Final Project
Ya-Hsuan Chuo

1. Research Topic:

   My research topic is based on a dataset of loan campaigns in a Portuguese banking institution. The marketing campaigns were based on phone calls. It recorded two waves of campaign, and also recorded the numbers of contact, and the outcome of both waves of campaign. In this dataset, I am looking for the factors that influence the subscription rate. Moreover, I suppose that the more contact with customers would increase or imply the higher rate of subscription.

2. Methodology

   Since the variables of the outcome are recorded with "yes" and "no" (yes=1, no=2), I would use logistic regression to identify the important factor. In the midterm project, I already recode all the variables into executable variables. For example, the outcome of 1st wave would be a dummy variable of yes=1, no=2. Setting this variable as a dummy would be a control to explain how the 1st wave outcome would influence the 2nd wave outcome.

   As other papers showed, the most important factors that affect the subscription would be the duration. Thus, my first model would be fitting the "y" with "duration". For the further steps, I added the variables that from the most relevant to the least. I took the results of other papers to determine which variables are important. In Enhancing Bank Direct Marketing through Data Mining (Sérgio Moro,2012) stated that Calls duration, Contact months, and the history of credits of a person are some critical values for predicting the influence of per customer buying loans. Moreover, the economic index also had more influence than personal background. Thus, in model 4, I added the economic index first, then in model 5 I added the personal information.

   In the model selection, I first see the significance of each variable. If it is significant in the regression, I will keep it remain in the regression and vice versa. I also compare the R square and the AIC and SBC. As the model being revised, the R square are getting larger and AIC, SBC getting smaller. Furthermore, the likelihood ration test should all variables are significant in the model 6. This shows they have stronger ability to explain this data.

		Model 1	Model 2	Model 3	Model 4	Model 5	Model 6
Important variables	Duration	0.00681*	0.00697*	0.00702*	0.00812*	0.00802*	0.0088*
	Campaign		-0.1786*	-0.1763*	-0.0724*	-0.1461*	-0.076*
	Previous		1.2062*	0.62568*	-0.249*	0.2695*	-0.214*
	poutcome[0]			-1.2662*	-0.32*	-0.5493*	-0.395*
	Pdays				-0.0012*	-0.0014*	-0.001*
Economic	emp.var.rate				-1.264*	-2.461*	-2.351*

index	cons.price.idx				0.754*	2.758*	2.8458*
	cons.conf.idx				0.004	---	---
	euribor3m				0.53	---	---
	nr.employed				-0.013*	0.0093*	0.0139*
Personal informtaion	Age					0.00527	---
	Job[][][]…					*	*
	Martial[1][2][3]					0.00822	---
	Education					*	*
	Housing					Biased	---
	Loan					Biased	---
	Month[][][]…					*	*
	R-square	0.0892	0.1882	0.2386	0.4481	0.3621	0.4929
	AIC	16576.6	14778.8	13864.6	10064.8	9295.7	9298.89
	SBC	16593.6	14812.8	13907.1	10158.2	9618.3	9604.58

3. Explanation

As the model 6 has a higher R square and lower AIC, SBC, I would choose this my final model and analyze with it.

Duration (+)

First, with a longer duration of calls, customers would have higher possibility to buy the loan. Since the coefficient is small, there are no really large difference between having longer duration of calls or not. That is to say, it doesn't really imply that longer duration with clients would attract them to buy.

Campaign (-)

This variable measures the numbers of contact within the campaign. The negative coefficient imply that more contact would not bring customer a positive perspective.

Previous(-), Poutcome[0](-), Pdays(-)

The "previous" variable measures the numbers of contact in the 1st wave campaign. It changes a lot in the changing of models. It is because this variable is related with poutcome. If the customer didn't buy the loan in the previous campaign (poutcome=0), he will have a higher potential not to buy the loan again. Moreover, holding that this client didn't buy the loan last time, more contact in the previous campaign could not bring more clients to buy the loan. The coefficient of Pdays is small, thus I don't think it has a good ability to explain.

Economic index (+)

A higher consumer price index and higher employment variation rate would motivate the sales of loan. The higher CPI could be an inflation in the economic, they may spend too much money and their deposit is low. When the economic is in depression that people can't find jobs, they also need money. Thus, the more fluctuate the economy, the more possibility to sale the loan.

Look at the job variables, the people who retires, and students are the target customers of the bank. They should focus on these groups to promote. People in the blue-collar or service sector are less likely to buy the loan, and the bank should consider not giving them to many contact else it would result in more pressure to them.

People who are well-educated are not easily affected by the telephone contact. While the results not really significant, where p-value is large, I think education didn't affect a lot.

The campaign started in March and end in November. It shows that the best promotion time would be the start and the end. People who really need a loan would buy it at the first phase when they receive this information. In the beginning of Jun or July, people who are considering at the first phase that whether to buy or not may forget about it. Thus, at the end of the loan selling, it is important that you contact them again and make sure they start to think about it.

