American Express Campus Analyze This 2017

Submission Deck by Closed&Bounded,DSE

Printed:

Team Details

Team Name : Closed&Bounded

Name	Campus	Roll No.	Mobile No.	Email Id
Suraj Kumar*	DSE	75205	9654506719	surajdhunna@gmail.com
Khalid Mushtaq	DSE			khalid.meer19@gmail.com

*** -** DISCLAIMER: All the work related to AMEX Analyze This 2017 was done by Suraj Kumar. Although, the team had two members but due to unforeseen exam/midterm, Khalid Mushtaq was not able to work on the project

Estimation Technique Used

3 – STEP ESTIMATION STRATEGY

STEP	Aim
Right Variable Identification	Steps were taken to find the right set of variables to be used in the model for estimation
Right Customer Identification	The purpose of this stage is get the list of 1000 customers to call.
Right Offer Identification	After generating the list of customer who are highly likely to respond, the most probable offer is being made to them.

STEP - 1

Right Variable Identification

Details of each Variable used in the logic/mode/strategy

Variable	Default Name	Notes and Inference from Individual Variable Graph (Histogram, Scatterplots)and Regression
cm_key	cm_key	- identifier for a particular customer
fam.size	mvar2	
total.active.cards	mvar3	- -Right Skewed
account.age	mvar5	- Right Skewed
cm.fees	mvar6	- Right Skewed
high.spend.affinity	mvar7	- Right Skewed
int.inf.score	mvar8	- Its always less then 1

income	mvar9	- Income is very weird (data imbalance) as value 2918974 is coming up with 13238 times out of 40,000 observations
		- After removing this troubling observation (2918974) and 0(Not data available) values, the income is following roughly normal distribution.
		- For income roughly 1400 observation were missing. They were imputed by running the following algorithm in r
		
		`library(dplyr)`
		`data$income[data$income == 0] = NA`
		`data$cust.spend.cap[data$cust.spend.cap == 0] = NA`
		`library(DMwR)`
		`data.Out <- knnImputation(data)`
		– This kind of knnImputation was also done for Leaderboard and Final Dataset

Details of each Variable used in the logic/mode/strategy

Variable	Default Name	Notes and Inference from Individual Variable Graph (Histogram, Scatterplots)and Regression
plcard	mvar10	- used as factor variable in r
int.busi.exp.score	mvar11	- not a normal distributions
spend.indus.code	mvar12	
freq.payments	mvar13	- normal distribution
cm.number	mvar14	
air.miles.mem	mvar15	
elec	$elec = mvar16 + mvar17 + mvar18 + mvar19$	- Its combines total electronic related spend of four quarter - There is not much information loss as electronic spend quarter variables are highly correlated among themselves - Its roughly follows normal distribution
trav	$trav = mvar20 + mvar21 + mvar22 + mvar23$	- Its combines total travel related spend of four quarter - There is not much information loss as electronic spend quarter variables are highly correlated among themselves - Its roughly follows normal distribution
hou	$hou = mvar24 + mvar25 + mvar26 + mvar27$	- Its combines total household related spend of four quarter - There is not much information loss as electronic spend quarter variables are highly correlated among themselves - Its roughly follows normal distribution

Details of each Variable used in the logic/mode/strategy

Variable	Default Name	Notes and Inference from Individual Variable Graph (Histogram, Scatterplots)and Regression
car	$car = mvar28 + mvar29 + mvar30 + mvar31$	- Its combines total car related spend of four quarters - There is not much information loss as electronic spend quarter variables are highly correlated among themselves - Its roughly follows normal distribution
retail	$retail = mvar32 + mvar33 + mvar34 + mvar35$	- Its combines total retail related spend of four quarters - There is not much information loss as electronic spend quarter variables are highly correlated among themselves - Its roughly follows normal distribution
total.spend	$total.spend = mvar36 + mvar37 + mvar38 + mvar39$	- Its combines total spend of four quarters - There is not much information loss as electronic spend quarter variables are highly correlated among themselves - Its roughly follows normal distribution
freq_extend_Supp	mvar40	
freq_extend_Elite	mvar41	
freq_extend_Credit	mvar42	
freq_accept_Supp	mvar43	
freq_accept_Elite	mvar44	
freq_accept_Credit	mvar45	

