FIN3210 Week 2 Assignment Report
Ma Kexuan 120090651

Abstract

This report provides a descriptive summary statistic for the dataset given, and construct several regressions to discover how borrower characteristics affect the outcome of default likelihood, the number of bids. Furthermore, we discover the relation between the platform default likelihood and platform characteristics.

Data Preprocessing

The preprocessing procedures and some interpretations of the code are described in each code blocks in the appendix, please check.

Questions

1) Present two tables for the summary statistics of the key variables in Renrendai loans.xlsx and p2p lending platforms.xlsx

In the table of Renrendai loans.xlsx, I choose 'BIDS', 'DEFAULT', 'AMOUNT', 'INTEREST', 'MONTHS','CREDIT', 'HOUSE', 'CAR', 'HOUSE_L', 'CAR_L', 'EDUCATION', 'WORKTIME', 'INCOME', 'AGE' as the key variables, the summary statistics is described as the left chart below. It's worth noting that the mean value of default is 0.15, meaning that about 85% percent of people have been rejected from loaning the money. In the table of p2p lending platforms.xlsx, I choose 'OnlineTime_YMD', 'Bankrupt_WDZJ', 'Collapse','Benign', 'Fraud', 'RegCapital', 'Capitaldeposit', 'Obtaininvest', 'Joinasso', 'Autobid', 'Transright','Riskdeposit','Thirdguarantee' as the key variables, the summary statistics is described as the right chart below. The Collapse variable has a mean of 78%, meaning that 78% percent of the platforms have already collapsed, indicating a high risk of default.

	count	mean	std	min	25%	50%	75%	max
BIDS	10000.0	24.150600	41.342608	1.0	9.0	15.0	24.0	592.0
DEFAULT	10000.0	0.151300	0.358359	0.0	0.0	0.0	0.0	1.0
AMOUNT	10000.0	24545.835000	38280.756524	3000.0	8000.0	14400.0	26000.0	500000.0
INTEREST	10000.0	12.621900	2.273689	5.0	11.0	12.0	13.0	24.4
MONTHS	10000.0	12.237300	8.091090	3.0	6.0	12.0	12.0	36.0
CREDIT	10000.0	2.146300	1.530990	1.0	1.0	2.0	3.0	7.0
HOUSE	10000.0	0.564500	0.495847	0.0	0.0	1.0	1.0	1.0
CAR	10000.0	0.391700	0.488155	0.0	0.0	0.0	1.0	1.0
HOUSE_L	10000.0	0.228400	0.419823	0.0	0.0	0.0	0.0	1.0
CAR_L	10000.0	0.082200	0.274683	0.0	0.0	0.0	0.0	1.0
EDUCATION	9996.0	2.165966	0.818108	1.0	2.0	2.0	3.0	4.0
WORKTIME	9994.0	2.838003	0.992755	1.0	2.0	3.0	4.0	4.0
INCOME	9998.0	4.309162	1.335842	1.0	3.0	4.0	5.0	7.0
AGE	10000.0	34.755500	6.682708	24.0	30.0	33.0	38.0	53.0

	count	mean	std	min	25%	50%	75%	max
Collapse	1000.0	0.782000	0.413094	0.0	1.0	1.0	1.0	1.0
Benign	782.0	0.098465	0.298134	0.0	0.0	0.0	0.0	1.0
Fraud	782.0	0.246803	0.431427	0.0	0.0	0.0	0.0	1.0
RegCapital	1000.0	596.064330	2328.221711	2.0	100.0	300.0	500.0	50000.0
Capitaldeposit	1000.0	0.191000	0.393286	0.0	0.0	0.0	0.0	1.0
Obtaininvest	968.0	0.026860	0.161756	0.0	0.0	0.0	0.0	1.0
Joinasso	968.0	0.054752	0.227613	0.0	0.0	0.0	0.0	1.0
Autobid	1000.0	0.244000	0.429708	0.0	0.0	0.0	0.0	1.0
Transright	1000.0	0.177000	0.381860	0.0	0.0	0.0	0.0	1.0
Riskdeposit	968.0	0.021694	0.145758	0.0	0.0	0.0	0.0	1.0
Thirdguarantee	968.0	0.034091	0.181557	0.0	0.0	0.0	0.0	1.0

2) Perform a logit regression and examine the relation between the default likelihood and borrower characteristics such as credit, house, car, education, work time, etc.

In this Logit regression, I use 'CREDIT', 'HOUSE', 'CAR', 'HOUSE_L', 'CAR_L', 'EDUCATION', 'WORKTIME', 'INCOME', 'AGE' as the independent variables, the result at the left below shows that all the variables chosen except for CAR_L and WORKTIME have 99% significance level. Among the significant coefficients, CREDIT, CAR, HOUSE_L and

EDUCATION are negatively correlated with the dependent variable default. This can show that the platform has an accurate credit rating for users, and the higher the credit, the lower the default risk. Users with cars are likely to have good living conditions and a low probability of default, just as those with mortgages are. Higher education may mean higher quality and lower probability of default. However, for the positive ones, the larger the AGE, the more likelihood to default, since they may not earn enough money. But for the INCOME, it's quite weird to get the result, to explain it, maybe we should do more research to figure out the logic underneath.

