below). Details of this selection process are shown in appendix 2.

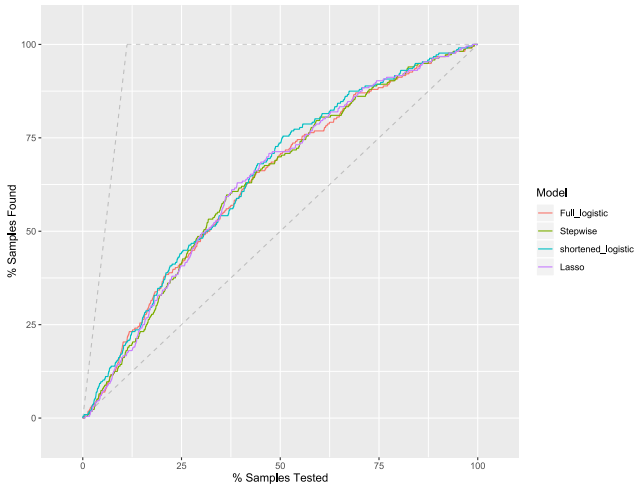
Predictor	80% CI	Predictor	80% CI
Accept time max	[-0.516, -0.078]	#different drivers/#trips	[-0.034, 0.972]
Accept time average	[-0.172, 1.223]	Saturday ratio	[-14.94, -0.525]
Number of trips	[-0.017, -0.034]	Sunday ratio	[-14.36, -0.001]
Big city	[-0.52, -0.19]		

❖ Performance evaluation

To evaluate the performance of the four proposed prediction models, we built predictor data from 3 months (7-9/2019) and outcomes from October 2019. There are 4928 regular customers after data pre-processing. After that, we randomly selected 60% of the customers to use as training data, and the holdout set includes 40% customers. We evaluate the performance of 4 prediction models based on the ROC curve and Lift chart. In general, the performance of the four models is not too different. However, the shortened logistic model is slightly better than the remaining models. This shows that the predictors chosen in the shortened logistic model have a greater influence on the customer decision to leave the service.