4. Conclusion & Recommendation

My hypothesis in the beginning is assuming the longer or more frequently contact with customer will have positive effect to selling the loan. Because I think loan is a special product that customer must have full understanding before buying it. Thus more contact would help. However, after my analyses, this hypothesis doesn't hold. Business should aware that the number of contact should not be too many.

Conduct Cluster Analyses and customize their needs.

They should more focus on understanding each customer, such as their job, salaries. Moreover, understanding the economic situation is also important. They could further do more market research to understand other customers and attract new members. Also, they should do cluster analyses and study different group of their customers. After that, they can customize products to meet each group's goal. Last but not least, it is not a useful idea to contact everyone by phone. Some people who stay at home working or retiring may easily be contact by phone. And some group of people are really busy and don't have time to contact with phone, bank should also take this into consideration and contact them with other ways such as e-mails or face-to-face explanation.

Final Project
Ya-Hsuan Chuo

Appedix

Source	LogWorth	PValue
duration	315.650	0.00000
month	129.038	0.00000
emp.var.rate	46.116	0.00000
cons.price.idx	29.880	0.00000
job	9.724	0.00000
nr.employed	8.956	0.00000
previous	5.737	0.00000
pdays	5.059	0.00001
campaign	4.194	0.00006
education	3.266	0.00054
poutcome	3.063	0.00086

Whole Model Test

Model	-LogLikelihood	DF	ChiSquare	Prob>ChiSq
Difference	4484.5799	35	8969.16	<.0001*
Full	4613.4089			
Reduced	9097.9888			

RSquare (U)	0.4929
AICc	9298.89
BIC	9604.58
Observations (or Sum Wgts)	36081

Measure	Training	Definition
Entropy RSquare	0.4929	1-Loglike(model)/Loglike(0)
Generalized RSquare	0.5557	(1-(L(0)/L(model))^(2/n))/(1-L(0)^(2/n))
Mean -Log p	0.1279	$\sum$ -Log(ρ[j])/n
RMSE	0.1959	$\sqrt{\sum(y[j]-\rho[j])^2/n}$
Mean Abs Dev	0.0753	$\sum \lvert y[j]-\rho[j]\rvert/n$
Misclassification Rate	0.0528	$\sum (\rho[j]\neq\rho Max)/n$
N	36081	n

Final Project
Ya-Hsuan Chuo

Lack Of Fit

Source	DF	-LogLikelihood	ChiSquare
Lack Of Fit	33161	4552.5683	9105.137
Saturated	33196	60.8406	**Prob>Chiq**
Fitted	35	4613.4089	1.0000

Parameter Estimates

Term	Estimate	Std Error	ChiSquare	Prob>ChiSq
Intercept	-341.88824	34.356491	99.03	<.0001*
job[admin.]	0.04957359	0.0670712	0.55	0.4598
job**[blue-collar]**	-0.4533347	0.0969921	21.85	<.0001*
job[entrepreneur]	-0.1665221	0.1634069	1.04	0.3082
job[housemaid]	0.07571153	0.1764299	0.18	0.6678
job[management]	-0.088364	0.1033786	0.73	0.3927
job**[retired]**	0.39529648	0.0959118	16.99	<.0001*
job[self-employed]	0.00715512	0.140843	0.00	0.9595
job**[services]**	-0.3739358	0.1140053	10.76	0.0010*
job**[student]**	0.35799009	0.1106886	10.46	0.0012*
job[technician]	0.06495843	0.0838315	0.60	0.4384
job[unemployed]	0.23401652	0.1410556	2.75	0.0971
education**[basic.4y]**	-0.2967785	0.1424374	4.34	0.0372*
education[basic.6y]	-0.2810944	0.1731484	2.64	0.1045
education**[basic.9y]**	-0.4826077	0.1402793	11.84	0.0006*
education**[high.school]**	-0.2859788	0.126228	5.13	0.0235*
education**[illiterate]**	1.58184965	0.7497292	4.45	0.0349*
education[professional.course]	-0.1119223	0.135388	0.68	0.4084
education[university.degree]	-0.0191264	0.1235069	0.02	0.8769
month**[mar]**	2.04368063	0.1120549	332.63	<.0001*
month**[apr]**	-0.3862356	0.0777326	24.69	<.0001*
month**[may]**	-1.2263577	0.0731318	281.20	<.0001*
month**[jun]**	-1.2156868	0.1442545	71.02	<.0001*
month**[jul]**	-0.2591508	0.1048913	6.10	0.0135*
month**[aug]**	0.58411228	0.0888054	43.26	<.0001*
month**[sep]**	0.63035841	0.1332224	22.39	<.0001*
month**[oct]**	0.26809088	0.098205	7.45	0.0063*
month**[nov]**	-0.4702137	0.0904556	27.02	<.0001*
duration	0.00886207	0.0002519	1237.4	<.0001*

