Blabla Car Market and Pricing Analysis

TABLE OF CONTENT

Executive Summary

Blablacar is a large corporation and it follows the trend of carpooling. However, the pricing strategy for Blablacar is to instrument low price to attract more people and the price for Blablacar is floating. Therefore, although the company is big, the speed of the profit increase is slow. As a result, one task for this report is to establish the pricing model to find the balance point for the pricing(not too high or low) to maximize the profits. Moreover, the report would discuss a different segment of the Blablacar market and provide an objective and practical suggestion for Blablacar to expand their market. To reach these goals, I would adopt different algorithms and methods which including random forest, linear regression, elbow method, k-means, decision tree, Adaboost, gradient boost, SVM, voting method and neural network.

Through the experiment of these algorithms, I establish the pricing model and the mean squared error reaches only 10.9. Moreover, I find two variables that are very interesting and crucial which are the distance of the trip and how many seats in cars. Actually, there are very few drivers who would provide many seats and long-distance service, however, the customers of Blablacar have such a kind of need. As a result, Blabla car can improve their service in this direction and it would satisfy the need for their passenger.

Introduction

Nowadays, more and more people incline to choose carpool as their major transportation on account of environmental issue and cheap price. As a result, it is an crucial time for Blablacar to increase their profit and expand their market. However, according to the information from the Blablacar, we can know that Blablacar did not increase their profit very fast because their strategy is to use the low price to attract customers and the pricing for Blablacar is floating. Therefore, despite of the fact that the company profit increases monotonically but slowly. In this report, I would combine different algorithms and method to help the Blablacar to tackle profit problem.

Research objectives

Pricing is an important factor to define whether the corporation grows up fast or not. Furthermore, if the pricing is over-high, then the company would lose the customer but if the pricing is low then the company would face the low profit problem.

Consequently, pricing should be very careful and the company should spare no efforts to find the appropriate pricing tactics. Besides, which factors should the company improve is another prominent issue. Because maybe there are some factors that when the company take care and improve, then it would cause a strong positive effect on the company.

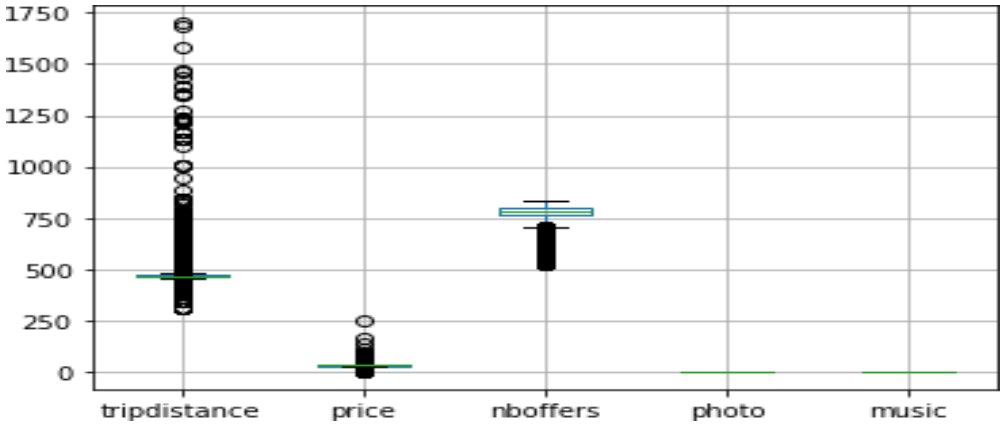
In this report, I would endeavor on these two goals. Firstly, I would establish the pricing model to maximize the profit. After that I would find which factors can be improved to help the company to expand its market.