Details of each Variable used in the logic/mode/strategy

Variable	Default Name	Notes and Inference from Individual Variable Graph (Histogram, Scatterplots)and Regression
IV_accept_Supp	mvar49	- used as factor variable in r - Used as dependent variable in some logistic regression, random forest and support vector machines
IV_accept_Elite	mvar50	- used as factor variable in r - Used as dependent variable in some logistic regression, random forest and support vector machines
IV_accept_Credit	mvar51	- used as factor variable in r - Used as dependent variable in some logistic regression, random forest and support vector machines
IV_accept	$IV_{accept} = mvar49 + mvar50 + mvar51$	- dependent variable - used in Right Customer identification stage

Offer.Accepted		`data$Offer.Accepted =`
		`mapply(offer_accept,mvar49,mvar50,mvar51)`
		`offer_accept <- function(s,e,c){`
		`  if(s == 1){`
		`    return('Supp')`
		`  } else if(e == 1){`
		`    return('Elite')`
		`  } else if(c == 1){`
		`    return('Credit')`
		`  } else{`
		`    return('No Offer')`
		`  }`
		`}`

Details of Dropped Variable

Variable	Default Name	Notes and Inference from Individual Variable Graph (Histogram, Scatterplots)and Regression
card.type	mvar1	- Just one category, so not much useful for further analysis
cust.spend.cap	mvar3	- 24345 out of 40,000 observations were missing in this variables - Plus This variable information could be capture in other variable such as income, total.spend, high.spend.affinity variables - Thus due to non-availability and correlation with other model variables, it was decided to drop this variable
IV_extend_Supp	mvar46	- All of these IV_extend_XXXX variable are correlated with IV_accept_XXXX variables. - Plus the details of this information is not available in leaderboard or final dataset. - And all of the 40,000 observation have one of these IV_extend_XXXX as positive so it not much useful in further analysis - Individual logistic model of these variables didn't result in good observation
IV_extend_Elite	mvar47	
IV_extend_Credit	mvar48	

STEP - 2

Right Customer Identification

Right Customer Identification

Components	Details
Model Formula	`IV_accept ~ fam.size + total.active.cards + account.age + cm.fees + high.spend.affinity + int.inf.score +income + plcard + int.busi.exp.score +spend.indus.code + freq.payments + cm.number + air.miles.mem + elec + trav + hou + car + retail + total.spend + freq_extend_Supp + freq_extend_Elite + freq_extend_Credit + freq_accept_Supp + freq_accept_Elite + freq_accept_Credit`
Step Details	- In this Step, Logistic Regression, Random Forest and Support Vector Machines were used to calculate the $P(IV_{Accept} = 1)$ in each model. - Then these three models were combined in following way $P(IV_{accept} = 1) = P(IV_{accept} = 1 \| Logistic\ Regression) + P(IV_{accept} = 1 \| Random\ Forest) + P(IV_{accept} = 1 \| Support\ Vector\ Machines)$ - Then the observations were sorted in descending order in terms of $P(IV_{accept} = 1)$, to get the observations which has maximum $P(IV_{accept} = 1)$ - Then out of these ordered dataframe, first 1000 observations were selected to be send to next step of Right offer identification - Train Data =

Model Parameters	**Logistic Regression**
	`Accept.log <- glm(IV_Accept ~. -cm_key,`
	`data = dat,`
	`family = binomial(link = 'logit') )`
	Random Forest
	`Accept.rf <- randomForest(IV_Accept ~. -cm_key,`
	`data = dat,ntree = 500,`
	`importance = TRUE)`

