Northeastern University

Project Title: In-Vehicle Coupons and its Successful Targeting

Team Members

Sichu Chen (chen.sic@northeastern.edu)

Yiying Liu (liu.yiyin@northeastern.edu)

Shimeng Wang (wang.shim@northeastern.edu)

Crispin Sujith Cletus (cletus.c@northeastern.edu)

1. Abstract

The spending on the commercial campaign is enormous and has been growing every year. In North America, the advertisement expenses amounted to 248 billion in 2020(Statista). However, the effectiveness of advertisement is very hard to quantify and has never been used to justify its cost. As we entered the big data era, we can gather information about potential target customers' opinions and therefore may achieve better customer targeting. In this project, we intend to build machine learning models to distinguish targeting customers from others, which ultimately may help businesses to attract more customers with less cost. The dataset is generated from a survey, containing information about customers, coupons, and business characteristics. The dataset is processed and split into training and test dataset. Naïve Bayes, Logistic regression, and SVM will be trained and evaluated to see which model does the best job in classifying potential customers. The overall goal of this project is to help businesses better identify their target customers, generating more revenue with less cost.

2. Introduction

The advertisement is the backbone of Capitalism. There is a well-known quote saying that half my advertising spend is wasted; the trouble is, I don't know which half (John Wanamaker). This famous quote points out the inefficiency of advertisement. Money spent on advertising could have been used more efficiently or practically to create more job opportunities or devote to R&D.

Fortunately, with the rise of the internet and enhanced computational power, the availability of big data grants businesses the opportunity to exploit the value behind data and to capture their targeted customers with more efficient commercial distribution. In this project, we will focus on coupons that are specifically available to the drivers. We will dive into the dataset to explore the characteristics of targeted customers, coupon type, business characteristics, and its successful targeting. To correctly identify successful coupon targeting, we will develop several machine models to help businesses to achieve better usage of business resources.

3. Data Description

This dataset we use is Data of In-vehicle Coupon Recommendation. The data was collected through a survey posted on Amazon Mechanical Turk. This Survey asks a person if he or she will use the coupon in different driving scenarios, for example, destination, weather, companion, and business, etc. The dataset contains 12684 instances of 25 attributes. 25 attributes are all categorical variables. The dataset is balanced as we have roughly two equal percentages of labeled groups. The information of successful coupon applications was gathered by asking the coupon audience

whether they would accept the coupon. Data can be free to access and download through the UCI database.

4. Methods

In this project, we use three machine learning models to train and classify successful commercial distribution—Naïve Bayes, Logistic Regression, and SVM. Overall, we are trying to utilize machine learning models to help businesses identify or classify a successful coupon distribution—here we define a successful coupon distribution as the customer's intention to apply coupon and to make purchases. By integrating all available information related to businesses, customers, and external conditions such as weather and distance, we want to help businesses to distribute their advertisements more wisely by distributing their coupons at the right time and to the right customers.

Naïve Bayes is our baseline model as it makes the most assumptions. Generally, Naïve Bayes is a probabilistic machine learning model that uses Bayes' theorem to perform classification tasks. Naïve Bayes is simple to implement and can be very fast since no iteration is needed in this algorithm. To implement the naïve Bayes method, we need to make two assumptions. Firstly, we suppose all the features are independent. We also assume that all the predictors have an equal effect on the outcome, which means one predictor does not have more importance in deciding than the other predictor. In our Naive Bayes model, we use Laplace smoothing to avoid zero probability.

Logistic Regression is another classification algorithm. It is used to predict a binary outcome based on a set of independent variables. First, we need to calculate the probability of a class happening given certain feathers; then the cost function called cross-entropy is calculated by taking the log of the probability. Then by trying to minimize the cross-entropy loss, we will be able to obtain the parameters such as w and b which maximize the log probability of the true label in the training data. The parameters trained will be applied to predict the probability of a certain class happening given a set of features. In our logistic regression model, we performed both Lasso and Ridge regularization, separately, to control the overfitting problem.

Support Vector Machines (SVMs) are a batch of supervised learning models used for data classification and regression. It was a widely used machine learning algorithm because it produces significant accuracy with less computation power. As for our soft-margin SVM model, we did not apply the kernel trick as our computation power does not allow for that. We use only 15% of the training data to train the model for the same reason.

