Data Processing / Machine Learning Assignment

Jason McCoy

Northwestern University

Introduction

The aim of the marketing division at Outdoor Power Equipment depot is to increase cross-selling and improve the effectiveness of their marketing campaign. This essay aims to provide recommendations on the use of machine learning algorithms that may be used for creating predictive models towards increasing cross selling and improving the campaign. A new product ‘Lawn Mower Toro Model XL234’ would be brought into the existing product line soon. The essay focuses on two scenarios.

Scenario A: What is the probability that an existing customer would buy the new product?

Scenario B: When customers buy the new product, what are the other accessories or products they would tend to purchase?

Scenario A

In order to find the probability that an existing customer would buy the new product, supervised machine learning algorithm would be used. In particular, logistic regression can be used. The dependent variable is binary – whether the existing customers buys a new product or not. There are several explanatory variables that can be considered, for example, when the last product was bought, whether it is still in warranty, what the nature of transaction was, what the equipment type and manufacturer was, etc. While building a logistic regression for this problem, the idea should be to maximize the likelihood function and include variables that are significant predictors. Overall, the predictive power of the logistic regression should be high. Once the logistic regression is done, concordance and discordance and other performance measures can be checked. In the logistic regression, the link function is logit. If however, a person is not comfortable with specifying the link function, a Generalized Linear Model (GLM) can be used instead of logistic regression.

While regression is the most common supervised learning algorithm that can be used in this scenario, some classification algorithms can also be effectively used since our dependent variable is binary. Decision trees and random forest can be used for classification by creating a multitude of ‘random’ decision trees. However, here I describe Support Vector Machines (SVM) for classification. SVM is a supervised machine learning algorithm where we would plot each of the data point in n-dimensional space, where n is the number of features which can be used. We would have two classes – one where the existing customer buys a new product and the other where the existing customer doesn’t buy a new product. SVM finds a hyperplane which differentiates these two classes. SVM maximizes the distance between the nearest data point from either class (called margin) and then decides on an appropriate hyperplane. If our data turns out to be complex and non-linear separation is required, then SVM uses kernel trick. SVM would help in separation of the two classes but would not directly provide the probability estimates, i.e. probability of an existing customer buying a new product. For that, logistic regression or GLM have to be used.

For this scenario, we selected supervised machine learning algorithms because we could use the existing data in an organized manner to derive insights. In particular, we could take the set of inputs and get output from it. Most importantly, we knew the algorithm’s possible outputs which makes it appropriate to use logistic regression, GLM or SVM.

Scenario B

In Scenario B, the focus is to find that when the customers buy a new product, what are the other accessories or products they would tend to purchase? This is cross-selling. Since Outdoor Power Equipment depot wants to increase cross-selling and improve the effectiveness of their marketing campaign, it is essential to have a suitable model for this scenario. In this scenario, we do not know the algorithm’s possible outputs. Hence, supervised machine learning algorithm can’t be used. So, unsupervised algorithms are used, which may be called true Artificial Intelligence. An algorithm that can identify complex procedures, patterns and processes without much human involvement would be ideal in this scenario. In particular, for identifying what other products or accessories the customer can buy, market based analysis or association rules can be used. For each customer, transaction/purchase history would be studied to find frequent patterns, associations, correlations, or causal structures among different set of items. These associations and correlations would then be used to *understand customers’ purchasing habits*. This would then make it easier to identify rules to predict the other products or accessories a customer can buy when buying a new product. Any rule would be in the form:

Antecedent → Consequent [support, confidence]

Support and Confidence are user-defined measures of interestingness. An example of a rule would be:

Buys(x, “new product”) → Buys(x, “another product B”) [15%, 60%]

This rule would mean that when a customer is buying a new product, he/she would tend to buy ‘another product B’. Here, support is 15% which means that probability of buying both “a new product” and “another product B” is 15%. It is the frequency of the rule within the transactions. Confidence is 60% which means that the probability of buying both given that the customer buys “a new product” is 60%. It is the conditioned probability. Lift can be calculated as the ratio of confidence and the probability that a customer would buy a “new product”. This is a very simplistic example. When association rules are used for answering Scenario B, there will be many transactions that needs to be considered before considering a rule to be statistically significant. In the end, the answers from the association rules would be that a customer buying a new product has 60% likelihood of also purchasing product X or 30% likelihood of also purchasing accessory Y, etc.

The market based analysis or association rules (which is unsupervised machine learning algorithm) can be effectively used for identifying cross-selling opportunities. Cross selling is important because it helps build customer equity, enhances market position, stimulates universe expansion, discourages customer attrition, and balances growth between new and existing customers.

Conclusion

The conclusion is that both supervised and unsupervised machine learning algorithms can be used under different scenarios to help gain insights on customer behavior, which in turn can be vital for the success of marketing campaigns.

References

Davenport, T. H., & Harris, J. G. (2017). Competing on Analytics The New Science of Winning. Boston, MA: Harvard Business School Publishing Corporation.