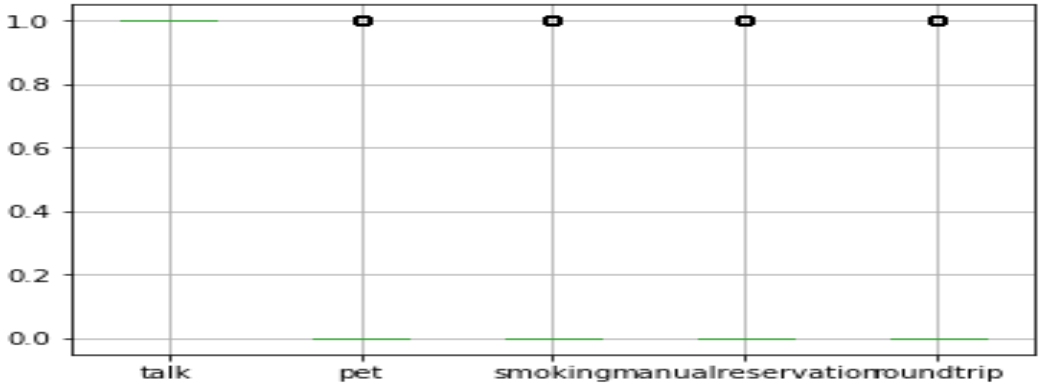
Literature review:

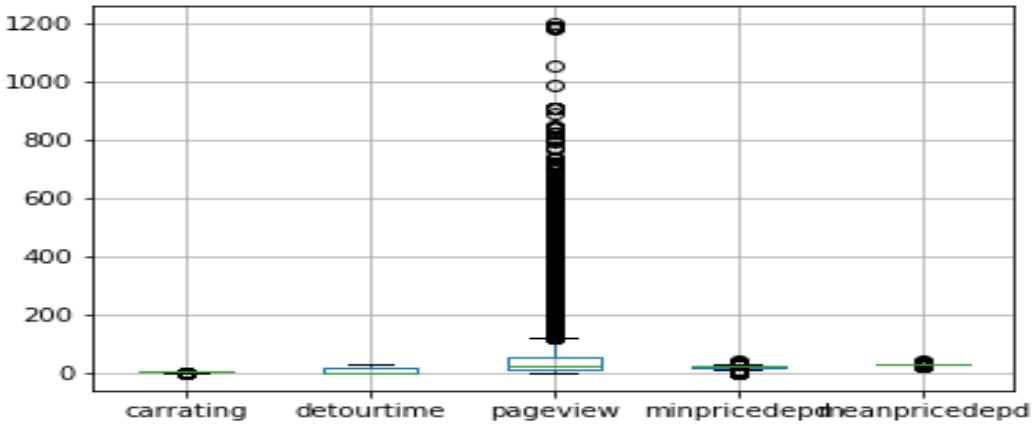
According to the paper(Susan, Shaheen, Adam Stocker, and Marie Mundler (2013). Online and App-Based Carpooling in France: Analyzing Users and Practices—A Study of BlaBlaCar), they find that the main Blablacar driver is highly probable to be high income and passengers is probable to be low income. The finding makes sense because one of the major groups of Blablacar customers are students and most of the students are not rich so Blablacar is a decent alternative for them. Additionally, the traveler and people who move due to work is also the customer for the Blablacar. In the paper (Moshe Ben-Akiva and Terry J. Atherton, (2015), METHODOLOGY FOR SHORT-RANGE TRAVEL DEMAND PREDICTIONS Analysis of Carpooling Incentives) mentioned several reasons for people to take carpooling as their transportation. It includes saving in cost, limit parking space, poor transit and except these normal reasons, also with other interesting reasons. For instance, reduce traffic congestion, fuel consumption and improve air quality. However, carpooling also has difficulties including drivers' will, location and time.