Right Customer Identification

Components	Details

Model Parameters	**Support Vector Machines**
	`Accept.svm = svm(IV_Accept ~. -cm_key ,`
	`                   data = dat ,`
	`                   kernel = 'radial',`
	`                   cost = 1e5,gamma = 1,`
	`                   decision.values =T, probability =T,`
	`                   scale =F)`
	
	NOTE regarding SVM:
	Running SVM on training data of 40,000 rows and 25 columns on author computer took around 4 to 5 hours on each run. So optimization of SVM parameters was not properly done

Reasons	**Reasons for Aggregating all accepting customers to IV_accept** - **Imbalance Data:** In the training data out of 40,000 customer about only 9000 have accepted the offers. In these 9000, the division is roughly 3000 in each card category. So running any machine learning algorithm on *[31000 (negative,IV_Accept = 0) vs 9000(postive,IV_Accept = 1)]* is better than running *[31000(negative, IV_accept = 0) vs 3000(positive, IV_accept_Supp = 1) vs 3000 (positive, IV_accept_Elite) vs 3000(positive,IV_accept_Credit =1)].* It is because 9000/40000 > 3000/31000 or 3000/40,000. Thus aggregation leads to more balance classes thus better classification. **Reasons for using Logistic Regression, Random Forest and Support Vector Machine** - **Boosting**: It was thought that each of ML algorithm has its own weakness and strengths. While Logistic regression has may underfit the data, the random forest may overfit the data on the training set. The logistic regression is highly affected by outliers but Support Vector machine is not. So combining the result of three different machine learning algorithm, should result better classification

STEP - 3

Right Offer Identification

Right Offer Identification

Components	Details
FORMULA	**Offer.Accepted** ~ fam.size + total.active.cards + account.age + cm.fees + high.spend.affinity + int.inf.score +income + plcard + int.busi.exp.score +spend.indus.code + freq.payments + cm.number + air.miles.mem + elec + trav + hou + car + retail + total.spend + freq_extend_Supp + freq_extend_Elite + freq_extend_Credit + freq_accept_Supp + freq_accept_Elite + freq_accept_Credit
STEPS	- **Training Data** = `data[data$IV_accept == 0,]` - **Test Data** = output of best 1000 observations from STEP – 2 (Right Customer Identification) <u>Steps are as follows</u> 1. Model was trained on Training Data, i.e subset of training data in which IV_accept = 1, that is we are training on data where each customer has accepted some offer. 2. Three ML algorithm such as Multinomial regression, Random Forest and Support Vector Machines were used to estimate the Probability of Offer.Accepted. 3. Then these probabilities from 3 ML algorithms were combined class wise to get the total probability for each class as follows $P(Offer.Accepted = Supp) = P(Offer.Accepted = Supp\|Multinomial) + P(Offer.Accepted = Supp\ \|Random\ Forest) + P(Offer.Accepted = Supp\|SVM)$ $P(Offer.Accepted = Credit) = P(Offer.Accepted = Credit\|Multinomial) + P(Offer.Accepted = Credit\ \|Random\ Forest) + P(Offer.Accepted = Credit\|SVM)$ $P(Offer.Accepted = Elite) = P(Offer.Accepted = Elite\|Multinomial) + P(Offer.Accepted = Elite\ \|Random\ Forest) + P(Offer.Accepted = Elite\|SVM)$

Right Offer Identification

Components	Details
STEPS	STEPS 4. The Maximum of these 3 class(offer probabilities) is used as predictor for an observation to belong to that class. $MaxP = \max(P(Offer.Accepted = Supp), P(Offer.Accepted = Elite), P(Offer.Accepted = Credit)$ $Category.Predicted = Categoryof(MaxP)$ 5. Then this Category.Preidicted is attached to test dataset to output the results.. These results were then uploaded to AMEX

