Use Case AGUC001

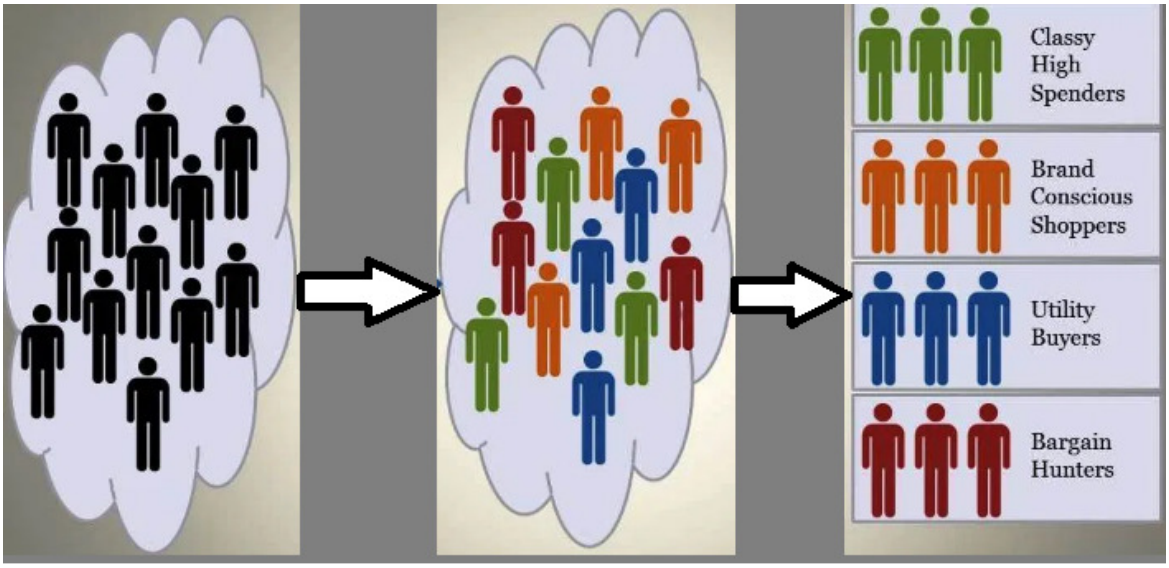
Agilytics Technologies Pvt Ltd.

Table of Contents

Introduction

Humans involuntarily behave like other similar humans around them. We all follow an involuntary pattern and this pattern detection is precisely the idea behind marketing analytics. The task for marketing analytics is to identify groups of people similar in their attributes to learn more about them. This method is also referred to as look-alike-modeling. This helps businesses to devise a sound marketing strategy for different groups of people based on their needs and consumption pattern.

To get a quick feel for workings of marketing analytics, consider the following schematic.



You may start with the raw customer data and then using data mining and advanced statistical tools, you can try to identify hidden patterns in customer spends behavior. Ultimately, you can superimpose business knowledge and rules on top of this **classified** portfolio to formulate business and marketing strategy for business growth.

In the present Use Case AGUC001, will get a better understanding of the Marketing Analytics Modeling.

Marketing Analytics

Suppose you have recently joined in as the chief of Analytics and Business Strategy at an online shopping store called FashionCart, that specializes in apparel and clothing. One day you had the chief marketing officer of the company come rushing to you office looking unusually worried . The board of directors has given him tough targets for sales and slashed his marketing budget into half at the same time. You immediately identify that you are dealing with a common business problem of improving business revenue with reduced cost. You have also realized that this is a great opportunity for you to establish analytics practices in the company since there is a quick opportunity for you to improve the P&L (Profit & Loss) income statement.

Additionally, the CMO informed you that last year they had carried out marketing campaigns with different offers on the product catalog. A direct mailing product catalog was sent to some hundred thousand customers from the base of over a couple of million customers last year with the response rate of 4.2%. The direct mailers were later followed up with SMS and email messaging.

To explain your strategy to the CMO, you drew a quick and dirty campaign P&L statement on the white board in your office. The following is a version of your drawing

Profit & Loss Statement for Marketing Campaign

Revenue Component	Cost Component
Cummulative purchases by customers during marketing campaign	Marketing Campaign Fixed Cost Variable Cost - Email/SMS/Direct Mailing of the catalogue
SUM OF (Value Generated by the customer)i	(Mailer Cost X Number of Customers)

The objective is to maximize cumulative value generated by customers while minimizing total mailer cost. You explained that the analysis will have an impact on the variable component of

the campaign i.e. you will reach out to the right set of customers and generate maximum value. Additional there is an intangible benefit of this exercise i.e. reduced customer dissatisfaction from unsolicited offers.