5. Exploratory Data Analysis (EDA)

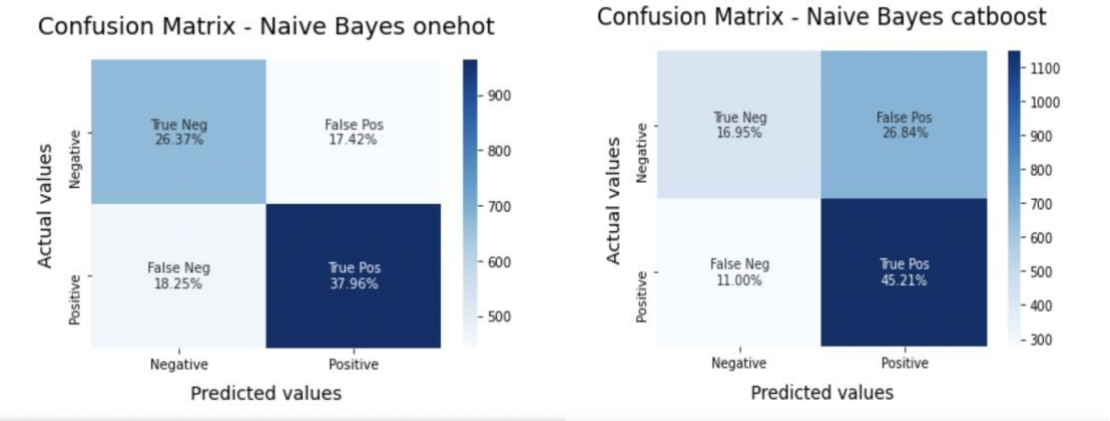
Among 25 attributes, the variable of Car has 12576 missing values, therefore being removed. Variable ToCoupon_GEQ5min is removed as it has a value of 1 for all cases. In addition, because variable Direction_opp has the same meaning as variable Direction_same, it will also be deleted.

Other variables that contain missing values were processed to grab the maximum frequency value respectively; the rationale behind that is to substitute the missing value with the most frequent observed value. All categorical variables are applied with one-hot encoding/Catboost encoding. Then the processed dataset was split into 80% and 20% for training and testing purposes.

After removing three useless ones, we still have 22 categorical variables. We use two methods – One-hot encoding and Catboost encoding –to transform these categorical variables so that they can be applied to models. One-hot encoding makes each category in one variable a vector. This is done for all categorical variables so that we have a total of 113 dummy variables after one-hot encoding. As for another Catboost encoding, starting from the first row, the running average of target labels grouped by each category is calculated and then replaces the categorical value. Both techniques use numerical variables to represent categorical variables.

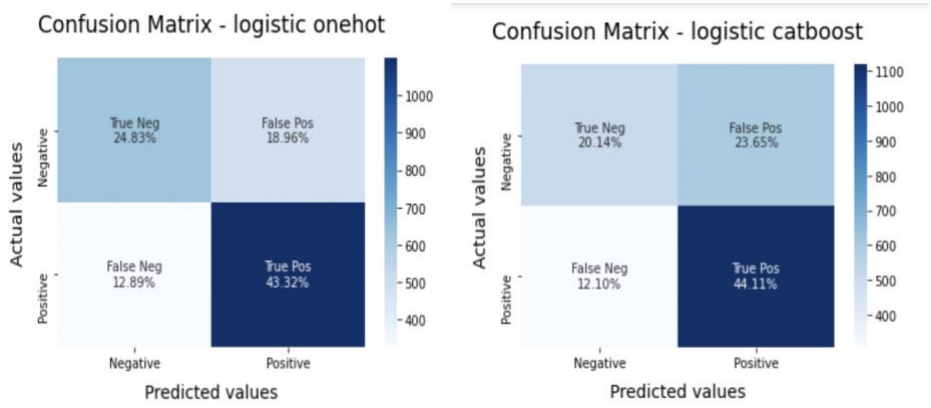
6. Results

Naive Bayes Performance

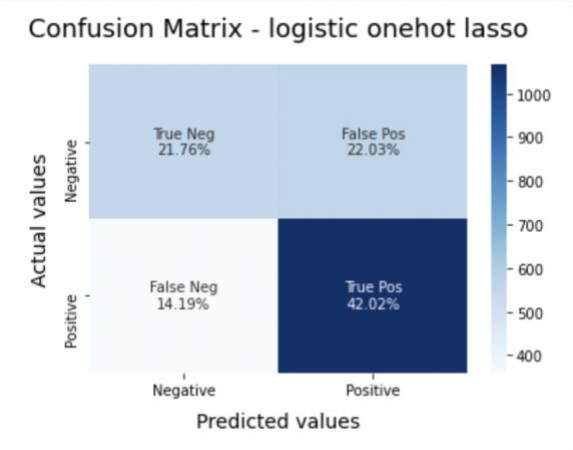


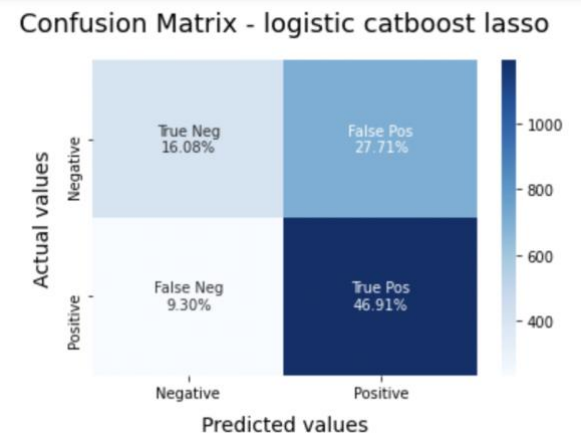
Naive Bayes with OneHot encoding has a total accuracy of 64.33%, whereas Naive Bayes with Catboost encoding has a total accuracy of 62.16%

