International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 06 Issue: 01 | Jan 2019 www.irjet.net p-ISSN: 2395-0072

Targeting for Up-Selling to Increase Revenue for Ride-Hailing Service using Machine Learning

Abhishek Kumar Tiwari[1], Govinda Bobade[2], Pratikkumar Dubey[3]

[1,2,3]*Manager Advance Analytics, TCS*

Abstract - *The rapid expansion of cities and urbanization has created a great load to the city traffic. This has led to large number of people finding comfortable mode of transport options - cab hailing service is one of the preferred options. Since past few years when we talk about Taxi services, the first thing we will think of is mobile applications like Uber or Ola or any other app - no more conventional (laborious, time consuming and risky) taxis! They are now an obvious transportation method in almost all urban areas.*

Now a days so many companies are undergoing a rapid digital transformation and trying to grab as much of market shares with innovative digital products, and taxi business is no more exception - instead, it is now leading and surprising everyone. These companies continuously strive to grow their business beyond exponential growth - which in turn result in a never-ending need for innovation. Since past few years, as we have evolved in Machine Learning and Data Estate - there is huge demand on how to use these techniques and tools to grow business.

This paper focuses on few aspects for cab hailing service industry on how they can leverage Machine learning techniques to target existing customers who are not very active but do possess a good scope of revenue generation. We will also be covering how can we-identify such customers as well as how to convert them from modest customer to ideal customer.

Key Words: Cab Hailing Service, Machine Learning, Revenue increase, opportunity prediction, XGBoost, KNN, Machine Learning, Modest Customer, Ideal Customer.

1. INTRODUCTION

Spending patterns has been analysed and with multiple cut-offs trials, got a finalized cut off 75th percentile of the spending. Use of cut off is to decide the high spending and low spending groups. Customers who spend more than cut off value are called "Ideal Customers" and others are called "Modest Customers". In the Modest Customers group some of these Customer have greater potential for higher spending and we want to tap this opportunity.

1.1 Solution Approach

The solution has 2 parts, one is targeting of the customers who has potential to move from Modest Customer group to Ideal customer group and other is finding the spending opportunity from modest customers.

To target we have tried multiple classification machine learning algorithms and we have mentioned the Random forest and XGBoost methods in this paper. The XGBoost method was giving better results so we zeroed down with XGBoost Approach. XGBoost machine learning algorithm helps us in deciding which customers to target. With the propensity score from XGBoost we got the probability of modest customers to move into Ideal customer group.

To address the spending opportunity, we can get maximum spending potential of a customers either by subtracting the current spend from the maximum spender (across segments) or maximum spender within the segment. For example, if a customer's monthly spending is 700 INR and customer with highest monthly spending is at 17000 INR. Then the maximum monthly spending potential for that customer will be 17000-700 which is 16300 INR. Or if customer with highest monthly spending in same group/segment is 12000 INR. Then the maximum monthly spending potential for that customer will be 12000-700 which is 11300 INR.

The problem with above approach is that it is over simplified, very optimistic and greedy. In case we consider this approach, we might give a false picture of KPIs and we can significantly miss the target. Most of the customers will never reach maximum spending either in overall spending or in specific segment as "No one size fits all", hence this paper proposes a unique solution for realistic spending opportunity.

To find out the similarities between modest Customers in lower spending group and Ideal customers in higher spending group we can treat them in test and train group respectively.

To find the similarities between train and test group we have considered the multiple and important features of the customers. Some of them are number of trips, average cost per trip, city, peak timings, average duration of trips, number of cancelations, mode of payment and others.

We have used nearest neighbour algorithm to find the opportunity spending of modest customers in the Ideal customers group. For any modest customer who's spending is less, then what is the (spending) opportunity that customer can have considering the neighbours in the Ideal customer groups.

In this analysis we have ignored the customers who's monthly spending is less than 500 INR.

1.2 Algorithms – Random Forest

Random forest (aka random decision forest) is an ensemble learning method for classification and regression problems. It is efficient ensemble machine learning algorithm that produces great results and avoids overfitting. Because of its efficient processing capability, it is widely used algorithm for classification as well as regression problems.

It takes multiple bootstrap samples and grows trees to the full extends. There is no pruning on the trees.

