M3S16 Coursework Part 2

Yuting Lu

December 9, 2020

1 Introduction

In this report, I perform a more in depth investigation of the given data by including more variables and doing data preparation and validation on them. Then, I selected the variables that contained more information about the data and proceeded to construct an initial model. I have then tried to improve this initial model by including interaction terms and implementing a segmentation. Finally, I've tested the model and calculated AUC to see how the model fit perform.

2 Data Preparation and Validation

During the exploratory data analysis, I noticed that two variables have missing values. Below are the methods I choose to deal with this problem for the 2 variables:

- first.time.home.buyer: To see what type of missing values they are, I've created a data frame where for this variable, all the values are missing.

  Then I compared it to the distribution of the original data and found out that the distribution did not change so much.

  In addition, for this variable, there are 22.6% missing values which is not a small percentage, so list-wise deletion can not be used. Also, using regression would not be efficient. Hence the method of variable deletion will be used.

- PPM: At first, I used the same strategy as for the first.time.home.buyer variable. It is noticeable that there are more default than non-default in the data frame where only consumers with missing PPM values are considered.

  In addition, the mean and median of their score are lower than the average value.

  Looking at the loan purpose, I noticed that in the original data, the percentages of the variables are in the relation Purchase mortgage > No Cash-out Refinance mortgage > Cash-out Refinance mortgage, while in the new data frame, it is seen to be Purchase mortgage > Cash-out Refinance mortgage > No Cash-out Refinance mortgage. Thus more people want a cash-out refinance mortgage, and therefore, the use of the loan amount is not limited to specific purpose.

Moreover, comparing to the original data, this one has less first time home buyer(63.2% and 78.5% respectively), and higher LTV.

Now consider the possibility of them missing by definition. The reason why I think they could not be is because within the data frame in which people have no PPM value, only 98 of them are first time home buyer while 587 people do not have PPM value. Due to the reasons above, I think PPM is an important variable in this data set and moreover, has non-ignorable missing values.

In the end, because of the nature of this variable (0 Yes), I decided to make all the unknown values another category.

3 Variable Selection

3.1 Stepwise backward selection

This method starts from a model with all available variables and then gradually removes them in order to show how much relevance they have to the AIC of the model. The stop criteria is set by a lack of improvement in the model fit. The table below shows the process.

Step	**Variable removed**	**AIC**	**Variable included**
0	-	22814	{score, DTI, LTV, UPB(log), OIR, PPM, channel, occupancy.status, loan.purpose, property.state, property.type, orig.loan.term, orig.year, number.borrowers}
1	orig.loan.term	22812	{score, DTI, LTV, UPB(log), OIR, PPM, channel, occupancy.status, loan.purpose, property.state, property.type, orig.year, number.borrowers}
2	Removing any more variables does not decrease AIC		Stop

From this table, we conclude that the variable orig.loan.term should be excluded.

3.2 LASSO penalty

In addition, I included a LASSO penalty as part of the model fit optimization process. Here the penalty term will shrink the sizes of coefficient estimates. Below is the table showing the coefficient estimate with linear regression and linear regression with LASSO penalty.

With the aid of this table, we exclude the variables: occupancy.status and property.type.

Predictor variable	**Data Type**	**LR: coefficient estimate**	**LR: Pr(>|z|)**	**LR with LASSO coefficient estimate**
(intercept)		-9.1714023	<0.0001	7.27586
score	continuous	-0.0132426	<0.0001	0.01172
DTI	continuous	0.0359019	<0.0001	-0.030339
LTV	continuous	0.0513606	<0.0001	-0.03433
UPB(log)	continuous	0.5383418	<0.0001	-0.3601
OIR	continuous	0.9796303	<0.0001	-0.9209
PPM				
=Yes	0/1	-	-	0
=Unknown	0/1	1.0925389	<0.0001	-0.6925
=No	excluded			
Channel				
=Broker	0/1	-0.1747533	0.001338	0.0298
=Correspondent	0/1	0.2927722	<0.0001	-0.0165
=Unspecified	0/1	0.4611204	<0.0001	-0.3573
=Retail	excluded			
Occupancy status				
=Investment	0/1	0.0851199	0.155841	0
=Second home	0/1	0.4229274	<0.0001	0
=Owner occupier	excluded			
Loan purpose				
=Cash-out refinance	0/1	0.1481953	0.002467	-0.0686
=Purchase	0/1	-0.6042785	<0.0001	0.2773
=No cash-out	excluded			
Property state				
=TX	0/1	-1.1675499	<0.0001	<0.0001
=CA	excluded			
Property type				
=Condo	0/1	-0.1539554	0.015053	0
=Planned unit	0/1	-0.1861631	0.000227	0.0115
=Others	0/1	-0.5352374	0.200669	0
=Leasehold	excluded			
Original year				
=09	0/1	-0.2500496	<0.0001	0.1782
=08	0/1	excluded		
Number of borrowers				
=2	0/1	-0.8906232	<0.0001	0.6624
=1	excluded			

