Intelligent Advertising

Business Analytics using Data Mining

BADM Group 3B

Achintya

Aditya

Ankit

Kausambi

Susila

12/27/2012

Index

Executive Summary

Targeted advertising can be a big challenge for businesses especially retailers. For any business, advertising is a necessary evil. They need it to retain the customer, but they have to pay the heavy costs of advertising and irritating customers by promoting them products they might not need. Targeted advertising, which ensures that customers get advertisements / promotions for only those products they need or are most likely to buy, helps businesses on two ways. One, it reduces the advertising cost – since now you are sending your promotions to lesser but more relevant number of people, and second, it reduces the chances of irritating customers since you are sending them the promotions they really want.

In this project, we have tried to solve this challenge of targeted advertising for the retailer. We have tried to predict the most popular dairy level item in a basket based on customer's demographics information and past purchase patterns. We have used 'K-nearest neighbours' and 'Categorical and regression trees' Methods to create a classification model. Both the models have helped us achieve prediction with around 27% overall error.

We recommend the retailer to use this prediction model to run targeted promotions such as discount coupon campaigns for the predicted product. The predicted intelligence can also be used to run promotions for new product launches for the item a customer is most likely to buy.

Business Goal

To improve the effectiveness of targeted advertisement and thus consequently optimize advertisement costs.

Data Used

The plan is to use data and flesh out some insights that can provide tangible or implementable business level insights. We query the raw data table to find out the following information:

- *Age of Customer:* A simple excel formula that calculates the difference between the current year (2012) and the year of Customer's DOB
- *Revenue:* Extended Price * Quantity Sold
- *Tot_Rev_tillM-1:* The field quantifies the total revenue earned by the Hypermarket from the particular Customer; calculated from transactions/purchases made until M-1th time
- *Hi_Rev_Class_onM-1 :* The field lists the "Class" that contributed for the highest share in revenue earned by the Hypermarket through the customer's last purchase (M-1th time).

Data Visualizations

We started exploring the data by creating visualizations in order to observe any hidden trends or patterns that might be useful in our later analysis. Creating scatter plots of customers vs. city and quantities sold vs. customer, we see that the latter scatter plot indicates that some customers had made large purchases while for most of the others, purchase quantities were pretty much uniform. Zooming into the number of transactions per customers, we now see that indeed the range of data is very broad. Analyzing the most common sub-classes sold and their corresponding quantities (using cross tables) we see that, as expected, salt is the winning sub-class. Another thing we

note while plotting maximum and minimum quantities per customer is that there are negative quantities in this data - indicating that some baskets were billed incorrectly and then those items were removed from the bill by subtracting the quantity. However such anomalies are very small.

Data Mining Goal

We have a plethora of data available. Our data mining goal is to extract relevant information and data that will help us predict what the customer would be most likely to purchase. A corollary of this is to pin point the set of customers who are most likely to buy a particular product say in the event of a new product launch.

We are following a supervised model where we are specifying our input (predictor) variables and output variable. Our predictor variables are a combination of customer demographics and customer's past purchase pattern. Using this we predict the output variable, which is 'product most likely to be purchased' - in our case, it is the most popular dairy item in a basket.

Data Mining Methods

Our data had predictors which were either categorical or numerical and output which were categorical. As one can see, it is a supervised data, where input is the demographic data of the customer (age, gender, marital status) and his past purchase patterns (total revenue brought to the store, most popular purchase in last basket) and output is the most popular purchase in the current basket.

Since the output is categorical, we used 'K- Nearest Neighbours' and 'Categorical and Regression Trees' methods to get the prediction. Since KNN needs numerical inputs, we

used dummy variable to convert categorical data to numerical data. We partitioned the data into training (50%), validation (30%) and test data (20%) sets.

The KNN model gave us an overall error of 27.62% and the tool recommended a best K value of 16 (figure 1). CART gave us an overall error rate of 27.09% (figure 2) and the best pruned tree had 7 levels (exhibit 1).