The CMO left you as a much less worried man than when he entered your office. However, you know that you had your work cut out, and you need to think of the right approach to solve this problem. You are up for the challenge!

Model Selection

Through your rigorous exploratory data analysis, you have found several factors that play crucial roles in marketing campaigns' response for customers, some of these factors are:

- **Recency**: Number of recent visits to the company's website and purchases
- **Frequency**: The time lag between purchases in the last 6 months
- **Payment mode used**: cash on delivery, credit card, internet banking etc.
- **Marketing data**: life-stage segmentations (i.e. luxury buffs, up-scale ageing, first-time earners etc.)
- **Last year's expenditure trend**: amount spent last year
- **Coupon usage pattern of customer**

You have tried several multivariate models mentioned like logistic regression, SVM (Support Vector Machines, decision trees etc.) to model customers' behaviour and generate **purchase propensity scores**. The choice of right model selection depends on the following 2 factors i.e.

1. Predictive power of models
2. Business and operations integration

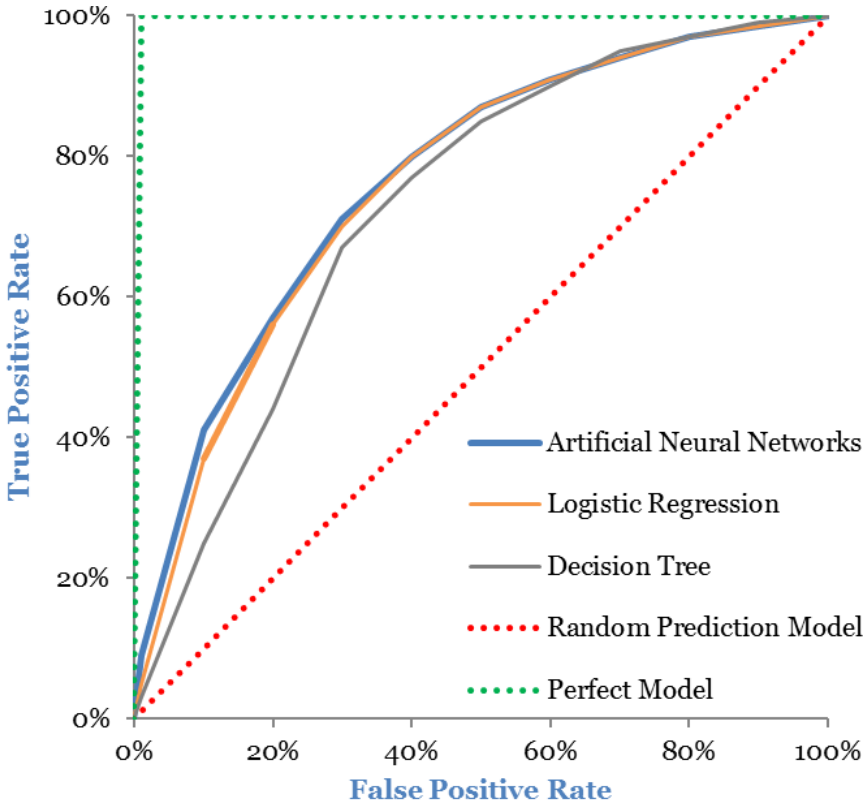
Predictive Power of Models

The first factor for model selection is the overall predictive power that the model has in comparison to other models. For this **classification problem**, the Area Under Receiver Operating Curve (AUROC) is possibly the best way to assess the predictive power of models.

Sometimes people also use Gini coefficient for assessing predictive power of models, Gini is another variant of AUROC and mathematically represented as:

Gini = 2 X AUROC - 1

In the following plot, AUROC is displayed for artificial neural networks, logistic regression, and CART decision tree.



Please note that the perfect model curve (in green) here is with 100% predictive power, and random model (in red) represents prediction through the flip of a coin. The AUROC values for the test sample for the three models are:

Model	AUROC
Decision Tree	72%
Logistic Regression	76%
Artificial Neural Networks	77%

Decision tree here is performing much below the other two models. This is often the case with decision trees, but they are still very useful and popular because of their highly intuitive and easy-to-explain solutions. Artificial neural networks are performing a notch above logistic regression in this case with a slightly higher area under ROC. Hence by the first criteria, artificial neural networks offer the best model among the three models.