Model Parameters	**Logistic Regression**
	`Cat.log <- multinom(Offer.Accepted ~.   -cm_key,`
	`data = data, maxit = 500)`
	Random Forest
	`Cat.rf <- randomForest(Offer.Accepted ~. -cm_key  ,`
	`data = data,ntree = 500,`
	`importance = TRUE)`
	Support Vector Machines
	`Cat.svm = svm(Offer.Accepted ~.  -cm_key,`
	`data = data,`
	`kernel = 'radial',`
	`cost = 1e3,gamma = 1,`
	`decision.values =T,`
	`probability=TRUE)`

Right Offer Identification

Components	Details
REASONS	**Reasons for Training on data where IV_accept = 1** 1. It is done to better train the model to identify the features where IV_accept =1. It leads to better classification as we predict 3 classes, each of which has roughly 3000 observations. So it leads to better balance training data. 2. Output of Step 2 which is top 1000 observations which are most likely to accept will be more likely to say yes to some offer. So here the problem is not customer identification but the offer identification. And better offers can be made when we all the data such that IV_accept = 1 for training observation. 3. Leaderboard Score didn't improve even if I considered 5000 or 9000 observation where IV_accept = 0. I did this to better model those 1000 observation which may consist where consist around 30 – 40% of observation where customer will not accept anything. But considering them didn't gave me better leaderboard score. So I finally kept my training data to be where IV_accept = 1 **Reasons for combining the Multinomial, Decision Forest and SVM** 1. Boosting: By combining the probabilities from the these 3 models we will rule the possibilities of underfitting and overfitting the data, affect by outliers. It leads to better bias-variance tradeoff. It also leads to better classifier by concept of boosting where we combine individual classifier output to predict the final result.

Model Growth

LeaderBoard Score	Technique Used
19000	1. Randomly choosing 1000 observations out of 10000 observation 2. Randomly assigning 4 categories such as Supp, Credit, Elite or No Offer to these Observations
32000	1. Building 3 Separate logistic Regression on three categories such as (Supp/No Supp),(Credit/No Credit) and(Elite/No Elite) 2. Then combining there results in following way 3. Category.Predicted by finding out which Probability is greatest. Suppose if Supp is greatest then Category.Predicted will be Supp 4. Best 1000 observations were selected by summing the Supp, Elite and Credit Probabilities. Let then 1000 observations were selected by arrange the observation in decreasing order of this Sum
33000	Same as above but instead of running logistic regression. I ran three random forest models for each offer category such as Supp, Credit or Elite. But this lead to overfitting the training data.
35000	This method is fully described in the document
34000	I tried to build a more complex model. Such as follows 1. First selecting right customers by running IV_Accept by using Logistic, random forest and support vector machines. Same as decribed in STEP- 2 of this document. 2. Then I choose three class of Models such as Logistic, Decision Forest and Support Vector Machines 3. In Each of these 3 classes, I build 3 models for 3 categories such as (Supp/No Supp), (Credit/No Credit) and (Elite/No Elite). 4. Then I combine the output these three classes in similar way as STEP-3 of this document 5. Then I choose the best 1000 observation by sum of right customer probability and sum of all three categories probability. And I choose those 1000 observations for which sum was maximum

Reasons for Technique(s) Used

Why do you think this is the best technique(s) for this particular problem?

Keyword	Reasons
Two Step Process	- Instead of directly targeting the customers according to whether they should be offered Supp, Credit, Elite or nothing. This technique first good customers and then target the offers to these good customers. It solves the problem of data imbalance
Boosting/ Multi Model	- By combining the Logistic Regression, Support vector machine and Random Forest, this technique is very robust for wide range of datasets. It basically implementation of boosting concept.
Future Improvements	- The model could be made more robust by inclusion of neural network as well in trio of logistic, rf and svm. - Model could be better tuned, had if the author had access to a faster computer. SVM tuning was a very big problem in the data