Dep. Variable:	DEFAULT	No. Observations:	9990
Model:	Logit	Df Residuals:	9980
Method:	MLE	Df Model:	9
Date:	Thu, 28 Sep 2023	Pseudo R-squ.:	0.2236
Time:	00:01:17	Log-Likelihood:	-3298.0
converged:	True	LL-Null:	-4247.9
Covariance Type:	nonrobust	LLR p-value:	0.000

	coef	std err	z	P>|z|	[0.025	0.975]
const	0.5155	0.212	2.427	0.015	0.099	0.932
CREDIT	-1.8927	0.082	-23.044	0.000	-2.054	-1.732
HOUSE	0.1438	0.073	1.968	0.049	0.001	0.287
CAR	-0.4586	0.080	-5.708	0.000	-0.616	-0.301
HOUSE_L	-0.3307	0.091	-3.633	0.000	-0.509	-0.152
CAR_L	0.1620	0.134	1.207	0.228	-0.101	0.425
EDUCATION	-0.4156	0.040	-10.426	0.000	-0.494	-0.337
WORKTIME	0.0090	0.034	0.264	0.792	-0.058	0.076
INCOME	0.1160	0.025	4.592	0.000	0.066	0.165
AGE	0.0254	0.005	4.936	0.000	0.015	0.036

Dep. Variable:	BIDS	R-squared:	0.173
Model:	OLS	Adj. R-squared:	0.172
Method:	Least Squares	F-statistic:	232.1
Date:	Thu, 28 Sep 2023	Prob (F-statistic):	0.00
Time:	00:01:17	Log-Likelihood:	-50383.
No. Observations:	9990	AIC:	1.008e+05
Df Residuals:	9980	BIC:	1.009e+05
Df Model:	9		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-50.8110	2.479	-20.497	0.000	-55.670	-45.952
CREDIT	1.8652	0.257	7.248	0.000	1.361	2.370
HOUSE	1.6099	0.926	1.738	0.082	-0.206	3.426
CAR	4.2582	0.918	4.637	0.000	2.458	6.059
HOUSE_L	-7.1289	1.030	-6.924	0.000	-9.147	-5.111
CAR_L	-7.1951	1.482	-4.854	0.000	-10.101	-4.290
EDUCATION	-2.0042	0.475	-4.218	0.000	-2.936	-1.073
WORKTIME	2.4355	0.426	5.721	0.000	1.601	3.270
INCOME	9.2260	0.308	29.918	0.000	8.622	9.831
AGE	0.8126	0.066	12.235	0.000	0.682	0.943

Omnibus:	11602.380	Durbin-Watson:	1.743
Prob(Omnibus):	0.000	Jarque-Bera (JB):	1282780.294
Skew:	6.139	Prob(JB):	0.00
Kurtosis:	57.139	Cond. No.	239.

3) Perform an ols regression and examine the relation between the number of bids and borrower characteristics such as credit, house, car, education, work time, etc.

The OLS result is shown on the right above. By observing the p-value of all the independent variables, we find that they are all 99% significant except for HOUSE. For HOUSE_L and CAR_L, since there's a loan on them, there may not be many investors to give money to them. For CREDIT, HOUSE, CAR, WORKTIME, INCOME, AGE, these characteristics describe the social status for certain person, thus if they're larger, it implies that there should be larger probability for them to give the money back, hence more BIDS for them. Nevertheless, the negative relation between BIDS and EDUCATION is quite elaborate. We may need further research to figure out the abnormal phenomenon.

model	lifelines.CoxPHFitter
duration col	'deltatime'
event col	'Collapse'
baseline estimation	breslow
number of observations	774
number of events observed	774
partial log-likelihood	-4321.49
time fit was run	2023-09-27 16:25:06 UTC

Concordance	0.62
Partial AIC	8658.98
log-likelihood ratio test	114.14 on 8 df
-log2(p) of ll-ratio test	67.34

	coef	exp(coef)	se(coef)	coef lower 95%	coef upper 95%	exp(coef) lower 95%	exp(coef) upper 95%	cmp to	z	p	-log2(p)
RegCapital	0.00	1.00	0.00	-0.00	0.00	1.00	1.00	0.00	0.03	0.98	0.03
Joinasso	-0.59	0.56	0.22	-1.03	-0.15	0.36	0.86	0.00	-2.62	0.01	6.81
Autobid	-0.24	0.79	0.09	-0.41	-0.06	0.66	0.94	0.00	-2.61	0.01	6.80
Capitaldeposit	-0.71	0.49	0.14	-0.99	-0.44	0.37	0.65	0.00	-5.05	<0.005	21.07
Obtaininvest	-0.35	0.71	0.27	-0.88	0.19	0.42	1.21	0.00	-1.27	0.20	2.29
Transright	-0.37	0.69	0.11	-0.59	-0.16	0.56	0.86	0.00	-3.37	<0.005	10.40
Riskdeposit	-0.14	0.87	0.27	-0.67	0.38	0.51	1.46	0.00	-0.54	0.59	0.76
Thirdguarantee	-0.06	0.94	0.23	-0.51	0.39	0.60	1.47	0.00	-0.28	0.78	0.36

4) Perform the Cox model (Proportional hazards model) and examine the relation between the platform default (survival) likelihood and platform characteristics such as RegCapital, Joinasso, etc.

RegCapital has slight effect on the likelihood of collapse, the rest of the variables except Riskdeposit and Thirdguarantee have negative relationship with the likelihood of collapse, and the exp(coef) are all smaller than 0.8, indicating that there's a significant impact. For Riskdeposit and Thirdguarantee, they have little impact on the likelihood. The concordance value is 0.62, indicating the model's ability to distinguish between different survival times. The p-value for the log-likelihood ratio test is very close to zero, reflecting the strong statistical significance of the model.

Below is the appendix for the code: