MDS6212 Fintech Theory and Practice
Assignment 2

Yihang Li
220041006

1 Question1

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

1.1 For Renrendai loans.xlsx

Table 1 is the summary statistics of the key variables in Renrendai loans.xlsx.

Table 1: Summary Statistics of the Key Variables in Renrendai loans.xlsx

	loanId	*BIDS*	*DEFAULT*	*AMOUNT*	*INTEREST*	*MONTHS*	*CREDIT*	*HOUSE*	*CAR*	*HOUSE_L*	*CAR_L*	*EDUCATION*	*WORKTIME*	*INCOME*	*AGE*
count	10000.000000	10000.000000	10000.000000	10000.000000	10000.000000	10000.000000	10000.000000	10000.000000	10000.000000	10000.000000	10000.000000	9996.000000	9994.000000	9998.000000	10000.000000
mean	418846.947200	24.150600	0.151300	24545.835000	12.621900	12.237300	2.146300	0.564500	0.391700	0.228400	0.082200	2.165966	2.838003	4.309162	34.755500
std	446432.580191	41.342608	0.358359	38280.756524	2.273689	8.091090	1.530990	0.495847	0.488155	0.419823	0.274683	0.818108	0.992755	1.335842	6.682708
min	2.000000	1.000000	0.000000	3000.000000	5.000000	3.000000	1.000000	0.000000	0.000000	0.000000	0.000000	1.000000	1.000000	1.000000	24.000000
25%	84635.250000	9.000000	0.000000	8000.000000	11.000000	6.000000	1.000000	0.000000	0.000000	0.000000	0.000000	2.000000	2.000000	3.000000	30.000000
50%	321945.000000	15.000000	0.000000	14400.000000	12.000000	12.000000	2.000000	1.000000	0.000000	0.000000	0.000000	2.000000	3.000000	4.000000	33.000000
75%	582930.500000	24.000000	0.000000	26000.000000	13.000000	12.000000	3.000000	1.000000	1.000000	0.000000	0.000000	3.000000	4.000000	5.000000	38.000000
max	2086049.000000	592.000000	1.000000	500000.000000	24.400000	36.000000	7.000000	1.000000	1.000000	1.000000	1.000000	4.000000	4.000000	7.000000	53.000000

1.2 For p2p lending platforms.xlsx

Table 2 is the summary statistics of the key variables in p2p lending platforms.xlsx.

Table 2: Summary Statistics of the Key Variables in p2p lending platforms.xlsx

	OnlineTime_YMD	*Bankrupt_WDZJ*	*Collapse*	*Benign*	*Fraud*	*RegCapital*	*Capitaldeposit*	*Obtaininvest*	*Joinasso*	*Autobid*	*Transright*	*Riskdeposit*	*Thirdguarantee*
count	1000.000000	782.000000	1000.000000	782.000000	782.000000	1000.000000	1000.000000	968.000000	968.000000	1000.000000	1000.000000	968.000000	968.000000
mean	20148498.958000	20163303.048593	0.782000	0.098465	0.246803	596.064330	0.191000	0.026860	0.054752	0.244000	0.177000	0.021694	0.034091
std	11351.943281	13043.220563	0.413094	0.298134	0.431427	2328.221711	0.393286	0.161756	0.227613	0.429708	0.381860	0.145758	0.181557
min	20090409.000000	20120601.000000	0.000000	0.000000	0.000000	2.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
25%	20140917.750000	20151117.000000	1.000000	0.000000	0.000000	100.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
50%	20150401.000000	20160801.500000	1.000000	0.000000	0.000000	300.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
75%	20151116.250000	20171107.000000	1.000000	0.000000	0.000000	500.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
max	20180524.000000	20190904.000000	1.000000	1.000000	1.000000	50000.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000

Logit Regression Results

Dep. Variable:	DEFAULT	**No. Observations:**	9990
Model:	Logit	**Df Residuals:**	9980
Method:	MLE	**Df Model:**	9
Date:	Sun, 04 Oct 2020	**Pseudo R-squ.:**	0.2236
Time:	12:30:47	**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	-3.1979	0.091	-35.272	0.000	-3.376	-3.020
x1	-2.8978	0.126	-23.044	0.000	-3.144	-2.651
x2	0.0713	0.036	1.968	0.049	0.000	0.142
x3	-0.2239	0.039	-5.708	0.000	-0.301	-0.147
x4	-0.3401	0.033	-10.426	0.000	-0.404	-0.276
x5	0.0090	0.034	0.264	0.792	-0.058	0.076
x6	-0.1389	0.038	-3.633	0.000	-0.214	-0.064
x7	0.0445	0.037	1.207	0.228	-0.028	0.117
x8	0.1549	0.034	4.592	0.000	0.089	0.221
x9	0.1700	0.034	4.936	0.000	0.102	0.237

(a) Result of Logit Regression

Logit Regression Results

Dep. Variable:	DEFAULT	**No. Observations:**	9990
Model:	Logit	**Df Residuals:**	9982
Method:	MLE	**Df Model:**	7
Date:	Sun, 04 Oct 2020	**Pseudo R-squ.:**	0.2234
Time:	12:30:47	**Log-Likelihood:**	-3298.8
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	-3.1951	0.091	-35.288	0.000	-3.373	-3.018
x1	-2.8951	0.126	-23.050	0.000	-3.141	-2.649
x2	0.0725	0.036	2.014	0.044	0.002	0.143
x3	-0.2053	0.036	-5.707	0.000	-0.276	-0.135
x4	-0.3400	0.033	-10.451	0.000	-0.404	-0.276
x5	-0.1364	0.038	-3.574	0.000	-0.211	-0.062
x6	0.1546	0.034	4.598	0.000	0.089	0.220
x7	0.1710	0.032	5.328	0.000	0.108	0.234

(b) Result of New Logit Regression

Figure 1: Result of Logit

2 Question2

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

2.1 Dealing missing values

As wee see in Table 1, EDUCATION, WORKTIME and INCOME have slightly degree of missing, here we just delete these missing values. Now we have 9990 observations in total.

2.2 Perform logit regression

Now let's perform logit regression by using statsmodels. And Figure 1a is the result.

```
import statsmodels.api as sm
scaler = StandardScaler()
q2_X = scaler.fit(q2_X).transform(q2_X)
q2_X = sm.add_constant(q2_X)
q2_model = sm.Logit(q2_y, q2_X)
q2_result = q2_model.fit()
```

2.3 Rebuild Model by removing non-significant variables

We can see from the result summary, x5('WORKTIME') has much high p-value(0.792), which means it is not significant, so as x7('CAR_L'). Thus we may rebuild our model by removing them. Figure 1b is the result of new Logit Regression. We can see now all variables are significant.

3 Question3

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.

3.1 Perform OLS

Here we perform OLS by using the following Python Code. And Figure 2 is the result.

```
q3_y = q3_Data[['BIDS']]
q3_X = q3_Data[[
    'CREDIT', 'HOUSE', 'CAR', 'EDUCATION', 'WORKTIME', 'HOUSE_L', 'CAR_L',
    'INCOME', 'AGE']]
q3_X = scaler.fit(q3_X).transform(q3_X)
q3_X = sm.add_constant(q3_X)
q3_model = sm.OLS(q3_y, q3_X)
q3_result = q3_model.fit()
```

4 Question4

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.

4.1 Coxs proportional hazard model

The idea behind Coxs proportional hazard model is that the log-hazard of an individual is a linear function of their covariates and a population-level baseline hazard that changes over time. Mathematically:

$$\underbrace{h(t \mid x)}_{\text{hazard}} = \overbrace{b_0(t)}^{\text{baseline hazard}} \quad \underbrace{\exp\overbrace{\left(\sum_{i=1}^{n} b_i\,(x_i - \overline{x_i})\right)}^{\text{log-partial hazard}}}_{\text{partial hazard}}$$

Note a few behaviors about this model: the only time component is in the baseline hazard, $b_0(t)$. In the above equation, the partial hazard is a time-invariant scalar factor that only increases or decreases the baseline hazard. Thus changes in covariates will only inflate or deflate the baseline hazard.

Dep. Variable:	BIDS	**R-squared:**	0.173
Model:	OLS	**Adj. R-squared:**	0.172
Method:	Least Squares	**F-statistic:**	232.1
Date:	Sun, 04 Oct 2020	**Prob (F-statistic):**	0.00
Time:	12:30:47	**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	24.1163	0.375	64.242	0.000	23.380	24.852
x1	2.8557	0.394	7.248	0.000	2.083	3.628
x2	0.7982	0.459	1.738	0.082	-0.102	1.698
x3	2.0788	0.448	4.637	0.000	1.200	2.958
x4	-1.6398	0.389	-4.218	0.000	-2.402	-0.878
x5	2.4175	0.423	5.721	0.000	1.589	3.246
x6	-2.9933	0.432	-6.924	0.000	-3.841	-2.146
x7	-1.9772	0.407	-4.854	0.000	-2.776	-1.179
x8	12.3210	0.412	29.918	0.000	11.514	13.128
x9	5.4282	0.444	12.235	0.000	4.558	6.298

Figure 2: Result of OLS

Table 3: Result of the Cox Model

	coef	*exp(coef)*	*se(coef)*	*coef lower 95%*	*coef upper 95%*	*exp(coef) lower 95%*	*exp(coef) upper 95%*	*z*	*p*	*-log2(p)*
RegCapital	0.00	1.00	0.00	-0.00	0.00	1.00	1.00	0.11	0.91	0.13
Joinasso	-0.63	0.53	0.23	-1.08	-0.19	0.34	0.83	-2.79	0.01	7.55
Capitaldeposit	-1.35	0.26	0.14	-1.62	-1.08	0.20	0.34	-9.94	<0.005	74.89
Obtaininvest	-0.14	0.87	0.27	-0.66	0.39	0.52	1.48	-0.51	0.61	0.71
Autobid	-0.19	0.83	0.09	-0.36	-0.01	0.70	0.99	-2.07	0.04	4.70
Transright	-0.55	0.58	0.11	-0.76	-0.34	0.47	0.71	-5.11	<0.005	21.54
Riskdeposit	-0.13	0.88	0.27	-0.65	0.40	0.52	1.48	-0.48	0.63	0.66
Thirdguarantee	-0.20	0.82	0.23	-0.65	0.24	0.52	1.27	-0.90	0.37	1.44