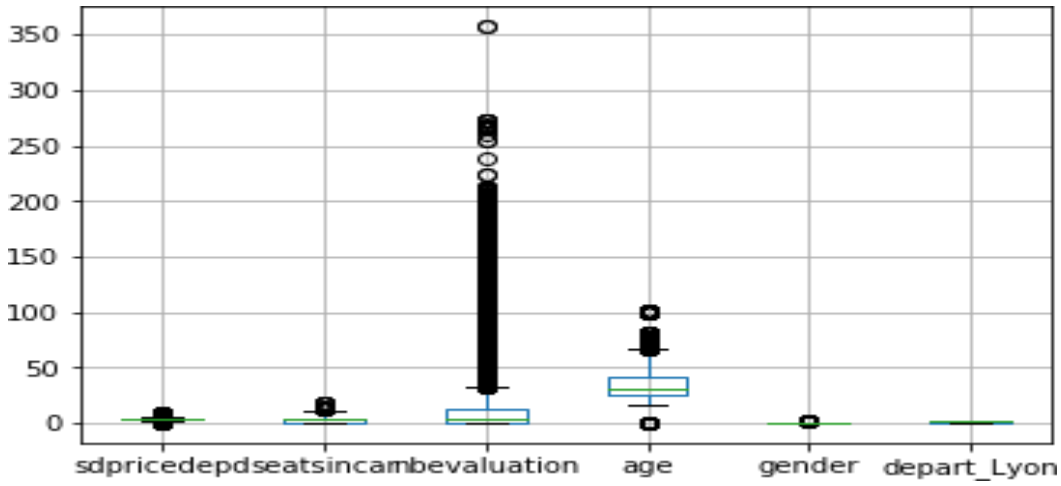
Method

Firstly, I eliminate the strong relationship variables then observe the distribution of variables.


1750
1500
1250
1000
750
500
250
0
tripdistance
price
nboffers
photo
music

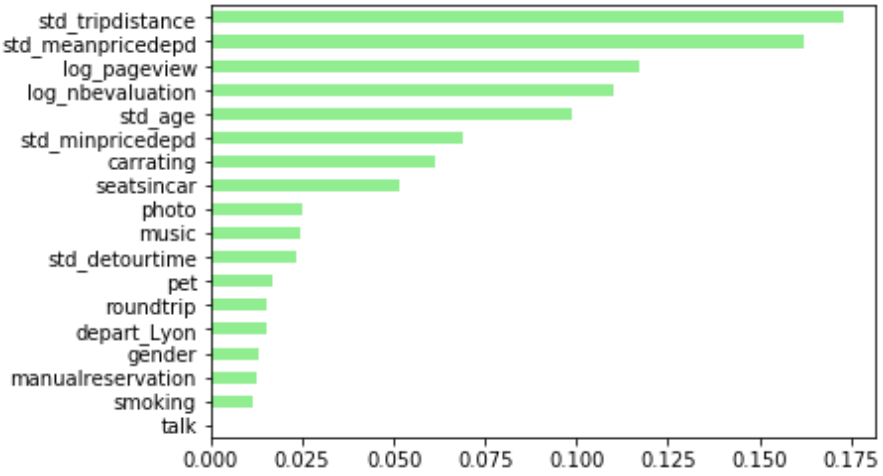

1.0
0.8
0.6
0.4
0.2
0.0
talk
pet
smoking
manualreservation
roundtrip


1200
1000
800
600
400
200
0
carrating
detourtime
pageview
minpricedepd
meanpricedepd



According to these graphs, some of them have strong right skew, for example, "nbevaluation","pageview".In this case, I would use Log to standardize them to change their distribution. Additionally, if the distribution is close to a normal distribution, then I just standardize them to prepare for the algorithm. Also, I eliminate the outliers.

Data description



I would use the top eight variables to conduct further analysis because there is a gap between number eight and night.

Std_tripdistance: This means the distance between departure point and arrival point and this variable is the most important factor for pricing. I standardize the distance to ensure the scale of each variable is similar.

Std_meanpricedepd: This is the mean price for the departure date and I also standardize it.

Log_page view: It means the number of page views for the offer and I use Log to change the distribution in this case.

Log_nbevaluation: It means the number of evaluations the drive has and on account of right skew, I use Log as well.

Std_age: It means the age of the driver and I standardize it.

Std_minpriced: It is the minimum price for the departure date and I standardize it.

Carrating: Each driver can be distinguished to five different levels. 0 means not be revealed, 1means basic, 2 means normal, 3 means comfortable and 4 means luxury.

Seats in car: It means how many seats are available in the car.

RandomForest

In the first stage, I choose to use the Random forest algorithm to distinguish which variable is important to price. There are three important variables in Random Forest, which are min smaples leaf, max feature and tree numbers.

Linear Regression

The first model that I used is linear regression because I want to observe how does this regression defines the relationship between dependent variance and independent variable. In addition, I also want to know whether these parameters are significant or not and how is the R-squared for the model.

K-means Clustering

After the linear regression, I speculate that there are some variables have a prominent and interesting relationship. As a result, I would like to use these variables to separate the market through the elbow method and K-means. Furthermore, after seeing the distribution of the market, I would provide some objective and practical bits of advice for Blablacar to increase their profit and expand their market.