Business and Operations Integration

This aspect of model selection is equally important. The model selection must be based on **productization** of the model for business usage in the long run. The following factors are useful to keep in mind at the beginning of modeling process:

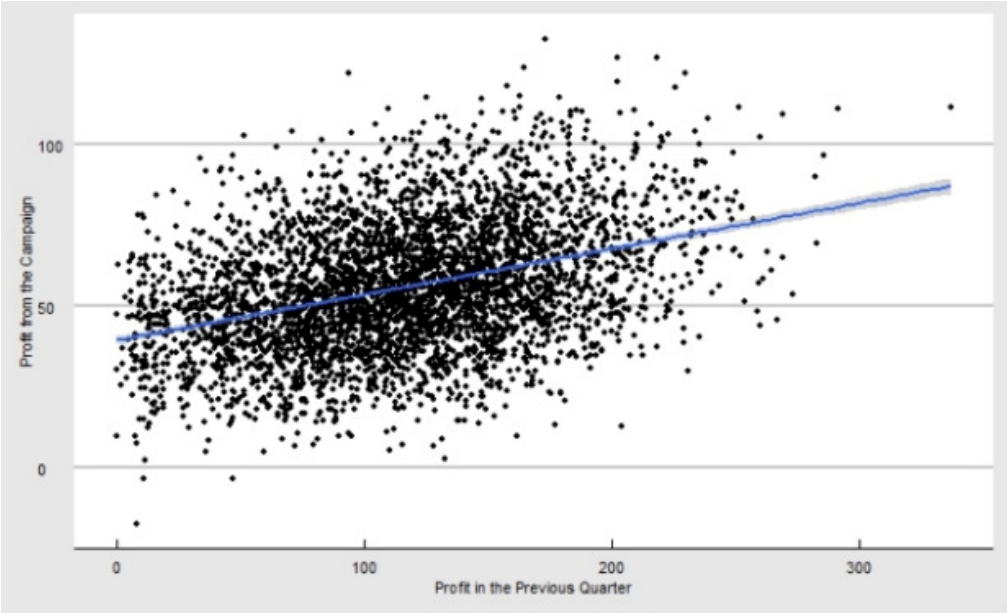
1. **Consistent availability of data for all predictor variables**: Many times models are developed by predictor variables that are hard to procure regularly and consistently. Keeping such variables in the model is not advisable even if they contribute to high predictive power. This is especially true for third party data which is purchased once in a while.
2. **The model should be simple enough to calibrate**: This factor is really important if the model will be used for a long duration i.e. more than two years. Certain models are relatively easy to calibrate or alter according to changes in market environment. This way analysts don't need to rebuild a new model every so often.
3. **Integration with information system and business process**: The goal of any model is to integrate well with IT systems used by business users. Analysts must think of

productionization of the model for business process integration at the beginning of the project to avoid unnecessary rework at the completion of the project.

4. **Business users' commitment for regular usage of models**: Data science is not just an intellectual exercise. The most important aspect of data science's success is the generation of business value through actionable insights, and business users' commitment to act on these insights. This commitment by business users come from their involvement in, and understanding of model building process. Data scientists need to communicate well with business users throughout to gain their trust.

Regression Model

Let's create a regression model to estimate the profitability of every customer for campaign management. We will examine a continuous variable 'profit generated by the customers in the previous quarter' to determine the profit they will generate through campaigns. The following is the scatter plot for these two variables:



Regression Model

There is a definite correlation between the above variables. If we calculate the correlation coefficient or Carl Pearson co-efficient, it's value is 0.372 (positive correlation).

The relationship between these two variables is mostly correlation. Profit in the previous quarter is definitely not causing profit from the campaigns. However, both these variables are governed by the same unobserved factors (driving forces) such as customers' affinity of purchasing from the online store, and their capability to spend. Hence, this correlation is not spurious or coincidental. As an analyst, it is absolutely important to distinguish between correlation, and coincidence through rigorous logic.

