Enhancing the Operation of Hsinchu13 GOOD Market by Ranking 5 Vendors

with Higher Probability for Extra Promotion

BADM Team 4
104078504 I-Chun Chao
104078509 Sherry Wu
104078514 Yi-Chun Chuang
104078515 Chia-Li Chien

Executive Summary

a. Market Introduction

Hsinchu 13 good market has been established in April, 2014, and opens every Saturday. The number "13" means the 13 townships in Hsinchu. That's, this market collects good, healthy products from 13 different area in Hsinchu. Most of the customers are the households living nearby, so the market will hold events for family as well.

The market has their own promotion channel "Facebook fan page". In January, 2016, the accumulative page "likes" is up to 4,950 in over 1 years.

Since this is almost the most important and the only way to publish their news and information to the customers. We wondering if there is any possible way to make good use of the fan page data to make a better promotion of the market.

b. Business Problem and Background

We took the manager of Hsinchu 13 GOOD market as our client.

The aim is to help the manager adapt a proper promotion strategy for the venders who really need. We try to classify the revenue fraction of each vender by whether it is above 3% or not. By doing that, the market manager can know if she should take extra promotional action to certain vender instead of spending extra cost on those who already got much attention. We hope we can save not only the advertising cost, but also the working hour of the staff who is in charge of FB fan page.

On every week, we can take a look on the revenue fraction of the vender we did promotion for to see its growth. If the fraction is growing, we can say that we are success.

c. Data Description

We get the sales data from 13 good market, including revenue, the product categories, and the activities they held. Besides market information, we also get the fan page data from Facebook insights. Then, we collect the temperature and other environmental data as factors as well. Finally, we combine all the dataset into one sheet.

We take "vender on each market day" as one record and the main output variable is if it need promotion (Yes / No).

d. Analytics Solution

According to the specialties of our data, we create a lot dummies first. Being terrified by so many columns, we run PCA. Then, we try to use several analytic methods, such like Logistic Regression, Naïve Bayes and x validation in Rapid Miner. In order to find some additional explanation for our data, we also take a shot at visualization.

e. Conclusions and Recommendation

Due to what we see from all the result, we recommend the market manager to take other more promotional channels and plan for the fall promotion activities. Although we did get a very large dataset at the end of the semester, we still get a good model to help the manager decide which vender should be promoted. And we hope that we can help our client save cost on advertisement and labor through a well performing model.

Detailed Report

a. Problem Description

I. Business Goal

Our client is the market manager of 13 GOOD market. We are going to help market manager to know about the performance of each vendor, and then help her take some promotional strategies. The vendor performance is evaluated based on revenue growth rate.

From the market manager's perspectives, he expects for an average 3% revenue growth every market day (revenue growth rate in 6 months will be 100%). The calculation of revenue growth rate is represented below:

$$\text{Revenue Growth Rate} = \frac{Revenue}{Average\ Revenue\ of\ last\ 3\ time} \times 100\%$$

If the revenue growth rate of a vendor is lower than 3%, it means that the vendor needs more promotions to achieve the selling goal. On the other hand, if the growth rate is more than or equal to 3%, no extra promotion is needed. Finding out the bad-performed vendors and then doing promotions to raise their revenue will be considered a success for our task.

In order to avoid the condition that all venders perform bad and all need extra promotion. We would like to rank the 5 worse-performed venders and provide extra promotion for them, considering the promotion usefulness and extra promotion cost, such as time, labor and advertising expenses.