Decision tree + Adaboost

Decision tree and AdaBoost is a non-linear algorithm. There are two important parameters for the decision tree, which is max depth and min sample leaf. However, only one tree is not enough, so I use AdaBoost to improve the model. Adaboost concept is similar to the random forest, however, each tree in AdaBoost would have different weights, which means some trees are important and some trees are not.

Gradient boosting method

Besides the decision tree and AdaBoost, I also want to use another algorithm and combine all of them to find the best result. As a result, I try to use the gradient boosting method here. The notion of gradient boost is that I would find an average of the independent variable and I would also create many residual trees. Each tree would be multiplied by the learning rate.

SVM

The goal of the SVM algorithm is to find a hyperplane to separate the data point correctly and this hyperplane can be a linear, non-linear, and high dimension.

Voting method

This method would combine the algorithm that I mentioned above together and establish the best model.

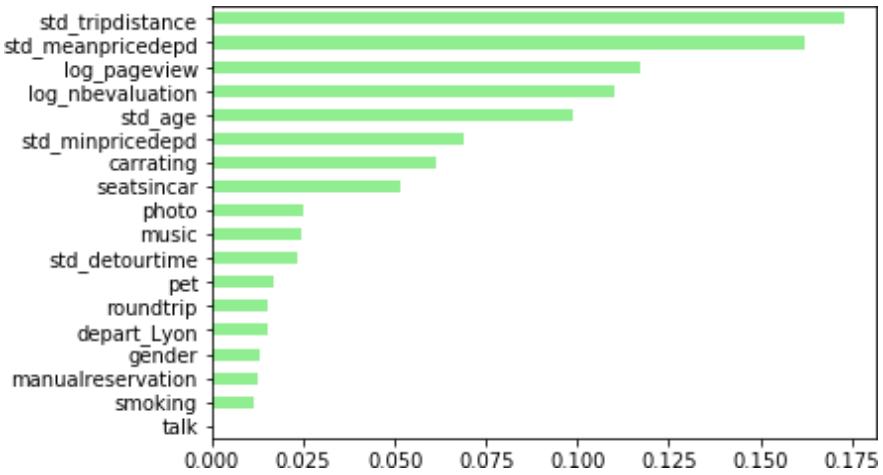
Neural Network

There are too many mediators may exist in the model, and it is hard to find everyone out. Therefore, I establish the neural network, because the neural network would try all the combinations of different variables with different weight.

Result

RandomForest:

After experiment,the best parameter that I get is 250 estimators, sqrt for max features and 4 for min samples leaf. Then I use these parameters to create this graph.



Linear Regression

OLS Regression Results

Dep. Variable:	price	R-squared (uncentered):	0.860
Model:	OLS	Adj. R-squared (uncentered):	0.860
Method:	Least Squares	F-statistic:	5.889e+04
Date:	Sat, 14 Mar 2020	Prob (F-statistic):	0.00
Time:	20:10:27	Log-Likelihood:	-2.9190e+05
No. Observations:	76528	AIC:	5.838e+05
Df Residuals:	76520	BIC:	5.839e+05
Df Model:	8		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
std_tripdistance	0.8162	0.040	20.158	0.000	0.737	0.896
std_meanpricedepd	1.0622	0.042	25.393	0.000	0.980	1.144
log_pageview	-0.1542	0.040	-3.867	0.000	-0.232	-0.076
log_nbevaluation	-2.5291	0.040	-62.924	0.000	-2.608	-2.450
std_age	0.5047	0.040	12.504	0.000	0.426	0.584
std_minpricedepd	-0.0873	0.041	-2.105	0.035	-0.169	-0.006
seatsincar	2.6736	0.020	134.986	0.000	2.635	2.712
carrating	8.0650	0.022	363.751	0.000	8.022	8.108