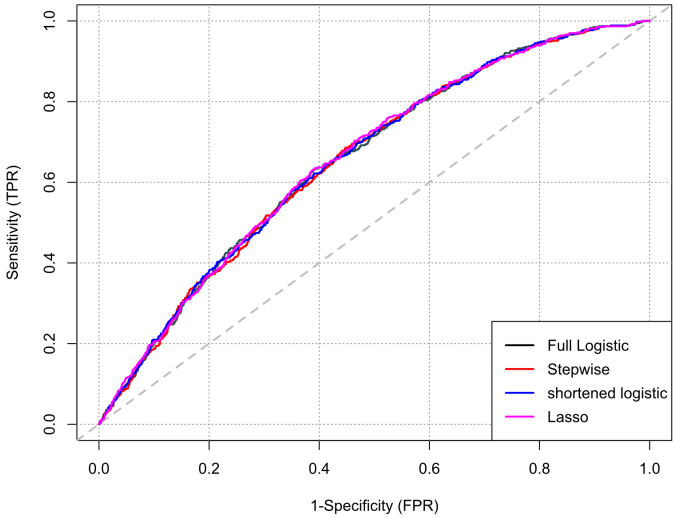
ROC curve of 4 model



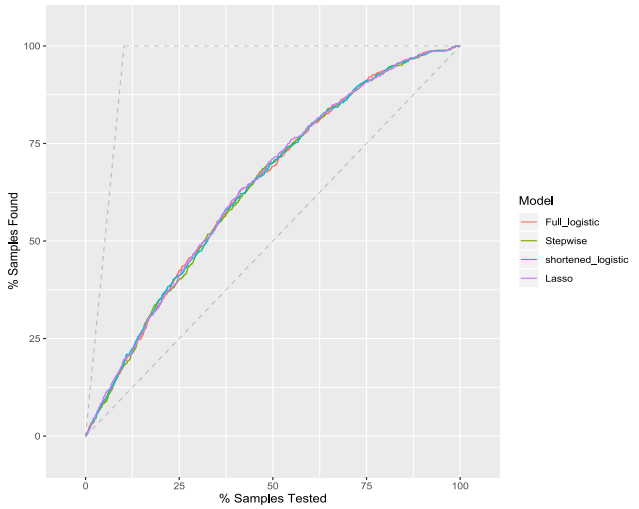
Lift chart of 4 model

❖ Predict for the next month and evaluation

In this section, we use the data of 4 months (7-10/2019) to train prediction models, then use this model to predict customers will leave the service in November based on data of 3 months (8-10/2019). There are 5152 regular customers from these three months of data. The predicted performance of the 4 models was assessed once again through the ROC curve and Lift chart. Overall the performance of the four models is similar. However, we do random resample of the customers, in some cases, the shortened logistic model gives a little more performance. Finally, we decided to choose a shortened logistic regression model as the prediction model.

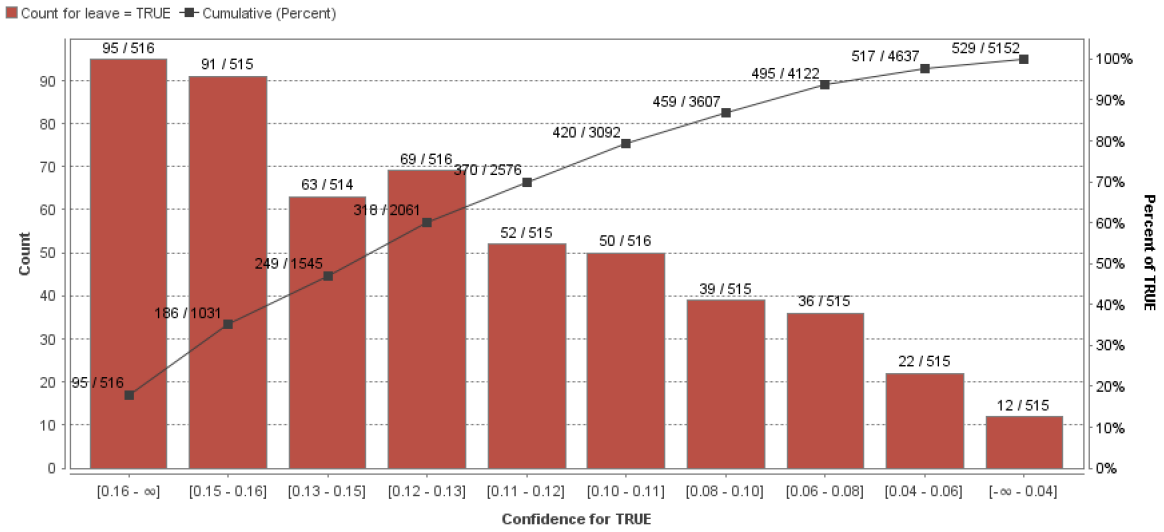


ROC curve of 4 model



Lift chart of 4 model

❖ **Marketing campaign based on prediction result**



The top of	10%	20%	30%	40%
The lift	1.8	1.75	1.56	1.5

The lift of choosing the top 10 to 40% of customers who have a higher probability of leaving shows that the prediction model gives better results than the random selection method. Depending on the marketing budget, using the prediction results, the marketing department can select from 10% to 40% of regular customers to implement the promotion, and it will reach leaving customers from 1.5 to 1.8 times higher than the random selection method.

5. Conclusions

In this project, we have not taken advantage of location data to measure the population density at the booking site as well as not yet categorized wait times according to different time frames (e.g., rush hours). However, our prediction model gave better results than the random selection method, and this result could be used for better reach to leaving customers.

The influence of predictors is different from our initial assumption. For example, we initially assumed that the waiting time pickup will have a great influence on the outcome, but the result showed that this variable is not a significant influence. The coefficients of logistic models show that higher drivers- customer familiarity or higher driver accept time, the higher probability of leaving service.