II. Data Mining Goal

Our data mining goal is to rank 5 vendors with worst performance. It is a supervised and classification task. The classification output will be RGR (Revenue Growth Rate) <3 % (Yes/No), and eventually ranked by its probability. We will do prediction 3 days before market day (every Saturday), that is, on every Wednesday.

b. Data Description

- Dimension: 43 variables
- Size: 1168 records
- A record represents a vendor on certain market day

The dataset is showed as following. (Figure 1 and Figure 2)

Input variables						Output
Date	_video link	last 1 time_revenue	Product category_ fresh	F.13-17	M.13-17	RGR<3%
Vender ID	# post comment	last 2 time_revenue	Product category_ cuisine	F.18-24	M.18-24	
Post ID	#post like	last 3 time_revenue	Product category _ handicraft	F.25-34	M.25-34	
#fb post	#post share	Temperature of Wed.	Product category _ other	F.35-44	M.35-44	
_photo	New product (1/0)	Special holiday(1/0)	Activity type _handmade	F.45-54	M.45-54	
_status	Activity host (1/0)	Distance from market	Activity type _speech	F.55-64	M.55-64	
_link	Revenue	season	Activity type _lunch	F.65+	M.65+	

Figure 1

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S
1	Date	Vender ID	Post ID	#fb post	_ photo	_ status	_ link	_ video link	#post comment	#post like	#post share	New product (1/0)	Activity host (1/0)	Revenue	last 1 time_ revenue	last 2 time_ revenue	last 3 time_ revenue	Temperature of Wed.	Special holiday (1/0)
2	41832	1	217636091780856_241974329347032 217636091780856_237888473088951 217636091780856_239079736303158	10	8	1	1	0	28	969	40	0	0	3650	3260	2040	3720	31.1	0
3	41839	1	217636091780856_241974329347032 217636091780856_240155586195573	3	3	0	0	0	11	230	12	0	0	3600	3650	3260	2040	30.4	0
4	41846	1	217636091780856_247372168807248 217636091780856_247147065496425 217636091780856_249037058640759	8	6	1	0	1	12	785	45	0	0	2270	3600	3650	3260	29.4	0
5	41853	1	217636091780856_252006981677100 217636091780856_251247551753043 217636091780856_250475945163537	10	7	1	2	0	19	812	27	0	0	520	2270	3600	3650	31.7	0

	T	U	V	W	X	Y	Z	AA	AB	AC	AD	AE	AF	AG	AH	AI	AJ	AK	AL	AM	AN	AO	AP	AQ
1	Product category _fresh	Product category _cuisine	Product category_ handicraft	Product category_ other	Activity type_ handmade	Activity type_ speech	Activity type_ lunch	F.13-17	F.18-24	F.25-34	F.35-44	F.45-54	F.55-64	F.65+	M.13-17	M.18-24	M.25-34	M.35-44	M.45-54	M.55-64	M.65+	distance from market	season	RGR<3%
2	1	1	0	0	1	0	0	24	45	370	663	341	173	21	9	20	94	183	177	102	17	18.7	summer	No
3	1	1	0	0	0	0	0	18	36	320	567	236	141	24	6	23	73	139	126	87	12	18.7	summer	No
4	1	1	0	0	0	0	1	21	41	307	505	185	130	20	4	20	58	122	109	68	9	18.7	summer	Yes
5	1	1	0	0	0	0	1	23	40	320	563	203	144	19	4	28	56	155	152	89	12	18.7	summer	Yes

Figure 2. Hsinchu 13 good market dataset

c. Data Preparation Details

I. Missing Value

- There were some records without last 1 time_revenue, last 2 time_revenue and last 3 time_revenue since every vendor had its very first time market day. We removed the records with these missing value.
- For few missing value of Revenue, we use median of its past revenue.

II. Create Dummy Variables

We created dummy variable for **season** (2 categories).

III. Calculation

We calculated revenue growth rate using **Revenue**, **last 1 time_revenue**, **last 2 time_revenue** and **last 3 time_revenue**, and then, generate the output variable.

IV. Visualization

We did visualization to figure out some structure or hints from data. (Figure 3)

We found that revenue will be higher in summer than in fall. And the revenue growth rate is not proportional to #fb_post. We think it is because the effect of post mentioning a certain individual vendor is better than an overall post (overall post will be counted to the #fb post of all vendors mentioned in the post).