Error Report			
Class	# Cases	# Errors	% Error
BUTTER	710	230	32.39
CHEESE	1007	97	9.63
DAHI &	385	137	35.58
FRESH MILK	244	65	26.64
ICECREAMS &	309	142	45.95
OTHER DAIRY	172	96	55.81
PANEER	193	67	34.72
Overall	3020	834	27.62

Figure 1: KNN

Error Report			
Class	# Cases	# Errors	% Error
BUTTER	710	301	42.39
CHEESE	1007	10	0.99
DAHI &	385	137	35.58
FRESH MILK	244	65	26.64
ICECREAMS &	309	142	45.95
OTHER DAIRY	172	96	55.81
PANEER	193	67	34.72
Overall	3020	818	27.09

Figure 2: CART

Evaluation Criteria

In our analysis we chose overall error as the performance metric of choice because that is what will measure the overall percentage of spillage in our advertising campaign. We could see that when compared to Naive method that gave an overall error of 66.7%, both methods – KNN and CART have given a better overall error rate of around 27%. Also we could see that both KNN and CART gave us almost similar results in terms of performance metric.

Recommendations

1. From this analysis we can predict the most popular dairy item in a basket. This intelligence can be used to create targeted promotion, such as customised discount coupons, for the customers.

2. Since this analysis is done at an item level, it can help us create customized promotions for new product launches. So if we know that somebody prefers cheese over other items, the store will to target him first for new product launches in the cheese category
3. To improve the campaign success rate, we will recommend collecting future feedback data on the promotions and using it as a predictor to build models for better prediction of most popular dairy item in the basket.

Exhibits

Exhibit A: CART tree

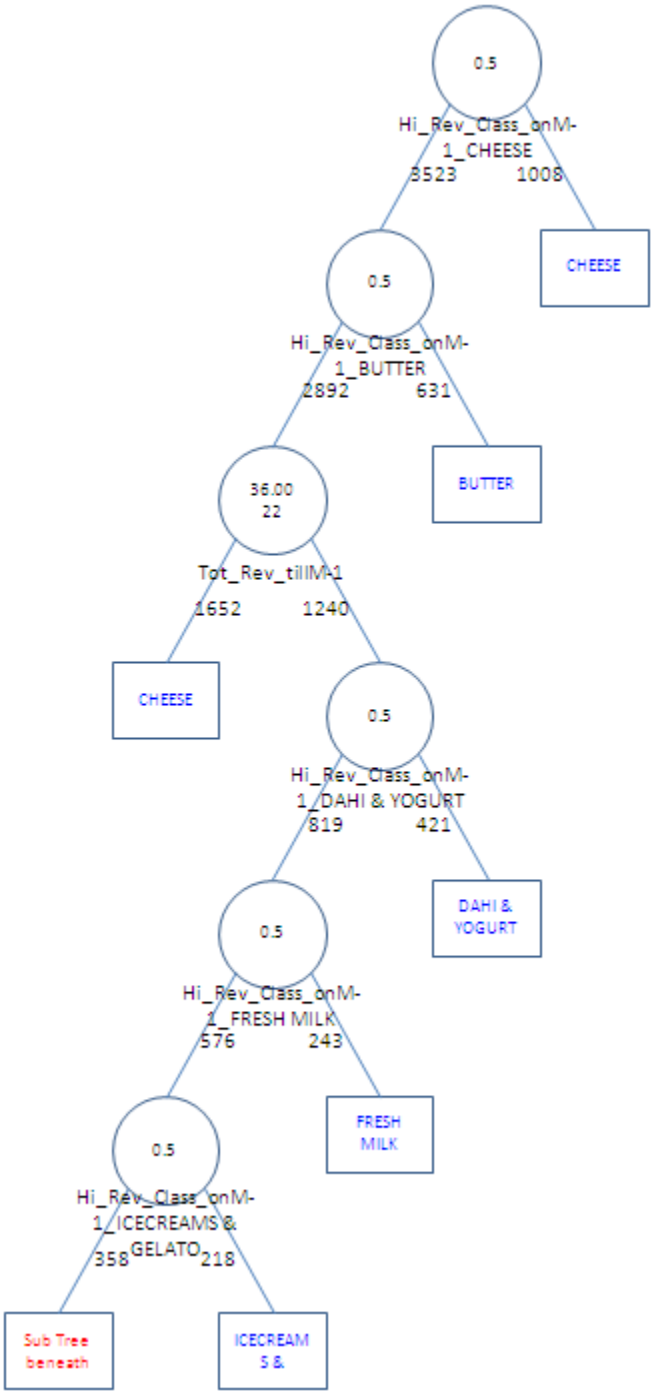
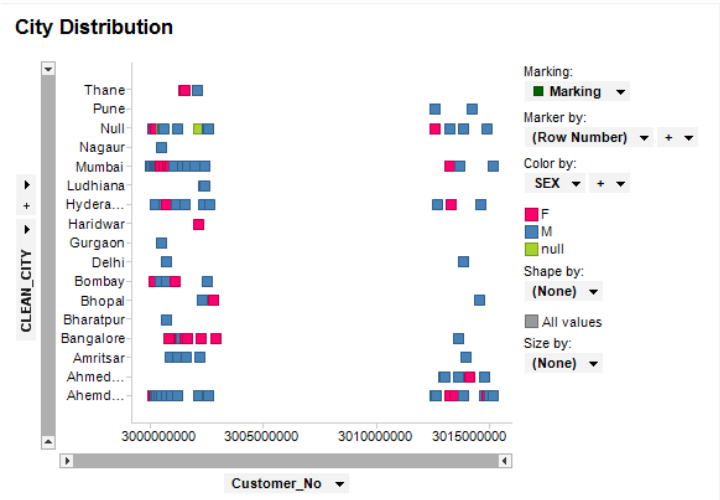


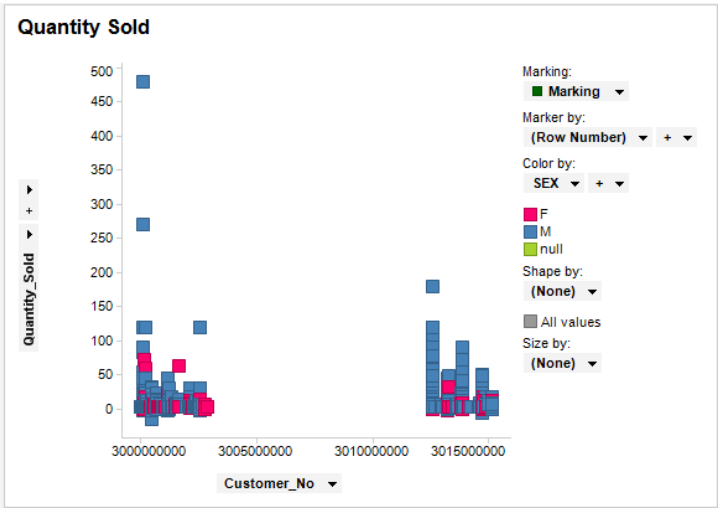
Exhibit B: Snapshot of how the data looks (limited columns displayed here)

Customer_No	DOB	SEX	Transaction_Date	Sku_Number	Quantity_Sold	Extended_Price	Item_Description	Revenue	Tot_Rev_tillM-1	Hi_Rev_Class_onM-1	Age of Cust
3000008858	20/09/82	F	17/10/11	100152313	3	135	NUTRALITE	405	0	NA	30
3000017362	21/10/67	M	17/08/12	100002570	3	435	AMUL BUTTE	1305	0	NA	45
3000039911	19/08/65	M	21/07/12	100002569	3	90	AMUL BUTTE	270	4455	CHEESE	47
3000039911	19/08/65	M	21/07/12	100100827	3	216	DAIRY CRAF	648	4455	CHEESE	47
3000039911	19/08/65	M	21/07/12	100002604	3	48	AMUL SHRIK	144	4455	CHEESE	47
3000039911	19/08/65	M	21/07/12	100226118	3	144	YAKULT PRO	432	4455	CHEESE	47
3000039911	19/08/65	M	16/06/12	100011989	3	360	NUTRA TABL	1080	981	BUTTER	47
3000039911	19/08/65	M	16/06/12	100133333	3	420	MD VANILLA	1260	981	BUTTER	47
3000039911	19/08/65	M	16/06/12	100002569	3	90	AMUL BUTTE	270	981	BUTTER	47
3000039911	19/08/65	M	16/06/12	100226118	3	144	YAKULT PRO	432	981	BUTTER	47
3000039911	19/08/65	M	16/06/12	100226118	3	144	YAKULT PRO	432	981	BUTTER	47
3000039911	19/08/65	M	15/10/11	100002811	3	297	BRIT CH SLI	891	90	CHEESE	47
3000039911	19/08/65	M	28/09/11	100101577	3	30	SARAS Dahi	90	0	NA	47