Now, let's create a simple regression model between these two variables

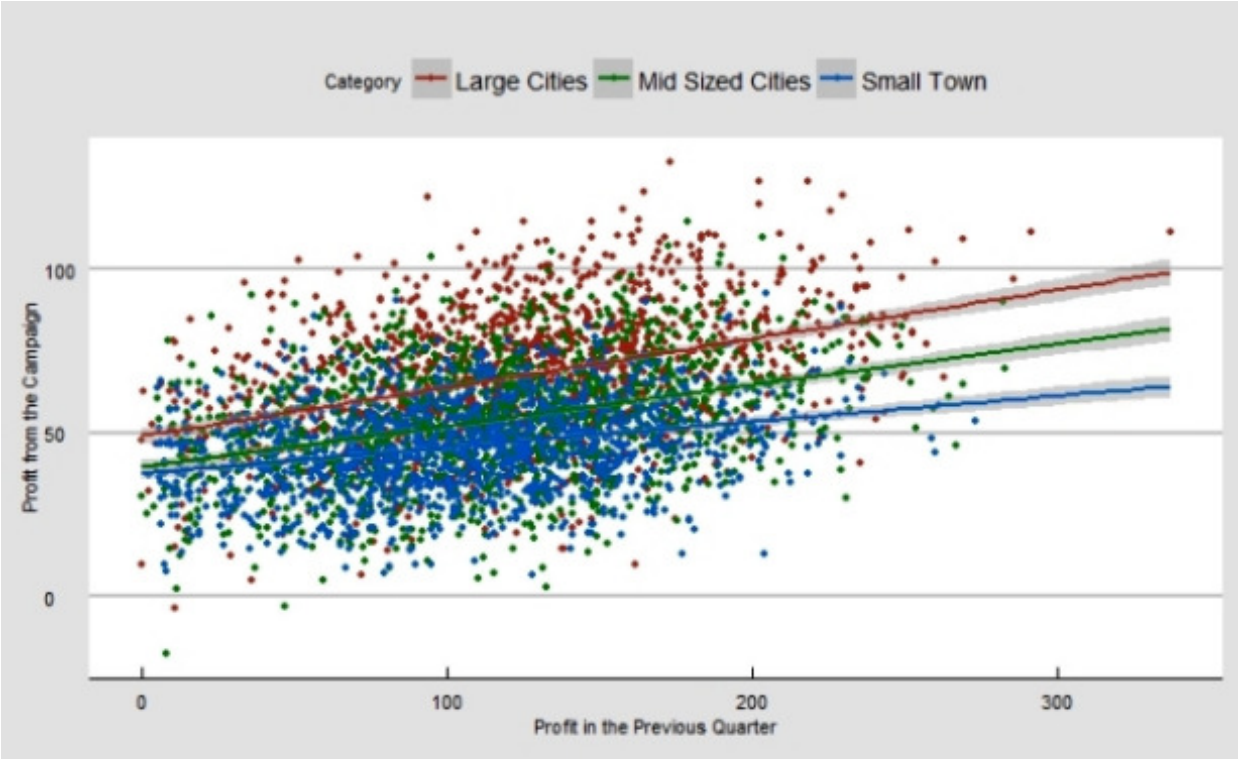
Regression Model	Estimate	Std. Error	t Value	Pr(>|t|)
(Intercept)	39.78	0.683	58.25	<2e-16
Profit in the Previous Quarter	0.14	0.005	25.96	<2e-16
Multiple R-squared:	0.138			
Adjusted R-squared:	0.138			
F-statistic (P Value)	2.2E-16			

The following is the linear equation for the above regression model

Profit from the campaign = **39.78** + **0.14** X **Profit in the previous quarter**

The model explains about 13.8% (R-square) variation in 'profit from the campaign'.

Now, let us extend this model by adding the categorical variable from the last time i.e. 'category of the location'. Let us first create the same scatter plot with the overlay of this categorical variable.



Regression Model by Category

In theory, you expect the above three lines for 'location category' to be perfectly parallel to each other. However, in practice, you will rarely find perfectly parallel (or zero interaction) lines. In our case the lines are following the same trend with very little interaction hence we can just add this categorical variable in our above model. The following is the new model after adding 'location category':

Regression Model	Estimate	Std. Error	t Value	Pr(>|t|)
(Intercept)	33.95	0.686	49.47	<2e-16
Large Cities	19.24	0.634	30.34	<2e-16
Mid-Sized Cities	7.29	0.626	11.64	<2e-16
Profit in the Previous Quarter	0.11	0.005	23	<2e-16
Multiple R-squared:	0.296			
Adjusted R-squared:	0.295			
F-statistic (P Value)	2.20E-16			

Notice, that the adjusted R-square value for this combined model (0.295) is greater than individual continuous (0.138) variable regression model. This is the process of regression model development where every incremental variable inclusion in the model will improve the R-squared value.

Conclusion

The primary task while initiating an analytics project is to clearly define the end goals/objectives of the exercise. The projects that influence the financials of a company goes a long way. Hence it is a best practice to link the end goals/objectives of your projects to the financial influence.

Website: https://www.agilytics.in

Contact: bd@agilytics.in

Phone: +91 9810884817