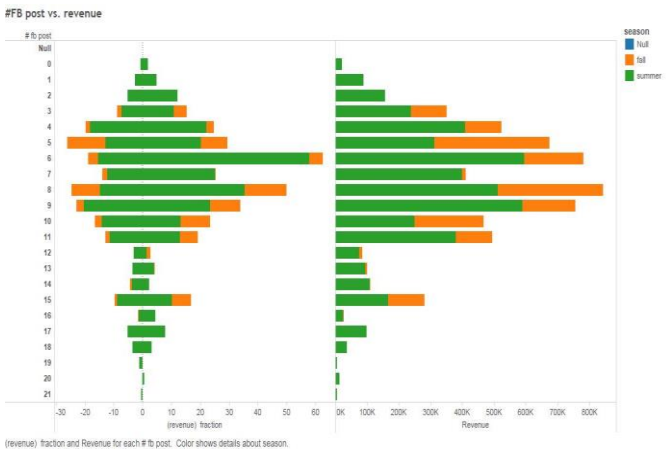


Figure 3

d. Data Mining Solution

- **Algorithms**

Since this project is supervised classification task, we tried many methods such as, Logistics regression, Classification Trees, K-NN and Naive Bayes. However both K-NN and Trees need great amount of dataset to generate good models, which implies our dataset is not large enough for these two methods. Therefore, we focus on doing LR and NB.

i. Logistics Regression

Logistics Regression is model-based so it doesn't require large dataset. First of all, we partition our dataset into training (50%), validation (30%) and test (20%) sets. Secondly, set the cutoff value as 0.5 and then choose "Stepwise" as the selection procedure. Finally, we get the LR model (Appendix).

ii. Naive Bayes

Since Naive Bayes can handle large amount of predictors and generate high performance accuracy when the classification goal is ranking. Firstly, we need to bin predictors. We try different numbers of bins for different variables either with equal count or equal interval and then we discover the performance will be better when we use the default value for numbers of bins with equal count. Secondly, partition the dataset into training (50%), validation (30%) and test (20%) sets. Thirdly, set "Yes" as successful class and the cutoff value as 0.5. From the summary report, we can rank the top 5 venders' probability for yes.

iii. Naive Bayes with X-Validation

Since our dataset is really small, so we would like to try X-Validation operator in Rapid Miner. (Figure 4 and Figure 5)

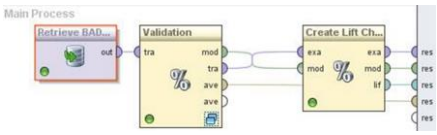


Figure 4. Main process

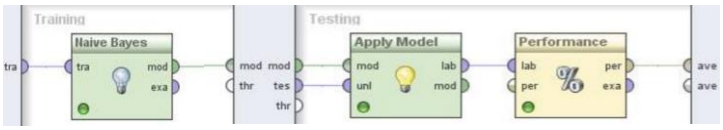


Figure 5. Training and testing process

- **Performance Evaluation**

i. Benchmark- Naive Rule

Naive Rule is classifying each observation as belonging to the majority class. If we classify all our dataset into the major class "Yes", needing extra promotion, and the overall error rate will be 47.5%. We set it as our benchmark. (Figure 6)