4 Model Structure

4.1 Interaction terms

By reasoning and building small models to test whether interaction terms are significant, I decided to include 3 interaction terms: DTI×OIR, OIR × UPB(log), DTI × loan.purpose.

4.1.1 Outcomes

Variable	Estimate	Standard Error	z	Pr(>|z|)
Intercept	-40.409702	4.784914	-8.445	<0.0001
score	0.012853	0.000443	29.016	<0.0001
DTI	-0.280599	0.018058	-15.539	<0.0001
LTV	-0.051043	0.001955	-26.115	<0.0001
UPB(log)	4.381906	0.385317	11.372	<0.0001
OIR	6.947635	0.768838	9.037	<0.0001
channel.Broker	0.132497	0.067228	1.971	0.0487
channel.Correspondent	-0.281770	0.058426	-4.823	<0.0001
channel.TPO	-0.461638	0.058321	-7.915	<0.0001
PPM.Unknown	-1.104628	0.135449	-8.155	<0.0001
loan.purpose.C	0.152800	0.229091	0.667	0.5048
loan.purpose.P	1.769466	0.224979	7.865	<0.0001
number.borrowers	0.812944	0.045062	18.041	<0.0001
orig.year09	0.118694	0.073719	1.610	0.1074
property.state.TX	1.168444	0.058944	19.823	<0.0001
DTI×OIR	0.042224	0.002964	14.247	<0.0001
UPB(log)×OIR	-0.794519	0.062302	-12.753	<0.0001
DTI×loan.purpose.Cashout	-0.006115	0.005069	-1.206	0.2277
DTI×loan.purpose.Purchase	-0.024379	0.004908	-4.967	<0.0001

"AUC: 0.910432765718511"

4.1.2 Interpretation

- After testing on the remaining data, the AUC given by this model is "0.9104" while the AUC from the model in the last report was "0.8819". In addition, I calculated the AUC of a model without interaction terms but with the categorical variables that were added to this model, the result is "0.9026". This shows that by including the interaction terms, the model fit does improve a bit.
- DTI×OIR: The positive value of the estimate shows a positive association between this product and creditworthiness but also a negative association with default. In addition, the effect of DTI on default is different for different values of OIR. The marginal effect of DTI is: $-0.28 + 0.04\times$OIR.

- UPB(log)×OIR: The negative estimate of the product indicates a negative association with creditworthiness, and a positive association with the default. As before, the effect of UPB on the default variable is different for different value of OIR. The marginal effect of UPB is: $4.38 - 0.79\times$OIR.
- DTI×loan.purpose.Cashout: This variable is not considered significant here.
- DTI×loan.purpose.Purchase: Having obtained a negative estimate for the value of this parameter we conclude that there is a negative association between product and creditworthiness, and a positive association with default. The marginal effect of having a DTI is: $-0.28 - 0.02\times$ loan.purpose.P. That is, -0.3 if the loan purpose is to purchase the mortgage, -0.28 if the purpose is not to purchase the mortgage.

4.2 Segmentation

To implement segmentation, I first plotted a decision tree to see how the variables are divided.