Out of the seven predictors that greatly influence the decision to leave the service, driver accepted time, and the driver- customer familiarity is two predictors that can be improved. therefore We propose the following recommendations to improve these two factors.

(1) Consider the driver accept time as a KPI to evaluate drivers and have appropriate bonus policies to reduce this time.

(2) Additional features to notify drivers on the App in case the driver serves customers for the second time or more. Besides, it is possible to train drivers to better communicate with customers, increasing customer relationships.

(3) Adding a feature to the Customer App which allows customers to add drivers to their favorites list and allows customers to select drivers within the appropriate distance.

APPENDIX

1. Data Preparation

Although the booking request table data has up to 66 columns, many of which contain duplicate information, many contain no information. In addition, since this table includes trips from phone booking, marketing points or customers waving a taxi on the road, and also including scooter booking service, there is a lot of information that is not related to our prediction model. We have discussed with domain experts from the taxi company and then build the data dictionary that includes the relevant information as below. We then analyzed and finally selected only nine columns that could be used as input data for the forecasting model. However, the 2 columns containing position information (longitude, latitude) are not used later, we only use the information from the 7 columns eventually.

Field	Description
province_id	the province id of the taxi request, Vietnam has 63 province
client_id	id of the customer, id=0 when the customer makes a phone call to the taxi call center to book a taxi
status_id	3: the trip is finished, 4: customer cancel, 5: request timeout, 6: driver cancel
distance	the distance of the trip
money	just available for bike service, the fare taxi service is calculated by taxi meter fare.
agency_id	id of taxi employee who processes the taxi request
company_driver	the company id of the drive (this corporation has a child company that serves in a certain area (province).
phone_client	the phone number of the customer.
phone_driver	the phone of the driver
time_client_resquest	the time customer send the booking request
content	the departure address
destination	the destination address, only available for customer booking from app.
driver_id	id of the driver
taxi_id	id of taxi (1 taxi may have several drivers to drive)
taxi_name	type of car.
time_driver accept	the time when drive accept the trip
time_up_taxi	time when customer get into taxi
tune_out_taxt	the time when customer get out the taxi
latitute_start	location of departure
longtitute_start	location of departure
latitute_end	loctaion of destination
longtitute_end	loctaion of destination
client_type	1: user VIP of taxi company, 2: user using booking App, 3: booking from call center, 0: others
request_from_app	1: request from customer app, 2: request from call center, 3: request from marketing/operation point
location_operate_id	id of marketing/operaton point if "request_from_app" =3