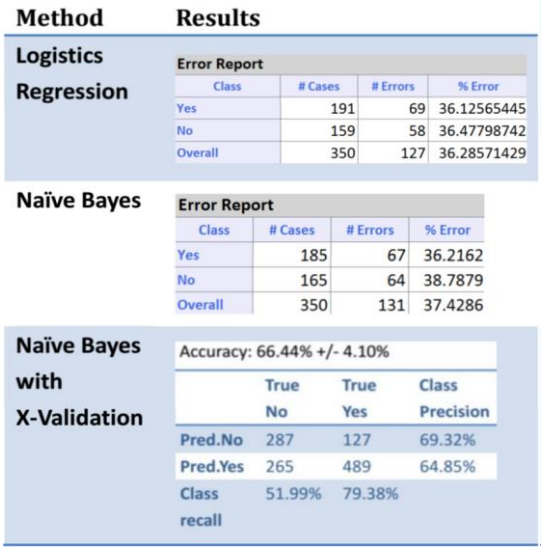
Naïve Rule

Error Report			
Class	# Cases	# Errors	% Error
Yes	616	0	0
No	552	552	100
Overall	1162	552	47.5043

Figure 6

ii. Error Comparison

From the validation summary reports, we can see that the error rate of X-Validation one is lower but it is quite similar to the other two. (Figure 7)



Method	Results
Logistics Regression	Error Report

Class	# Cases	# Errors	% Error
Yes	191	69	36.12565445
No	159	58	36.47798742
Overall	350	127	36.28571429

Naïve Bayes — Error Report

Class	# Cases	# Errors	% Error
Yes	185	67	36.2162
No	165	64	38.7879
Overall	350	131	37.4286

Naïve Bayes with X-Validation — Accuracy: 66.44% +/- 4.10%

	True No	True Yes	Class Precision
Pred.No	287	127	69.32%
Pred.Yes	265	489	64.85%
Class recall	51.99%	79.38%	

Figure 7

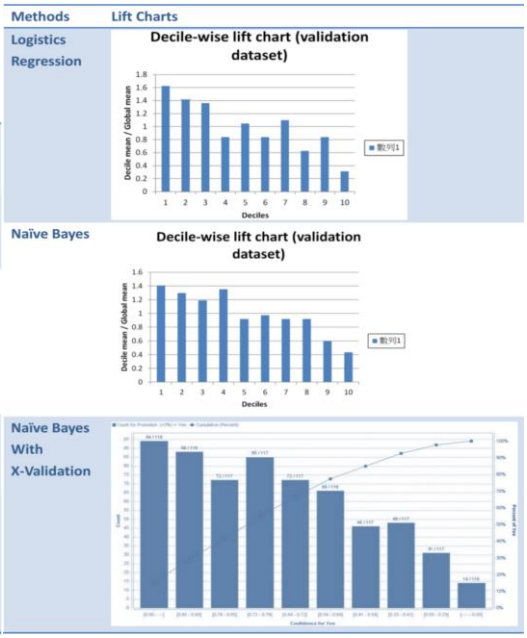


Figure 8

iii. Decile-wise Lift Chart Comparison

The first 10% of logistics output data performs 1.62 times better than Naive Rule and the second 10% performs 1.41 times better. In Naive Bayes, the first 10% is 1.4 times better than Naive Rule and the second 10% performs 1.29 times better. In NB with X-Validation, we would like to evaluate the first two 10% since we want to rank top 5 venders' "Yes" probability. (Around 25 venders would set up on every market days.) Therefore, we compute the first 20% accuracy rate and divided by the accuracy rate of Naive Rule, which is 52.5%, and then outcome will be 1.48. (Figure 8) Since our goal is ranking, we infer Logistics Regression performs better than the other two methods. In conclusion, we would like to choose Logistics Regression as our priority method.

e. Conclusions and Operational Recommendations

In this report, we have tried several methods to analyze our dataset and find out that logistics regression model performs better than naïve rule and other methods. In addition to the model, we also discover something relating to the promotion strategy from visualization parts. We think it can be taken as reference when the market manager is thinking about the promotion strategy.

- **Advantage**

  All the model we generated got better performance (over 60% accuracy rate) than naïve rule. So that we help the manager to concentrate on certain venders that need promotion accurately.

- **Limitation**

  Since Hsinchu 13 good market is too young to have large database. We believe we can build a lower error rate model with more records. Besides, single promotional channel could limit the information from other possible prospects. Final, the information of new vendors cannot be used due to the lack of last_time revenue.