Logistic Regression Performance

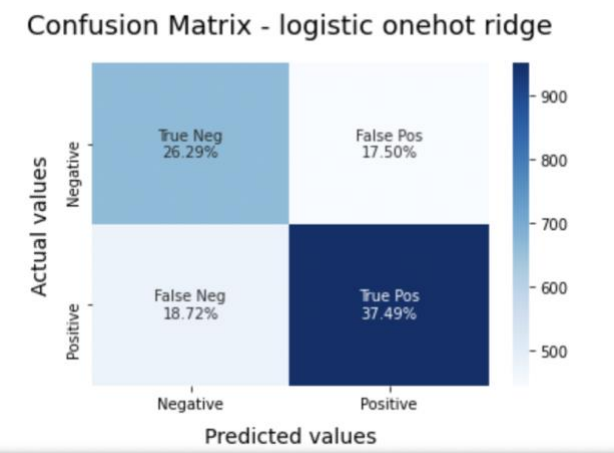


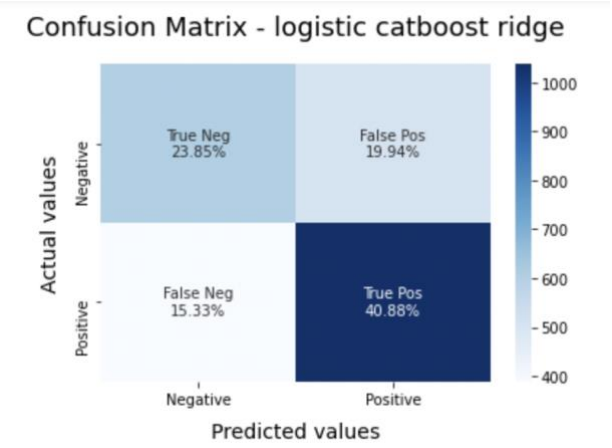
Logistic regression with OneHot encoding has a total accuracy of 68.15%, whereas Naive Bayes with Catboost encoding has a total accuracy of 64.25%.





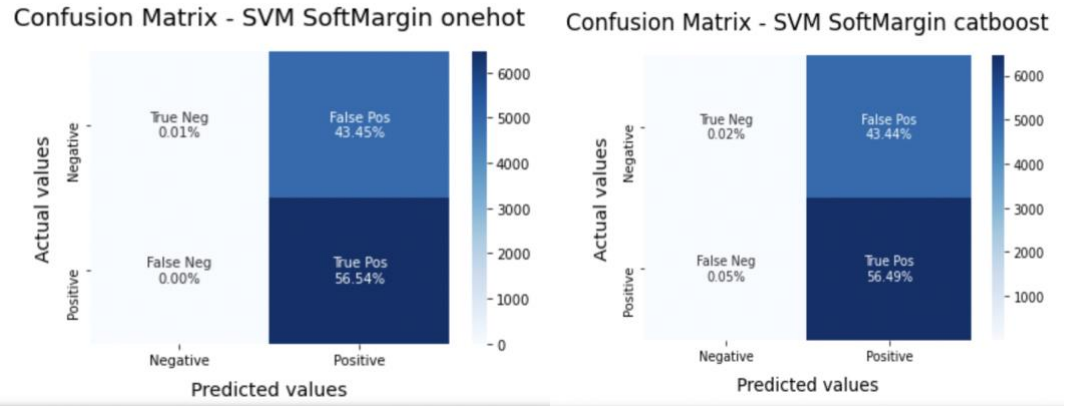
Logistic regression with OneHot encoding and Lasso Regression has a total accuracy of 63.78%, whereas Naive Bayes with Catboost encoding and Lasso Regression has a total accuracy of 62.99%.





Logistic regression with OneHot encoding and Ridge Regression has a total accuracy of 63.78%, whereas Naive Bayes with Catboost encoding and Ridge Regression has a total accuracy of 64.73%.