Exhibit C: Preliminary Visualizations During Data Exploration

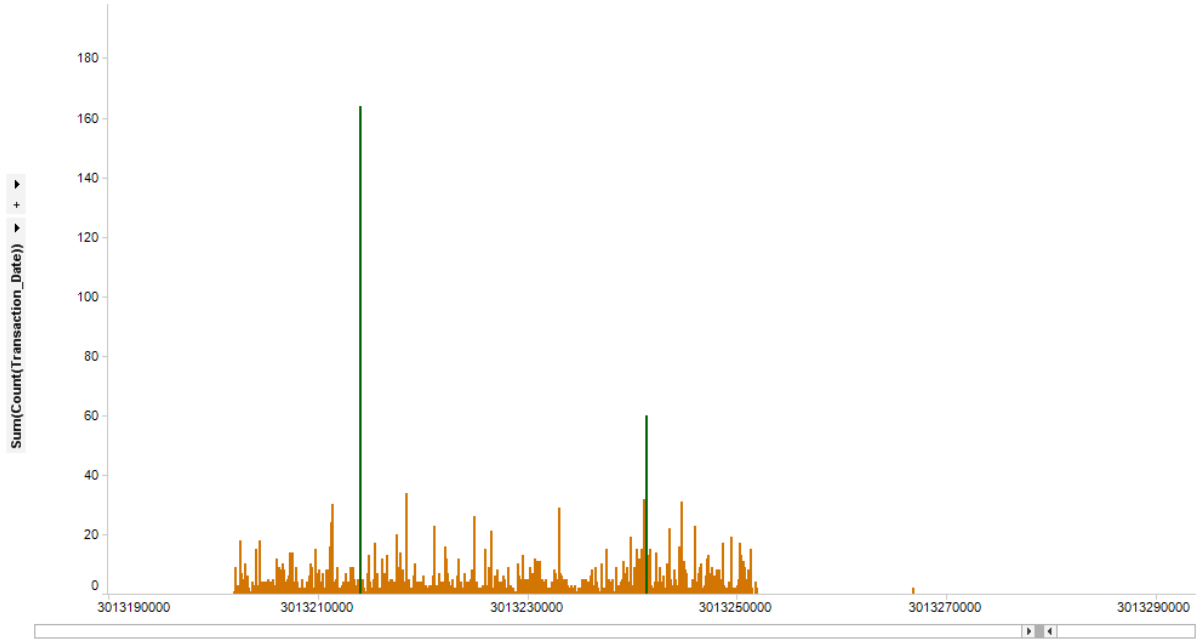
(1) Scatter plots for city vs. customer no. and quantity sold vs. customer no.





(2) Detailed zoomed in view of plot of #transactions vs. customer no.