- **Recommendation**

  From all the result, we recommend the market to take other promotional channels into consider to improve the analytic accuracy. And we strongly recommend the manager to plan for the fall, since the revenue in fall is averagely lower than in the summer. Besides, we also suggest that the market can try to keep the records of their customers (for instance, start a VIP program), so that the analytic can be done on more reliable data resources.

Appendix A: Logistics Regression Model & Stepwise

- **Logistics Regression Model**

Regression Model

Input Variables	Coefficient	Std. Error	Chi2-Statistic	P-Value	Odds	CI Lower	CI Upper
Intercept	0.674615	1.20266	0.314648967	0.574841306	1.96328	0.1859	20.7342
#fb post	-0.052886	0.12602	0.176107271	0.674740157	0.94849	0.7409	1.21424
_ photo	-0.053947	0.13372	0.162769474	0.686619424	0.94748	0.72904	1.23137
_ status	1.336817	0.27148	24.24766773	8.47096E-07	3.80691	2.23609	6.48121
_ link	-0.186851	0.17476	1.143185808	0.284980155	0.82957	0.58898	1.16844
_ video link	0.08517	0.54584	0.024346668	0.876006072	1.0889	0.37357	3.17402
#post comment	-0.008242	0.02355	0.122517838	0.726319575	0.99179	0.94706	1.03864
#post like	0.001221	0.00092	1.75202322	0.185622587	1.00122	0.99941	1.00303
#post share	-0.015403	0.0119	1.675987816	0.195458432	0.98472	0.96202	1.00795
New product(1/0)	-0.786485	0.564	1.944579012	0.163172213	0.45544	0.15079	1.37566
Activity host(1/0)	-0.298163	0.56848	0.275091444	0.599936657	0.74218	0.24357	2.26152
last 1 time_ reven	-1.81E-05	2.7E-05	0.461660921	0.496848761	0.99998	0.99993	1.00003
last 2 time_ reven	0.000107	4E-05	7.052362017	0.00791608	1.00011	1.00003	1.00019
last 3 time_ reven	3.48E-05	2.9E-05	1.415723268	0.234108617	1.00003	0.99998	1.00009
Temperature of W	-0.059091	0.04548	1.688359596	0.193817352	0.94262	0.86224	1.0305
Special holiday (1/	0.343491	0.51937	0.437395243	0.508382347	1.40986	0.50943	3.90181
Product category_	0.298798	0.25552	1.367404572	0.242258315	1.34824	0.81708	2.22468
Product category_	0.109593	0.24116	0.206520318	0.649508522	1.11582	0.69554	1.79007
Product category_	0.927854	0.29807	9.689980279	0.001852755	2.52908	1.41008	4.53607
Product category_	0.281876	0.39409	0.511592066	0.474450474	1.32561	0.6123	2.86991
Activity type_hanc	0.154024	0.26662	0.333719333	0.563477174	1.16652	0.69174	1.96717
Activity type_spee	-0.297487	0.33143	0.805665132	0.369404967	0.74268	0.38787	1.42205
Activity type_lunc	-0.51484	0.31336	2.699351255	0.100389088	0.5976	0.32335	1.10443
F.13-17	0.007156	0.02446	0.085589849	0.769860458	1.00718	0.96003	1.05665
F.18-24	0.012208	0.01221	1.00046136	0.317198898	1.01228	0.98835	1.03679
F.25-34	-0.014781	0.00534	7.672993038	0.00560534	0.98533	0.97508	0.99569
F.35-44	0.009285	0.00301	9.544692554	0.002005284	1.00933	1.0034	1.01529
F.45-54	0.000717	0.0061	0.013823541	0.906405558	1.00072	0.98882	1.01276
F.55-64	0.003533	0.01184	0.089089424	0.765337971	1.00354	0.98053	1.02709
F.65+	0.012149	0.03182	0.145813392	0.702568774	1.01222	0.95103	1.07735
M.13-17	0.160503	0.05065	10.03980357	0.001531935	1.1741	1.06313	1.29665
M.18-24	-0.027914	0.02711	1.060386683	0.303127008	0.97247	0.92215	1.02554
M.25-34	-0.017358	0.01124	2.385354131	0.122477135	0.98279	0.96138	1.00468
M.35-44	-0.020696	0.01218	2.886879353	0.08930375	0.97952	0.95641	1.00318
M.45-54	0.00901	0.008	1.266987485	0.260332587	1.00905	0.99334	1.02501
M.55-64	0.009267	0.01361	0.46391167	0.495801501	1.00931	0.98275	1.03659
M.65+	-0.05759	0.0267	4.653005901	0.030999219	0.94404	0.89591	0.99475
distance from mar	-0.001462	0.01402	0.010874844	0.916945149	0.99854	0.97147	1.02636
season_fall	0.02109	0.44612	0.002234956	0.962293801	1.02131	0.42601	2.44847