In this paper we had used random forest of 350 trees and 10-fold cross validation after hyperparameter tuning. This paper had also used advance tuning methods of Random forest for better performance.

1.3 XGBoost

Xgboost is short for eXtreme Gradient Boosting. Gradient boosting technique is used for regression and classification problems, which gives an output prediction as an ensemble of weak predicted models, mainly decision trees. It creates the model in the stage wise like most of boosting and then generalizes by using differentiable loss function.

XGBoost is a very efficient machine learning algorithm which use gradient boosting internally.

XGBoost's greatest advantage is in accuracy and handing of unbalanced datasets. It is over 10 times faster than classical GBM methods at the cost of multiple hyper parameters tuning. It enables parallel processing capability to process the data. XgBoost supports various objective functions including classification, regression and ranking.

$$obj = \sum_{i=1}^{n} l(y_i, \widehat{y(t)}_i + \sum_{i=1}^{t} \Omega(fi)$$

where L is the training loss function, and Ω is the regularization term.

1.4 KNN

The k-nearest neighbours are a non-parametric method which helps in for classification as well as regression problems. It is called as lazy algorithm as it doesn't learn from any function from training data but memorize the training dataset.

The input of this algorithm is checked against all the training example and depending on K neighbours it will give us the output. If it is used for classification, then we will get the predicted class by majority of the K neighbours. In case of regression problems, we get the average value of K neighbours as the predicted value.

We have finalized the K as 5 in KNN after iterations to get the spending opportunity of modest customers with respect to 5 closest customers in the ideal group. Here we will find out 5 nearest neighbors from the ideal customer segment (A, B, C, D, E as shown in Chart -1) for the modest customer to determine the realistic spending opportunity. Thus, the realistic spending opportunity for customer X = Average spending of A, B, C, D and E

Currently we can get the opportunity as the subtracted value of current spending of customer from the opportunity value from KNN.

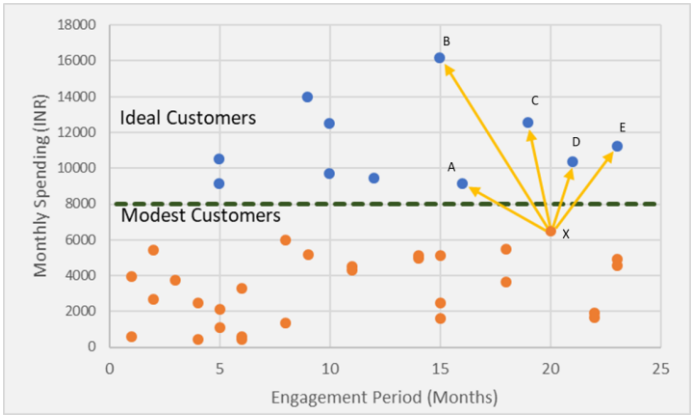

Chart -1: Illustration of KNN

From XGBoost algorithm we got probability of modest customers moving to ideal group customers and only top 20 percent customers by probability score should be targeted by the marketing team. And for those 20 percent people we generated opportunity value using KNN.

1.5 XGBoost Prediction Accuracy Matrices

Confusion Matrix:

		Actual	
		Positive	Negative
Predicted	Positive	True Positive	False Positive
	Negative	False Negative	True Negative

Formulae:

$$\text{Accurecy} = \frac{TP + TN}{TP + TN + FN + FP}$$

$$\text{Sensitivity(Recall)} = \frac{TP}{TP + FN}$$

$$\text{Specificity} = \frac{TN}{TN + FP}$$

$$\text{Precision} = \frac{TP}{TP + FP}$$

$$\text{F1 Score} = \frac{2 * TP}{2 * TP + FP + FN}$$

2. Model Results from XGBoost

In XGBoost model we have divided our dataset in Train, Holdout and Test datasets. Following table shows the model measuring parameters of the XGBoost algorithm.

Train data is 70 percent of the total training data and holdout is remaining part of training data. The test data set is data after a months' time. We had tested the model performance on the test data to ensure we get the consistent results in testing as we got in holdout.