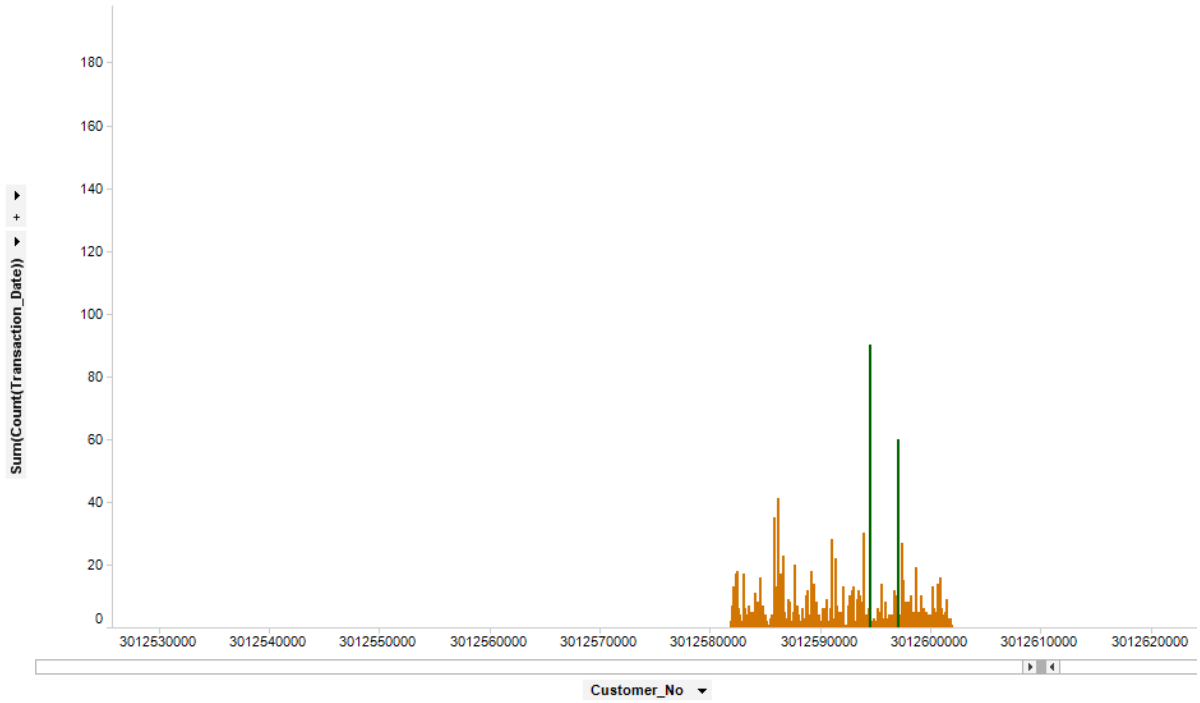
Number of Transactions per Customer



Details-on-Demand

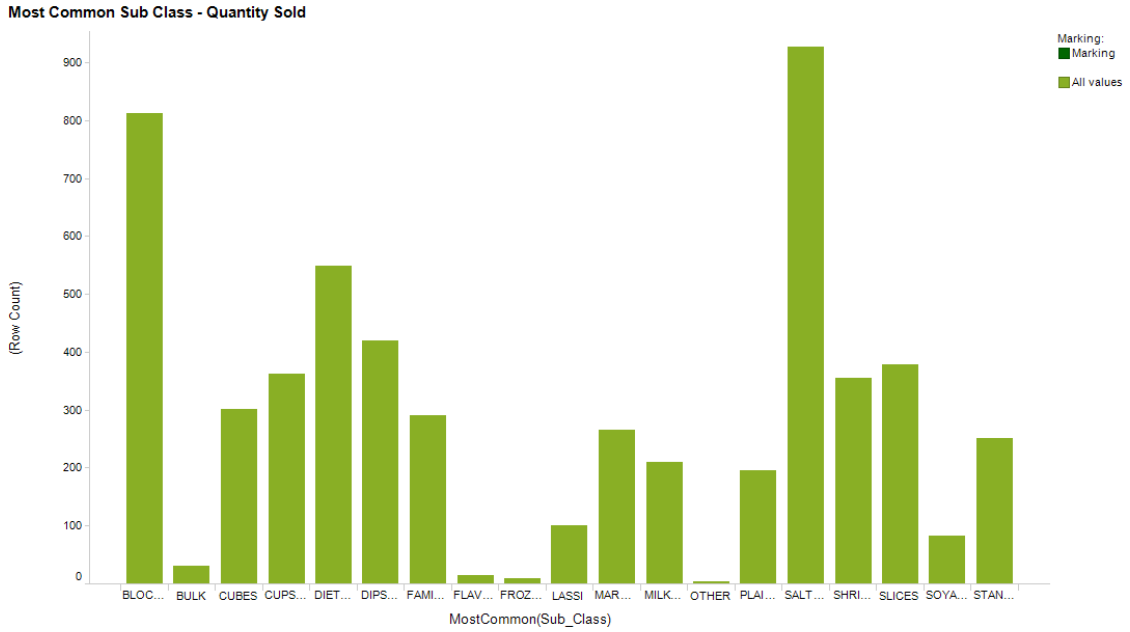
Customer_No	Count(Transac..
3013213883.00	164
3013241322.00	60

Number of Transactions per Customer



Customer_No	Count(Transac..
3012594473.00	90
3012596957.00	60

(3) Quantities sold for the most common sub classes



(4) Maximum and minimum quantities sold per customer