Residual DF	545
Residual De	698.944
# Iterations	2
Multiple R?	0.13661

- **Stepwise Variable Selection**

Variable Selection

					Model													
Subset Link	#Coeffi	RSS	Cp	Probability	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Choose Subset	1	684.3934	40.8495	0.0002	Intercept														
Choose Subset	2	676.0009	35.2118	0.0009	Intercept												last 2 time_ revenue		
Choose Subset	3	667.1667	29.1719	0.0045	Intercept			_ status									last 2 time_ revenue		
Choose Subset	4	660.7776	25.3574	0.0121	Intercept			_ status					#post share				last 2 time_ revenue		
Choose Subset	5	654.6124	21.7465	0.0297	Intercept			_ status					#post share				last 2 time_ revenue		

16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39
			Product category_ handicraft																				

Appendix B: Validation Summary Report

- ## Logistics Regression Validation Summary Report

Validation Data Scoring - Summary Report

Cutoff probability value for success (UPDATABLE)	0.5

Confusion Matrix		
	Predicted Class	
Actual Class	Yes	No
Yes	122	69
No	58	101

Error Report			
Class	# Cases	# Errors	% Error
Yes	191	69	36.12565445
No	159	58	36.47798742
Overall	350	127	36.28571429

Performance	
Success Class	Yes
Precision	0.67778
Recall (Sensitivity)	0.63874
Specificity	0.63522
F1-Score	0.65768

- ## Naive Bayes Validation Summary Report

Validation Data Scoring - Summary Report

Cutoff probability value for success (UPDATABLE)	0.5

Confusion Matrix		
	Predicted Class	
Actual Class	Yes	No
Yes	118	67
No	64	101

Error Report			
Class	# Cases	# Errors	% Error
Yes	185	67	36.2162
No	165	64	38.7879
Overall	350	131	37.4286

Performance	
Success Class	Yes
Precision	0.64835
Recall (Sensitivity)	0.63784
Specificity	0.61212
F1-Score	0.64305

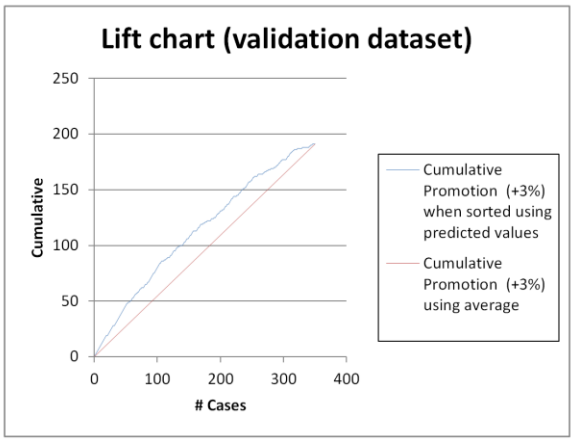
- ## Naive Bayes with X-Validation Performance Report