Final Project
Ya-Hsuan Chuo

Term	Estimate	Std Error	ChiSquare	Prob>ChiSq
campaign	-0.0764919	0.0199403	14.72	0.0001*
pdays	-0.0011182	0.0002482	20.30	<.0001*
previous	-0.2146275	0.0454383	22.31	<.0001*
poutcome[0]	-0.3952404	0.118885	11.05	0.0009*
emp.var.rate	-2.3515369	0.1633627	207.20	<.0001*
cons.price.idx	2.84580289	0.2450016	134.92	<.0001*
nr.employed	0.01395142	0.0022743	37.63	<.0001*

For log odds of 1/2

Effect Likelihood Ratio Tests

Source	Nparm	DF	L-R ChiSquare	Prob>ChiSq
job	11	11	69.004309	<.0001*
education	7	7	25.8227432	0.0005*
month	9	9	629.619325	<.0001*
duration	1	1	1445.89081	<.0001*
campaign	1	1	15.979313	<.0001*
pdays	1	1	19.7720951	<.0001*
previous	1	1	22.7619323	<.0001*
poutcome	1	1	11.0982838	0.0009*
emp.var.rate	1	1	206.578224	<.0001*
cons.price.idx	1	1	132.253255	<.0001*
nr.employed	1	1	37.127569	<.0001*

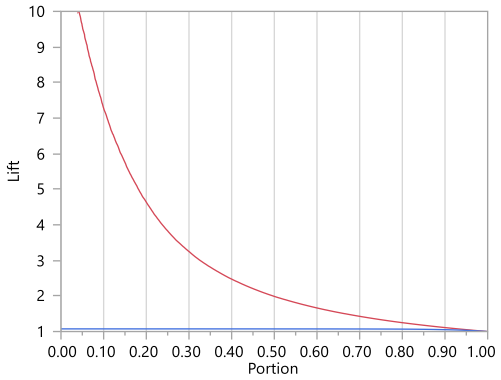
Confusion Matrix

Training

Actual	Predicted	
y	1	2
1	1136	1369
2	535	33041

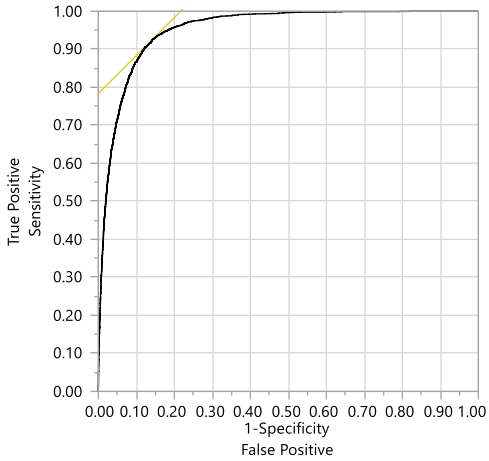
Final Project
Ya-Hsuan Chuo

Lift Curve



y
1
2

Receiver Operating Characteristic

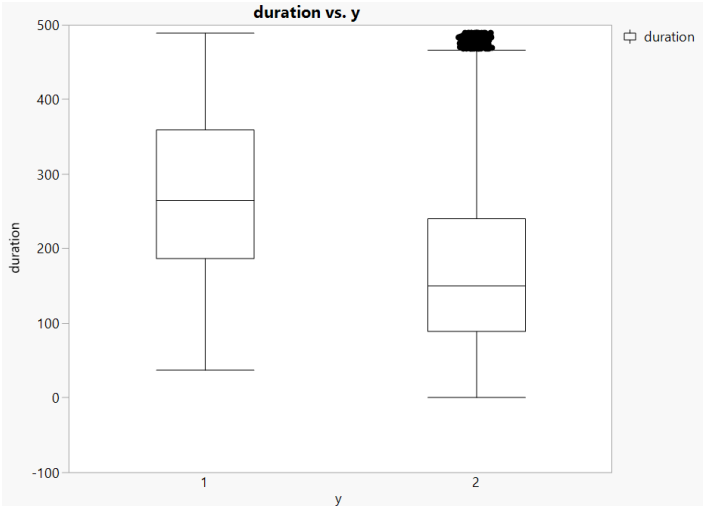


Using y='1' to be the positive level

AUC
0.95283

The ROC curve shows that this model is really good for explanation this dataset.

Final Project
Ya-Hsuan Chuo



The people who subscribe the loan (y=1) had a longer mean of duration on the phone call.

Tabulate

	1	2
y	2505	33576
Most Likely y	1671	34410

My prediction from my model shows that it is close to predict people who will not buy it, and a little bit worse on predicting people who are going to buy.