4.2 Calculate Duration

We need to calculate 'Duration' by subtracting 'OnlineTime_YMD' from 'Bankrupt_WDZJ'. Since not every platform bankrupt in this dataset and in order to calculate the duration, we may need to consider using today's date to filling all the missing values in the 'Bankrupt_WDZJ'

4.3 Perform the Cox model (Proportional hazards model)

By using the following Python Codes, we performed the Cox model. The result is as in Table 3. we can see here, most of the covariates are not significant. And curiously, the coef of 'RegCapital' is 0, then what does this mean?(Will Explain Below)

```
from lifelines import CoxPHFitter
cph = CoxPHFitter()
cph.fit(q4_Data, duration_col='Duration', event_col='Collapse')
cph.print_summary()
```

4.4 Plot the effect of varying a covariate

After fitting, we can plot what the survival curves look like as we vary a single covariate while holding everything else equal. This is useful to understand the impact of a covariate, given the model. To do this, we use the plot_partial_effects_on_outcome() method and give it the covariate of interest, and the values to display.

See from Figure 3, join an association(Joinasso=1, the red line) can survive longer, and may be a possible explanation for the coef of RegCapital being 0 is that there's no survival difference among different values of RegCapital, for the 'Capitaldeposit', there is a huge difference and we can analyse each of them by the same way.

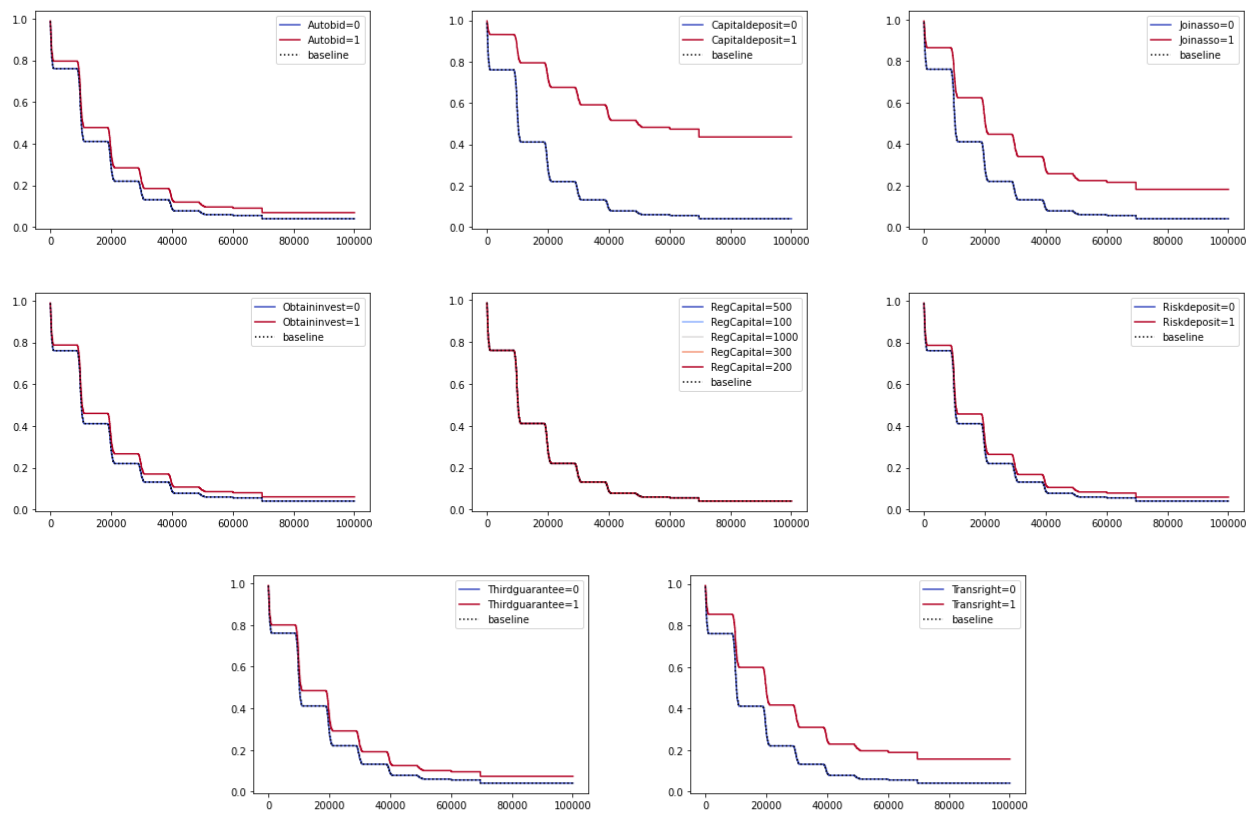

Figure 3: Kaplan-Meier Estimate