MDS6212 Fintech Theory and Practice
Assignment 1

Yihang Li
220041006

1 Qustion1

Present a table of summary statistics for the key variables including the borrowers age, gender, loan amount, interest rate, credit scores, a dummy whether the borrower has a frequent contact, approval dummy, and delinquency dummy.

1.1 Description for Numeric Variables

Table 1 is the summary statistics for the key numeric variables.

1.2 Description for Categorical Variables

Table 2 is the summary statistics for the key categorical variables(The number of True or False for each variable).

Table 1: Summary Statistics for Numerica Variables

	age	*instalments_amount*	*nominalrates*	*creditlevelasbuyer*	*tencentscore*	*gaodescore*
count	5000.000000	5000.000000	4997.000000	4031.000000	5000.000000	5000.000000
mean	27.675400	406201.420000	0.276058	53.119077	58.608168	0.201975
std	8.326146	130623.360240	0.085912	108.629757	14.218112	0.076724
min	18.000000	50000.000000	0.130080	0.000000	9.000000	0.023518
25%	21.000000	320000.000000	0.204560	0.000000	53.888889	0.192094
50%	25.000000	398000.000000	0.204579	14.000000	60.200000	0.192094
75%	32.000000	498000.000000	0.359347	58.000000	65.258929	0.192094
max	56.000000	869000.000000	0.494185	1830.000000	98.000000	0.732120

Table 2: Summary Statistics for Categorical Variables

	gender	*highcontact*	*deal*	*default*
False	4267	2539	2793	1280
True	733	2461	2207	925

2 Deal with Missing Values

As we see in Table 1 and Table 2, there are some missing values: 'nominalrates': 3, 'creditlevelasbuyer': 969, 'default': 2795. SO, before performing logit regression, we need to deal with these missing values.

For nominalrates, since there are only 3 missing values, we simply delete the row of those three. For convinient, we also delete the rows where creditlevelasbuyer is missing, because its missing values is not so much. By doing so, we can focus on dealing with the huge propotion missing values of default

Now we have 4029 observations in total. The idea to deal with 'default' is: Consider 'default' vs other variables in the Key_Data, use non-missing data to build a logistic classifier, and then use such classifier to predict the value of missing default.

By using Python, we successfully accomplished this process.

3 Question2

Perform a logit regression and examine the relation between the delinquency likelihood and credit scores

3.1 Answer

By using the following Python codes, we performed a logit regression between the delinquency likelihood and credit scores.

```python
import statsmodels.api as sm
q2_X = Key_Data[['creditlevelasbuyer', 'tencentscore', 'gaodescore']]
q2_y = Key_Data['default'].map({True:1, False:0})
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()
```

Figure 1 is the result of this Logit Regression. We can see that all the p-values are significant, and all the coefficients are positive, which means delinquency likelihood and these three credit scores have positive relationships.

Besides, We also performed logit regression on each variable separately, as a result, for each model, we got the coefficients: Model1('creditlevelasbuyer': 0.0294 with p-value: 0.388), Model2('tencentscore':

Logit Regression Results

Dep. Variable:	default	**No. Observations:**	4029
Model:	Logit	**Df Residuals:**	4025
Method:	MLE	**Df Model:**	3
Date:	Sat, 26 Sep 2020	**Pseudo R-squ.:**	0.01901
Time:	20:28:39	**Log-Likelihood:**	-2337.9
converged:	True	**LL-Null:**	-2383.2
Covariance Type:	nonrobust	**LLR p-value:**	1.605e-19

	coef	std err	z	P>|z|	[0.025	0.975]
const	-0.9781	0.036	-27.192	0.000	-1.049	-0.908
x1	0.1044	0.035	2.979	0.003	0.036	0.173
x2	0.3138	0.038	8.161	0.000	0.238	0.389
x3	0.1305	0.035	3.765	0.000	0.063	0.198

Figure 1: Logit Regression Result

0.3077 with p-value: 0.000), Model3('gaodescore': 0.1454 with p-value: 0.000). Here we see, for each single model, only 'tencentscore' and 'gaodescore' are significant.

4 Question3

Perform a logit regression and examine the relation between the loan approval likelihood and credit scores.

4.1 Answer

By using the following Python codes, we performed a logit regression between the loan approval likelihood and credit scores.

```
q3_X = Key_Data[['creditlevelasbuyer', 'tencentscore', 'gaodescore']]
q3_y = Key_Data['deal']
q3_X = scaler.fit(q3_X).transform(q3_X)
q3_X = sm.add_constant(q3_X)
```

Figure 2 is the result of this Logit Regression. We can see that all the p-values are significant, and the coefficients of x1(which is 'creditlevelasbuyer') is positive, which means loan approval has positive relationship with it. For the other two variables, there are negative relationships.

5 Question4

Perform a logit regression and examine the relation between the loan approval likelihood and the dummy whether the borrower has a frequent contact

5.1 Answer

By using the following Python codes, we performed a logit regression between the loan approval likelihood and the dummy whether the borrower has a frequent contact ('highcontact').

```
q4_X = Key_Data['highcontact']
q4_y = Key_Data['deal']
q4_X = scaler.fit(q4_X.values.reshape(-1, 1)).transform(q4_X.values.reshape(-1, 1))
q4_X = sm.add_constant(q4_X)
q4_model = sm.Logit(q4_y, q4_X)
q4_result = q4_model.fit()
```

Figure 3 is the result of this Logit Regression. We can see that all the p-values are significant, and the coefficients of x1(which is 'highcontact') is positive, which means loan approval has positive relationship with it.

Logit Regression Results

Dep. Variable:	deal	**No. Observations:**	4029
Model:	Logit	**Df Residuals:**	4025
Method:	MLE	**Df Model:**	3
Date:	Sat, 26 Sep 2020	**Pseudo R-squ.:**	0.04411
Time:	20:28:39	**Log-Likelihood:**	-2651.2
converged:	True	**LL-Null:**	-2773.5
Covariance Type:	nonrobust	**LLR p-value:**	9.324e-53

	coef	**std err**	**z**	**P>|z|**	**[0.025**	**0.975]**
const	-0.1993	0.033	-6.098	0.000	-0.263	-0.135
x1	0.2019	0.039	5.182	0.000	0.126	0.278
x2	-0.3976	0.035	-11.445	0.000	-0.466	-0.329
x3	-0.1589	0.035	-4.580	0.000	-0.227	-0.091

Figure 2: Logit Regression Result

Logit Regression Results

Dep. Variable:	deal	**No. Observations:**	4029
Model:	Logit	**Df Residuals:**	4027
Method:	MLE	**Df Model:**	1
Date:	Sat, 26 Sep 2020	**Pseudo R-squ.:**	0.001289
Time:	20:28:39	**Log-Likelihood:**	-2769.9
converged:	True	**LL-Null:**	-2773.5
Covariance Type:	nonrobust	**LLR p-value:**	0.007503

	coef	**std err**	**z**	**P>|z|**	**[0.025**	**0.975]**
const	-0.1961	0.032	-6.187	0.000	-0.258	-0.134
x1	0.0847	0.032	2.672	0.008	0.023	0.147

Figure 3: Logit Regression Result