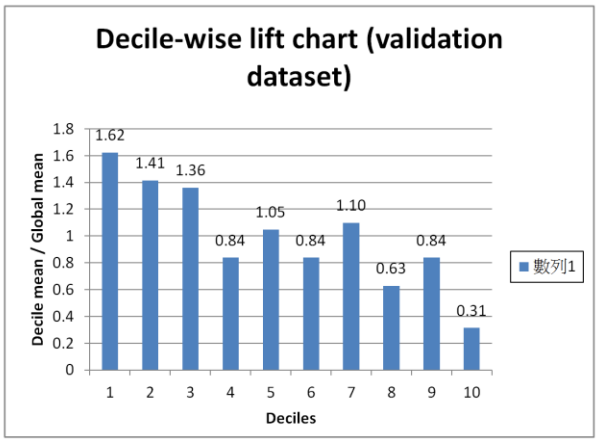
PerformanceVector

```
PerformanceVector:
accuracy: 66.44% +/- 4.10% (mikro: 66.44%)
ConfusionMatrix:
True:   No      Yes
No:     287     127
Yes:    265     489
precision: 64.93% +/- 3.55% (mikro: 64.85%) (positive class: Yes)
ConfusionMatrix:
True:   No      Yes
No:     287     127
Yes:    265     489
recall: 79.37% +/- 3.81% (mikro: 79.38%) (positive class: Yes)
ConfusionMatrix:
True:   No      Yes
No:     287     127
Yes:    265     489
AUC (optimistic): 0.711 +/- 0.044 (mikro: 0.711) (positive class: Yes)
AUC: 0.711 +/- 0.044 (mikro: 0.711) (positive class: Yes)
AUC (pessimistic): 0.711 +/- 0.044 (mikro: 0.711) (positive class: Yes)
```

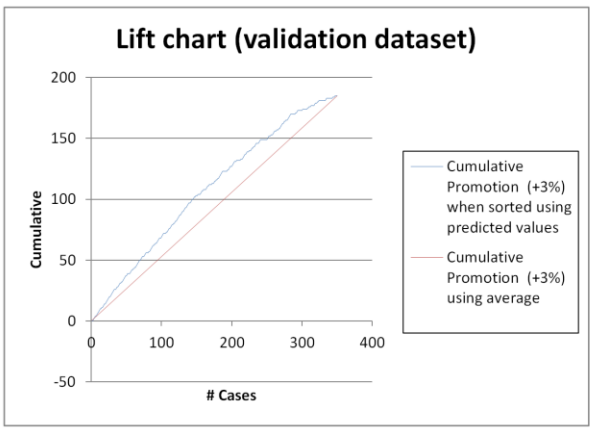
Appendix C: Lift Chart

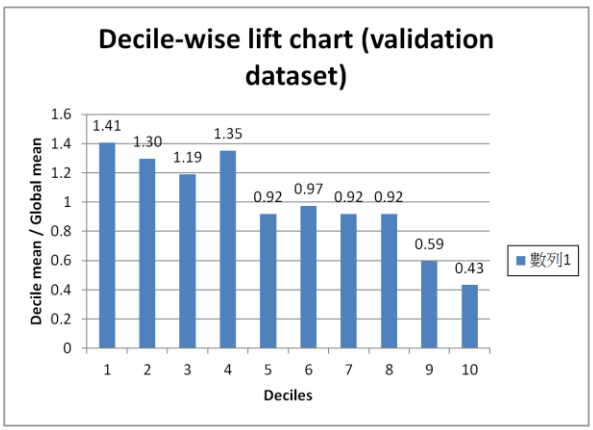
- **Logistics Regression Lift Chart**





- **Naive Bayes Lift Chart**





- **Naive Bayes with X-Validation**