2. Select the predictor in the shortened logistic regression model

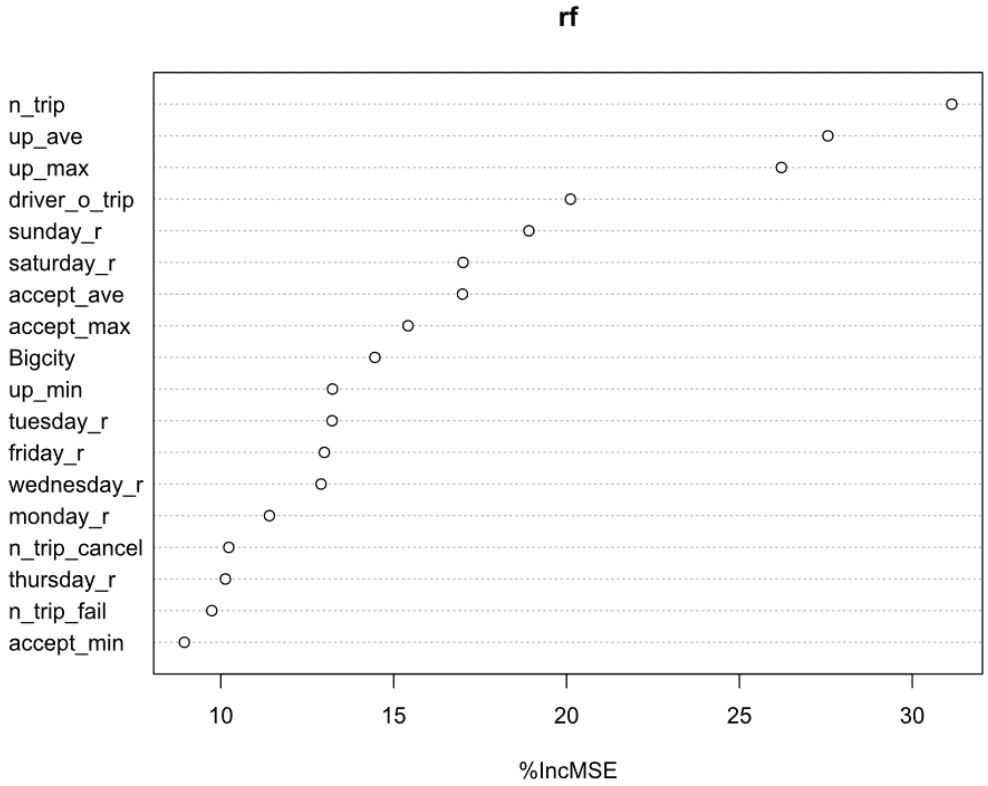
monday_r	tuesday_r	wednesday_r	thursday_r	friday_r	saturday_r	sunday_r	n_trip_cancel	n_trip_fail
-14.05012	-14.0462905	-14.2519010	-13.9971010	-14.1924510	-14.9444121	-14.36038225	-0.001880108	-0.009318397
0.39518	0.4325518	0.1554035	0.3991016	0.2052832	-0.5252968	-0.00974234	0.001072961	0.001485258

accept_max	accept_min	accept_ave	up_max	up_min	up_ave	n_trip	driver_o_trip	Bigcity1
-0.51587924	-1.841443	-0.1728123	-0.0083404290	-0.1023455	-0.02326219	-0.03388473	-0.03421737	-0.5197202
-0.07837795	1.115687	1.2227323	-0.0003118955	0.1345171	0.06360353	-0.01728613	0.97244041	-0.1855246

We performed bootstrapping 10,000 times the coefficients of the logistic equation and calculated the 80% confidence interval(quantile: 0.1-0.9). We eliminate the 10 predictors: *monday_r, tuesday_r, wednesday_r, thurday_r, friday_r, n_trip-cancel, n_trip-fail, accept_min, up_min, up_ave* because the 80% confidence interval distribute in both positive and negative sides. We alo eliminate up_max because the coefficient of this variable is quite small, the effect is not significant and this variable never appears when we use the stepwise method. Besides, we keep the predictor accept_ave because the positive side is larger than the negative side.

Compare with stepwise (5 predictors): n_trip, driver_o_trip, Saturday_r, accept_max, Bigcity.

Compare the importance of predictors using random forest methods.



3. The shortened logistic regression model result

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

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.7561  -0.5354  -0.4601  -0.3314   3.2847

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)   -1.220336   0.329962  -3.698 0.000217 ***
n_trip        -0.028398   0.003991  -7.115 1.12e-12 ***
sunday_r      -0.253746   0.408042  -0.622 0.534032
driver_o_trip  0.511268   0.369105   1.385 0.166005
saturday_r    -0.869404   0.432899  -2.008 0.044608 *
accept_ave     0.482487   0.511174   0.944 0.345230
accept_max    -0.305090   0.166191  -1.836 0.066391 .
Bigcity1      -0.370186   0.126170  -2.934 0.003346 **
---
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: 3230.9  on 4920  degrees of freedom
AIC: 3246.9

Number of Fisher Scoring iterations: 6
```

Waiting time (+)

#Driver / #trip unfamiliar degree (+)

Weekend (-)

#Success trip (-)

Bigcity booking (-)