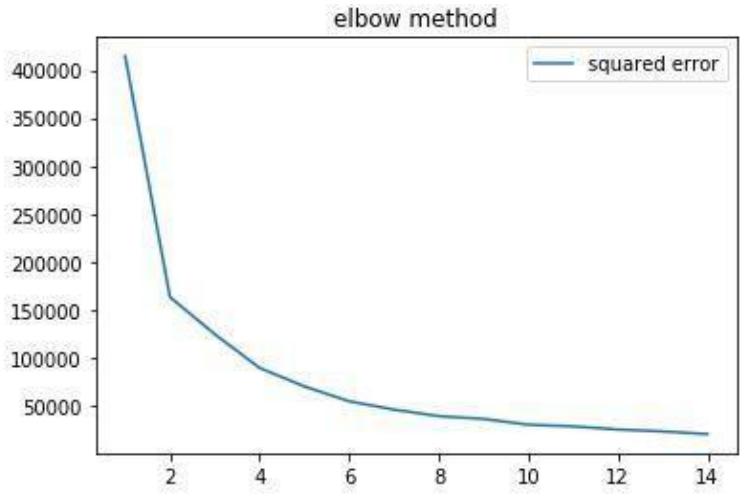
Omnibus:	8782.296	Durbin-Watson:	1.769
Prob(Omnibus):	0.000	Jarque-Bera (JB):	35013.986
Skew:	0.528	Prob(JB):	0.00
Kurtosis:	6.141	Cond. No.	4.40

According to the graph, we can observe that all of the P-value of variables are lower than 0.5, which means that all of them are significant. Moreover, the R-squared is 0.86 which means this model has a strong ability to explain the independent

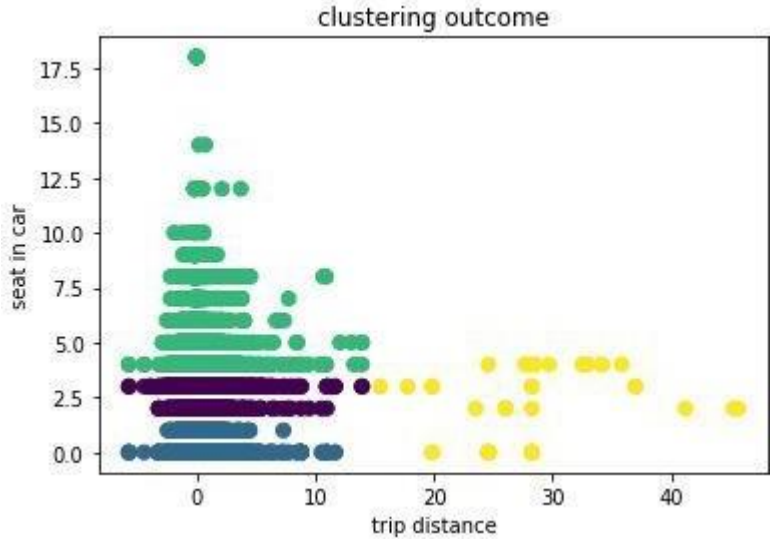
variable. We also can understand the relationship between the dependent variable and the independent variable here. The salient point we can find here is that page view, the number of evaluation and min price has a negative relationship with pricing. Therefore, we can speculate that more evaluation would easier to obtain a more negative score. Also, I calculate the mean squared error for this model which is 117.89, and it did not meet my expectations, so I decided to establish other nonlinear models.

Elbow method and K-means

There are two variables are interesting because we can manipulate it and it also causes a positive effect on the pricing. These two variables are trip distance and seat in a car. Therefore, I used these two variables to conduct the elbow method and clustering.



(Figure1)



(Figure2)

When observing the elbow method (Figure1), I chose to cluster into four groups. When I look at the figure2, I find that the upper right of the figure is empty, which means there are almost no Blablacar driver would take a few passengers or even many passengers for long-distance. In my point of view, it is a good penetration point for Blablacar. Corresponding to the literature review, the main customers are student, traveler, and business traveler and fortunately all of these people have the demand to travel for middle and long distances. Moreover, compare with other variables, seats in car and distance is relatively easy to improve. However, one of the barriers for carpooling is that there are not many drivers who have enough space to take a middle number of people or a large number of people and despite they have space they may not willing to take people when they are driving. As a result, I contemplate that Blablacar can establish some stipulation which would benefit these type of drivers, to attract more drivers for Blablacar in this market segment.