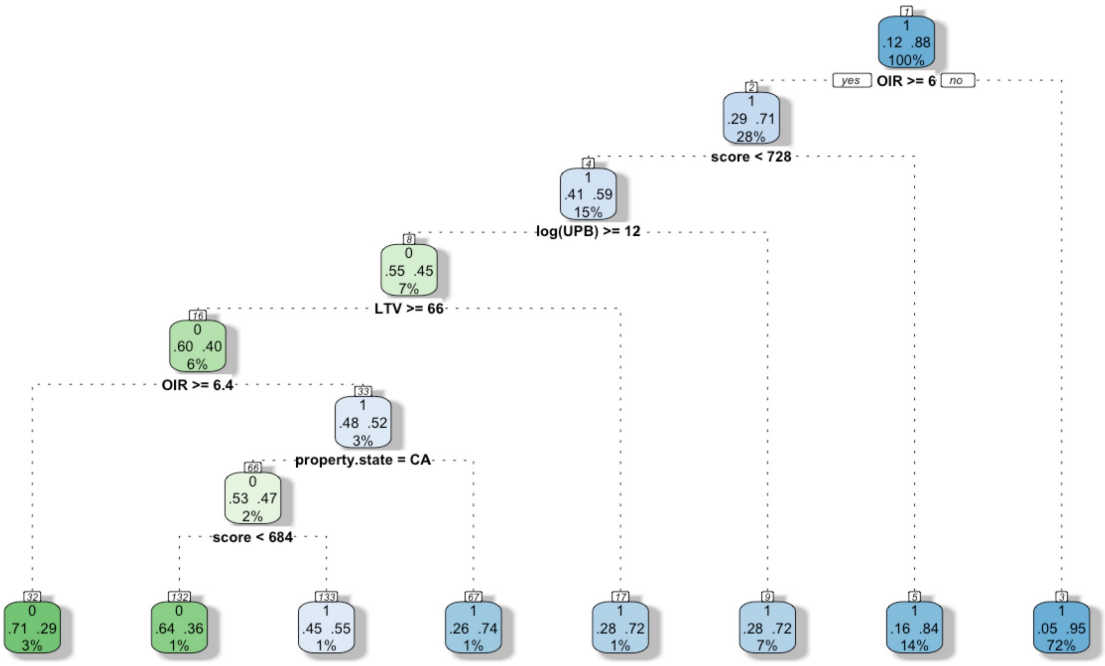

Figure 1: Decision trees for the selected data set

It can be seen that the first node OIR is divided into two parts : OIR$<$6 and OIR$\geq$6. Thus, I separated the data into two sets depends on this. Then I performed logistic regression on the two different training sets to build two models and then combine them to do a prediction in the combined model.

4.2.1 Outcome

	OIR<6		OIR>=6	
Variableas	**Estimated Coefficient**	**p-value**	**Estimated Coefficient**	**p-value**
Intercept	31.06	0.015540	11.117537	0.2744
score	0.01243	<0.0001	0.012325	<0.0001
DTI	-0.6174	<0.0001	-0.047411	0.2233
LTV	-0.04449	<0.0001	-0.056182	<0.0001
UPB(log)	5.007	<0.0001	-0.640222	0.4439
OIR	5.397	0.021289	-0.813751	0.5938
Channel.Broker	0.320	0.003032	0.062650	0.4678
Channel.Correspond	-0.2945	0.000854	-0.164880	0.0334
Channel.TPO	-0.2497	0.010295	-0.477701	<0.0001
PPM.Unknown	-0.9256	0.253640	-1.052860	<0.0001
loan.purposeC	0.3492	0.302082	0.026472	0.9316
loan.purposeP	2.432	<0.0001	1.287499	<0.0001
number.borrowers2	0.604	<0.0001	1.068635	<0.0001
orig.year09	0.003	0.971639	-0.060094	0.6857
property.stateTX	0.9783	<0.0001	1.210865	<0.0001
DTI × OIR	0.1031	<0.0001	0.005542	0.3514
UPB(log) × OIR	-0.9195	<0.0001	-0.024852	0.8434
DTI×loan.purposeC	-0.01072	0.145769	-0.002444	0.7248
DTI×loan.purposeP	-0.04012	<0.0001	-0.012140	0.0616

"AUC: 0.913574349174494"

4.2.2 Interpretation

- To begin with, the AUC of this model is a little higher than the last one. Which shows that the model fit is improved by performing segmentation.
- The association of OIR is much higher for the group where OIR<6.
- The first difference of association in direction is between OIR and UPB, when OIR $\geq$ 6, it starts to have a negative coefficient. This means a negative association with creditworthiness, and a positive association with default. So the higher OIR, the easier it is for people to default which is the same situation as for UPB.
- The second difference of association in direction is orig.year09: it is only a risk factor for people with OIR $\geq$ 6.
- For people with OIR<6, the risk of having default if their PPM value missed is lower than for people with OIR $\geq$ 6.
- For people with OIR $\geq$ 6, if they use Third Party Organization not specified, they are easier to default the loan.

5 Conclusion

To ensure testing is performed correctly, I separated my data into training set and testing set in the ratio of 2:1. In addition, I disrupt the order of the data to make it more random.

In conclusion, the last model got the highest AUC. It shows that the methods I used in this report did work. In addition, the segmented models provide some business information for people to understand how default is related to two groups of people with different OIR.