Support Vector Machine (SVM) Performance



SVM with OneHot encoding has a total accuracy of 56.55%, whereas SVM with Catboost encoding has a total accuracy of 56.51%

As for our project, the total accuracy is not the best measure of our interest target. We should take a further look into the recall as we are trying to target our potential customers as much as we can. Most importantly, the revenue from correctly targeting customers justifies the cost of distributing coupons.

	accuracy	precision	recall
logistic onehot	0.681514	0.69557	0.770687
logistic onehot lasso	0.637761	0.656	0.747546
logistic onehot ridge	0.637761	0.68172	0.6669
logistic catboost	0.642491	0.65096	0.784712
logistic catboost lasso	0.629878	0.628632	0.834502
logistic catboost ridge	0.647221	0.672067	0.727209
Naive Bayes onehot	0.643279	0.675316	0.685409
Naive Bayes catboost	0.6216	0.804348	0.627462
SVM SoftMargin onehot	0.565522	1	0.565484
SVM SoftMargin catboost	0.565084	0.99907	0.565305

From the graph above, Logistic Regression with Catboost Encoding and Lasso Regularization gives the best recall of 83.45%. Therefore, Logistic Regression with Catboost Encoding and Lasso Regularization is the model we should choose to distinguish between successful coupon distributions and unsuccessful ones. With the help of such a model, businesses will be able to find as much as 83% of successful coupon distribution/identify potential target customers. The generated revenue well justified the cost of the coupon, achieving a more effective and cost-saving advertisement campaign.

7. Discussion

We have already answered the first three questions (Q1, Q2, and Q3) through our prior explanations and they can be found in the previous pages. The questions that are left to be answered can be found below as follows:

Q1) Your understanding of the model selection - why and did you compare it with a baseline model?

We assessed/evaluated various candidate models like Logistic Regression, Naive Bayes, and other models to choose the final best one that addresses the problem at hand. We did this by considering a lot of key factors to evaluate if the model is "good enough". Some of the factors that we took under consideration were if the model meets the requirements and constraints of the project goals, if the model is as skillful as compared to naive models, and if the model is skillful relative to other tested models. In our project, the Naive Bayes Model is our baseline model as it takes the most assumptions. The criterion we use to conduct model selection is the Recall, which is the percentage of accurate identification of successful coupon applications among all successful coupon applications. The reason we choose this is that we want to identify as much as the successful coupon application and generate more revenue from that. Another reason for that is the low cost associated with distributing coupons. So by comparing the recall of other models with that of our baseline model, we will be able to find out the model we will choose.

After splitting the data into training and testing sets, we considered the following two types of model selection techniques, which are Probabilistic Techniques and Resampling Methods. We compared our results to a baseline model because the baseline helps us put a more complex model into context in terms of accuracy. Another advantage of comparing with a baseline is that it's easy to deploy, it is faster to train as few parameters can quickly fit in our data and it is very simple to detect problems easily.

Q2) Your understanding of Bias-Variance-trade-off and how you applied it to your project?

Bias-Variance-trade-off describes a tension between two error sources introduced by bias and variance in ML problems. Normally, a model tends to have high bias and low variance when the model is too simple; on the contrary, a model tends to have low bias and high variance when the model is too complex. To avoid oversimplification and over-fitting, we tried to use two different regularizations–Lasso and Ridge. The performances with two different regularizations are mixed–no single regularization performs dominantly better than the other. In our problem specifically, we

care more about the recall since the cost of distributing coupons to potential customers is relatively lower than the revenue generated from customers' purchases/consumption.

Logistic Regression with Catboost and Lasso regularization delivers the best recall–83.4% true customers could be identified among all true customers. This is reasonable as the Lasso Regression tends to decrease certain parameters towards 0 and performs auto-selection at the same time. On the other hand, Logistic Regression with Catboost and Ridge regularization delivers 72% recall, which is even lower than Logistic regression with Catboost and no regularization.

Q3) Your understanding of generalization and data-split and how you applied it to your project?

Generalization means how the model performs on the out-of-sample dataset. This is very important for machine learning problems as we are trying to either predict or classify based on new data. To ensure out-of-sample generalization, one of the keys is to ensure there are enough and different data points to train the models. This prerequisite is hard to be fulfilled if the entire dataset is not big enough; specifically, the instances are not enough considering the number of features. K-fold cross-validation is frequently used to solve problems associated with limited data points.

As for our project, the dataset contains 12684 data points. We divided the entire dataset to 80%/20% to conduct training and testing. Even with 113 dummy features after One-Hot encoding, our dataset is still valid to conduct training and testing. K-fold cross-validation is not necessary for our project.

References:

1. Statista,
https://www.statista.com/statistics/429036/advertising-expenditure-in-north-america/
2. John Wanamaker,
https://www.huffpost.com/entry/we-now-know-which-half-of-advertising-is-wasted_b_59a03a69e4b0a62d0987ae80
3. Data of In-vehicle Coupon Recommendation,
https://archive.ics.uci.edu/ml/datasets/in-vehicle+coupon+recommendation