Decision tree and Adaboost

I obtain that 8 is the best max depth and 2 is the best min sample leaf. After that, I combine the decision tree with Adaboost to establish the model. Finally, I obtain the mean square error outcome was 10.6, which is better than linear regression.

Gradient boosting and SVM

According to my experiment, I obtain the mean square error was 10.28. for gradient boosting. For the SVM algorithm, I obtain the mean squared error is 10.91.

Voting method:

According to this method, I obtain 10.09 for the mean square error and it is lower

than previous several algorithms so this is the best model that I got and I think it can help Blablacar for pricing goal.

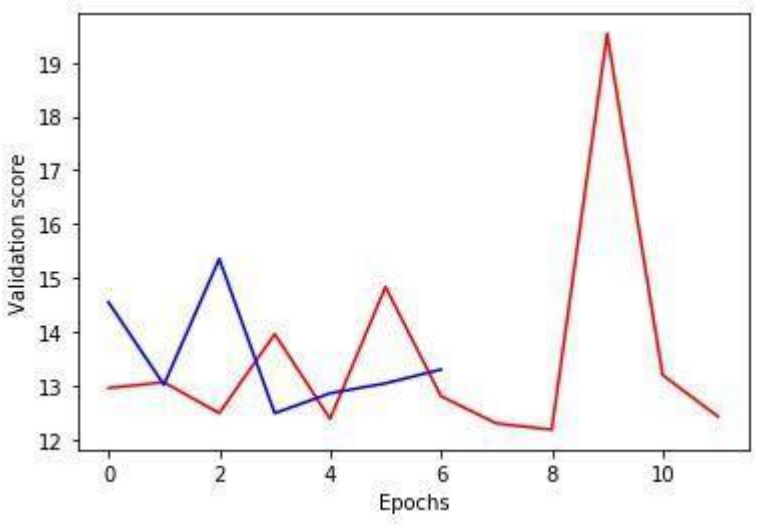
Neural Network

```
Train on 78100 samples, validate on 19525 samples
Epoch 1/15
78100/78100 [==============================] - 6s 79us/step - loss: 16.4501 - accuracy: 0.1446 - val_loss: 12.9519 - val_accuracy: 0.2035
Epoch 2/15
78100/78100 [==============================] - 6s 74us/step - loss: 12.8090 - accuracy: 0.1555 - val_loss: 13.0590 - val_accuracy: 0.1010
Epoch 3/15
78100/78100 [==============================] - 6s 74us/step - loss: 12.4719 - accuracy: 0.1624 - val_loss: 12.4840 - val_accuracy: 0.1552
Epoch 4/15
78100/78100 [==============================] - 6s 75us/step - loss: 12.3710 - accuracy: 0.1644 - val_loss: 13.9551 - val_accuracy: 0.0822
Epoch 5/15
78100/78100 [==============================] - 6s 77us/step - loss: 12.1704 - accuracy: 0.1663 - val_loss: 12.3782 - val_accuracy: 0.1492
Epoch 6/15
78100/78100 [==============================] - 6s 77us/step - loss: 12.0029 - accuracy: 0.1661 - val_loss: 14.8317 - val_accuracy: 0.0715
Epoch 7/15
78100/78100 [==============================] - 6s 76us/step - loss: 11.9014 - accuracy: 0.1696 - val_loss: 12.8015 - val_accuracy: 0.1145
Epoch 8/15
78100/78100 [==============================] - 6s 76us/step - loss: 11.7546 - accuracy: 0.1690 - val_loss: 12.2930 - val_accuracy: 0.1542
Epoch 9/15
78100/78100 [==============================] - 6s 78us/step - loss: 11.7058 - accuracy: 0.1680 - val_loss: 12.1786 - val_accuracy: 0.1394
Epoch 10/15
78100/78100 [==============================] - 6s 76us/step - loss: 11.6205 - accuracy: 0.1703 - val_loss: 19.5397 - val_accuracy: 0.0580
Epoch 11/15
78100/78100 [==============================] - 6s 75us/step - loss: 11.5321 - accuracy: 0.1711 - val_loss: 13.1954 - val_accuracy: 0.2465
Epoch 12/15
78100/78100 [==============================] - 6s 77us/step - loss: 11.4350 - accuracy: 0.1743 - val_loss: 12.4215 - val_accuracy: 0.1153
```