Table -1: Model measuring parameters of XGBoost

Segment	Data Set	Accuracy	Precision	Recall
Go cx	Train	0.873	0.867	0.866
Premier	Train	0.893	0.916	0.926
XL	Train	0.926	0.924	0.815
Hire Go	Train	0.943	0.918	0.875
Hire Premier	Train	0.953	0.926	0.881
Go	Holdout	0.864	0.848	0.865
Premier	Holdout	0.859	0.88	0.865
XL	Holdout	0.898	0.916	0.782
Hire Go	Holdout	0.925	0.88	0.896
Hire Premier	Holdout	0.901	0.858	0.894
Go	Test	0.852	0.835	0.847
Premier	Test	0.855	0.89	0.848
XL	Test	0.909	0.907	0.783
Hire Go	Test	0.917	0.887	0.871
Hire Premier	Test	0.915	0.876	0.829

Segment	Data Set	Negative Precision	Negative Recall	F1 Score
Go	Train	0.897	0.841	0.866
Premier	Train	0.893	0.831	0.921
XL	Train	0.877	0.886	0.866
Hire Go	Train	0.767	0.864	0.896
Hire Premier	Train	0.823	0.79	0.903
Go	Holdout	0.848	0.838	0.856
Premier	Holdout	0.853	0.843	0.873
XL	Holdout	0.855	0.865	0.844
Hire Go	Holdout	0.849	0.859	0.888
Hire Premier	Holdout	0.799	0.754	0.876
Go	Test	0.864	0.756	0.841
Premier	Test	0.865	0.759	0.868
XL	Test	0.867	0.784	0.84
Hire Go	Test	0.881	0.814	0.879
Hire Premier	Test	0.816	0.747	0.852

2.1 Hyper parameter Tuning of XGBOOST

With multiple runs of XGBoost and for different values of ETA following graph was plotted for ETA values. This enabled to get optimal cutoff.

Following Graph variable represents ETA values as 0.05, 0.1, 0.2, 0.5, 1 and value represents the train error.

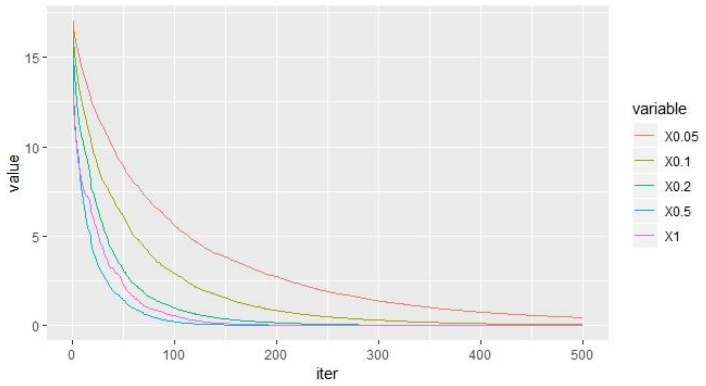

Chart -2: ETA chart

3. CONCLUSION

This paper makes a machine learning based attempt to transform modest customers to ideal customer groups. Accurate prediction of opportunity score can lead to better revenue for the organization. This paper makes an effort toward making the opportunity prediction machine learning based than that of simplified approach. Machine learning methods can help getting more consistent results. As checked in the test dataset. With the help of multiple model measuring parameters to verify the performance such as precision, recall, f1 score, negative precision, positive precision and others, this paper validate the results of test datasets. To compare the Random forest and XGBoost, accuracy measure was used. XGBoost had better accuracy so this paper zeroed down on XGBoost method. The KNN is

efficient as needed neighbor's information and enabled to get opportunity score.

REFERENCES

[1] XGBoost Documentation URL: https://xgboost.readthedocs.io/en/latest/tutorials/model.html

[2] Random Forest Learning sources: URL: https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html

[3] R documentation portal for detailed understanding: URL: https://www.rdocumentation.org/

[4] Research and education community portal from University of California (article by Leo Breiman and Adele Cutler) URL: https://www.stat.berkeley.edu/~breiman/RandomForests/cc_home.htm

[5] Learning books from Stanford University: https://lagunita.stanford.edu/c4x/HumanitiesScience/StatLearning/asset/trees.pdf

[6] Market Making with Machine Learning Methods by Kapil Kanagal, Yu Wu, Kevin Chen (https://web.stanford.edu/class/msande448/2017/Final/Reports/gr4.pdf)