(Figure1)

```
Epoch 1/15
78100/78100 [==============================] - 8s 97us/step - loss: 22.2970 - accuracy: 0.1397 - val_loss: 14.5471 - val_accuracy: 0.0829
Epoch 2/15
78100/78100 [==============================] - 7s 93us/step - loss: 13.1851 - accuracy: 0.1531 - val_loss: 13.0128 - val_accuracy: 0.1235
Epoch 3/15
78100/78100 [==============================] - 7s 93us/step - loss: 12.6735 - accuracy: 0.1621 - val_loss: 15.3511 - val_accuracy: 0.0739
Epoch 4/15
78100/78100 [==============================] - 12s 149us/step - loss: 12.2910 - accuracy: 0.1658 - val_loss: 12.4828 - val_accuracy: 0.1468
Epoch 5/15
78100/78100 [==============================] - 15s 186us/step - loss: 12.1287 - accuracy: 0.1666 - val_loss: 12.8543 - val_accuracy: 0.1056
Epoch 6/15
78100/78100 [==============================] - 8s 102us/step - loss: 11.9649 - accuracy: 0.1656 - val_loss: 13.0386 - val_accuracy: 0.1255
Epoch 7/15
78100/78100 [==============================] - 7s 95us/step - loss: 12.0237 - accuracy: 0.1673 - val_loss: 13.2960 - val_accuracy: 0.2471
```

(Figure2)



(Figure3)

Figure1 is the first neural network and figure2 is the second network. We can use figure three to compare them together. The y-axis for figure3 is the mean squared error for test data. Obviously, the red line is better than the blue line. During 8 epochs, the mean squared error reaches 12.29 which is the lowest and the mean squared error for the training data reaches 11.7 which is an acceptable outcome, so Blablacar can take also this model into account for the pricing tactics.

Conclusion

Market implication

The objective of this report is to find a model to define the price in order to increase their profit and find which variables can Blablacar improves to expand their market.

How do I expand the Blableacar market?

Through linear regression, I find trip distance and seats in the car are two controllable and important variables. Through K-means I find that there are very few cars which have middle and many seats would provide the middle and long-distance trips. However, the customers of Blablacar have demands in this segment. As a result, if Blablacar willing to establish some attractive condition for the driver who meets the requirement that I mentioned above, then they have a high probability to expand their market and increase their profit.

How to pricing?

The high price would lose customers and a low price would lose profit, so the maximum profit would appear in a balance point. I use a decision tree, AdaBoost, gradient boost, SVM to establish the first reliable pricing model with only 10.09 mean squared error. Moreover, I take into account all mediators when conducting a neural network and create a second reliable model for pricing with an 11.7 mean squared error.

Limitation

The first limitation is that I prepared more content than limit to explain the detail of report but I cancel them finally. The second limitation is one pricing model is not enough because there are many segments in the market. As a result, it is better to establish many models correspond to each segment in the market.

Self-reflective Statement

Actually, before this module, I have no idea how to analyze the market. As a result, I also do not know how to think about a business problem, market implication, etc... Through this course and group project, finally, I obtain some basic concepts about this field and understand what is customer lifetime value, how to calculate the market share, etc.. I know I still have a long way to go, but it is a good starting point for me.

References

Susan, Shaheen, Adam Stocker, and Marie Mundler (2013). Online and App-Based Carpooling in France: Analyzing Users and Practices—A Study of BlaBlaCar

Moshe Ben-Akiva and Terry J. Atherton, (2015), METHODOLOGY FOR SHORT-RANGE TRAVEL DEMAND PREDICTIONS Analysis of Carpooling Incentives

ROGER F. TEAL ,(1986), CARPOOLING: WHO, HOW AND WHY

Yuting. SU,(2016), BlaBlaCar-France's only unicorn and originator of the sharing economy(https://www.bnext.com.tw/article/39101/bn-2016-04-02-020242-174)

Chiara Bresciani, Alberto Colorni, Francesca Costa, Alessandro Luè Luca Studer,(2018) Carpooling: facts and new trends(https://www.researchgate.net/publication/328370623_Carpooling_facts_and_new_trends)